General Conditions for Bounded Relative Error in Simulations of Highly Reliable Markovian Systems

Marvin K. Nakayama

Department of Computer and Information Science
New Jersey Institute of Technology
Newark, NJ 07102

Abstract

We establish a necessary condition for any importance sampling scheme to give bounded relative error when estimating a performance measure of a highly reliable Markovian system. Also, a class of importance sampling methods is defined for which we prove a necessary and sufficient condition for bounded relative error for the performance measure estimator. This class of probability measures includes all of the currently existing failure biasing methods in the literature. Similar conditions for derivative estimators are established.

SIMULATION; IMPORTANCE SAMPLING; LIKELIHOOD RATIOS; GRADIENT ESTIMATION; RELIABILITY; MARKOV CHAINS.

AMS 1991 Subject Classifications: Primary: 65C05
Secondary: 60J10, 60K10

1 Introduction

There is an increasing demand for systems, such as computing systems or transaction processing systems, to be highly reliable. A designer faced with developing such a system usually constructs and evaluates a mathematical model of the system to determine if it will perform at an acceptable level. Analytic methods for evaluating models are often impractical due to the large state spaces that arise in reliability models, and frequently the designer must resort to simulation. However, standard simulation without the use of any variance reduction techniques is inefficient because of the rarity of system failures. Thus, variance reduction techniques must be employed to obtain efficient estimators which yield acceptable confidence intervals.

Importance sampling is one such variance reduction technique and is the focus of our study. The basic idea behind this method is to simulate the system under a modified probability distribution that is chosen in a way so that the important events (in our case, system failures) which are rare under the original probability measure occur more frequently. This technique is known as a "change of measure." By properly selecting the importance sampling probability distribution, we can significantly reduce the variance of the estimator.

In the literature a number of importance sampling schemes have been proposed for simulating highly reliable Markovian systems. These include all of the failure biasing methods, namely simple failure biasing (Lewis and Böhm [17], Conway and Goyal [4], Goyal et al. [11], Shahabuddin [25], Nakayama [18]), bias2 failure biasing (Goyal et al. [11]), balanced failure biasing (Shahabuddin [25], Goyal et al. [11]), and failure distance biasing (Carrasco [1, 2]). Shahabuddin [25] developed the mathematical framework to study the asymptotic properties of estimators obtained using importance sampling in simulations of highly reliable Markovian systems. In particular, Shahabuddin introduced the notion of "bounded relative error" in this problem setting. A simulation estimator of a performance measure has bounded relative error if the ratio of the expected half width of its confidence interval over the expected point estimate remains bounded as the component failure rates tend to zero and the repair rates remain fixed. If an estimator enjoys this property, then only a fixed number of samples is needed to obtain a confidence interval having a fixed expected relative width, independent of how rarely system failures occur. Otherwise, the sample size must increase as system failures become less frequent.

Previous theoretical work on importance sampling for highly reliable Markovian systems mainly focused on the asymptotic properties of particular changes of measure. Shahabuddin [25] developed a simple sufficient condition for certain importance sampling schemes to yield bounded relative error and used it to establish that balanced failure biasing always yields bounded relative error. It was also proved that if the system under consideration is "balanced"

(i.e., the transition rates of all of the failure transitions are of the same order of magnitude), then simple failure biasing produces estimators having bounded relative error. Moreover, Shahabuddin showed by example that if the system is not balanced, then simple failure biasing may not give bounded relative error. Nakayama [18] later demonstrated that simple failure biasing can yield bounded relative error for certain unbalanced systems, and a necessary and sufficient condition for bounded relative error was established for this method. Nakayama [19, 18] also examined the issue of bounded relative error for likelihood ratio derivative estimators obtained using balanced failure biasing and simple failure biasing.

While much of the previous work focused on the asymptotic efficiencies of specific failure biasing methods, the goal of this paper is to unify and generalize the existing theory by establishing a number of conditions for bounded relative error for large classes of importance sampling methods. To this end, we first establish a necessary condition for *any* importance sampling scheme to produce a performance estimator having bounded relative error for a given system. This result is quite general since the only required assumption is that the importance sampling probability measure is a valid change of measure; no other conditions are imposed on the structure of the distrubution. Thus, even though we are considering Markovian systems, our theorem can be used to examine importance sampling methods which are non-Markovian.

One consequence of our necessary condition is the following. To determine if an importance sampling estimator has bounded relative error for a given system, it may seem plausible that it is sufficient to analyze the behavior of the estimator on only the most likely paths to system failure. However, this is not the case. In fact, our result shows that in the context of highly reliable Markovian systems, the probability under the new measure of *every* path to failure must satisfy some condition for there to be bounded relative error. Nakayama [18] first noted this need to examine paths which are not the most likely ones in an analysis of the simple failure biasing method. Our current result generalizes this observation to any arbitrary importance sampling scheme. It also clearly demonstrates in general the role that each path plays in determining the relative error of an estimator. Moreover, as noted in Nakayama [18], this feature, in some sense, illustrates the difference between importance sampling schemes for the highly reliable systems considered here and those arising in the "large deviations" context. Specifically, the optimal change of measure for large deviations problems is selected solely with regard to the most likely path to failure; e.g., see Cottrell et al. [5] and Chang et al. [3]. However, our necessary condition shows that in the highly reliable Markovian system context, it is not sufficient to consider only the most likely paths to failure.

To obtain a necessary and sufficient condition for bounded relative error, we must impose additional structure on the importance sampling scheme considered. Thus, we define a certain

broad class of importance sampling methods and prove a theorem which establishes a necessary and sufficient condition for any importance sampling method in this class to give rise to performance measure estimates having bounded relative error for a given system. The class includes all of the failure biasing techniques currently in the literature. Our current result generalizes the work of Nakayama [18], which established a necessary and sufficient condition for bounded relative error for the special case of simple failure biasing.

The necessary and sufficient condition may be difficult to apply in practice, and so we provide some simple sufficient conditions for bounded relative error of the performance measure estimators. The first condition is due to Shahabuddin [25] and was used in [25] to show that balanced failure biasing always gives bounded relative error and that simple failure biasing does so when the system is balanced. We prove a theorem which demonstrates that a large class of importance sampling methods satisfy Shahabuddin's sufficient condition when the system is balanced. The class includes all of the currently existing failure biasing schemes.

We apply our results to study each of the failure biasing methods which are currently in the literature. In particular, we show that of all these techniques, only balanced failure biasing is guaranteed to always give bounded relative error. However, as Nakayama [18] showed by example, for a given model, the simple failure biasing method may yield estimators with smaller constants for the leading term in the asymptotic expansion of the variance. These constants are important in practice since they play an important role in determining the actual width of the resulting confidence interval. Thus, although balanced failure biasing is the most robust of the existing methods, it still may be more appropriate to use one of the other schemes for a particular model.

We also perform a similar analysis to determine when a given importance sampling method will give rise to likelihood ratio derivative estimators having bounded relative error for a given model. First, we establish a necessary condition for any importance sampling scheme to yield derivative estimators having bounded relative error. Then, we prove a necessary and sufficient condition for the class of importance sampling measures described above. Finally, we provide simple sufficient conditions for the derivative estimators to have bounded relative error. (For other work on derivative estimation which are not necessarily in the setting of highly reliable systems, see [6, 7, 16, 19, 20, 24, 27] and references therein.)

In addition to the previous work cited on importance sampling for highly reliable systems, others also have studied this problem. Shahabuddin [26] and Shahabuddin and Nakayama [27] analyzed the asymptotic properties of importance sampling estimators of transient performance measures and their derivatives for Markovian systems. Also, several importance sampling schemes have been proposed for highly reliable non-Markovian systems. Nicola et al. [21]

proposed an algorithm in which some of the clocks (i.e., future events in the event list) are rescheduled after certain events, and the technique was studied empirically. Nicola et al. [22] developed another method for estimating the system reliability based on the idea of uniformization and a concept which they call "exponential transform," and Heidelberger et al. [14] established under certain conditions that the method yields estimates with bounded relative error. Heidelberger et al. [13] and Nicola et al. [23] also studied the method in various settings.

The rest of the paper is organized as follows. In Section 2 we present the mathematical model of highly reliable Markovian systems developed by Shahabuddin [25]. We examine the asymptotic behavior of performance measure estimators obtained using importance sampling in Section 3. In Section 3.1, we establish our necessary condition for any importance sampling scheme to yield performance measure estimates with bounded relative error. Then, in Section 3.2 we define our class of importance sampling methods for which we prove the necessary and sufficient condition for bounded relative error of the performance measure estimate. Section 3.3 contains some sufficient conditions for bounded relative error. We apply our results to study each of the existing failure biasing methods in Section 3.4. Section 4 establishes results analogous to those in Section 3 but for derivative estimators. In Section 5 we state our conclusions and give some directions for future research. Finally, an appendix contains one of the longer proofs.

2 Mathematical Model

We now describe the mathematical model of highly reliable Markovian systems with which we will work. The model was originally developed by Shahabuddin [25] to study the asymptotic behavior of performance measure estimators and later was modified by Nakayama [19] to analyze likelihood ratio derivative estimators.

Our system consists of C, $0 < C < \infty$, different types of components, where there are n_i, $0 < n_i < \infty$, components of type i. We let

$$N = \sum_{j=1}^{C} n_j$$

be the total number of components in the system. The components are subject to random failures and, when failed, are sent to a repair facility having some number of repairpersons. The queueing discipline at the repair center is arbitrary.

We will model the evolution of the system as a continuous time Markov chain (CTMC) $\{Y_t : t \geq 0\}$ having some state space S, where we assume that $|S| < \infty$. Our analysis will be independent of the exact form of the state space. Note that any state $x \in S$ is an encoding of

the number of failed components of each type along with any information about the queueing at the repair facility. Let $n_i(x)$ be the number components of type i that are operational in state x. We decompose the state space as $S = U \cup F$, where U is the set of operational (or up) states and F is the set of failed (or down) states. We assume that if $x \in U$ and $y \in E$ with $n_i(y) \geq n_i(x)$ for all components types i, then $y \in U$. We also assume that the system initially starts in state 0, the state with all components operational, and $0 \in U$.

We allow for the possibility of failure propagation; i.e., the failure of one component causes other components to simultaneously fail with some probability. For example, consider a computer system with a processor and a power supply, and the failure of a power supply creates a power surge which causes the processor to fail. We model this in the following manner. Consider some component type i and state $x \in S$, and define $S_i(x) = \{y \in S : n_j(y) \leq n_j(x) \text{ for all } j \neq i,\ n_i(y) < n_i(x)\}$, which is the set of states y in which there is at least one more component of type i failed than in state x and every other component type has at least as many failed components as in state x. Also, let $p(\,\cdot\,; x, i)$ be some probability mass function on $S_i(x)$. Suppose that the system is in a state x with $n_i(x) > 0$ and a component of type i fails. Then the system immediately enters state $y \in S_i(x)$ with probability $p(y; x, i)$. In this situation, $n_j(x) - n_j(y)$ components of type j, $1 \leq j \leq C$, failed on the transition caused by the failure of the component of type i, and we say that the failure of the component of type i *triggered* the transition (x, y).

Similarly, we allow for the possibility of a repairperson to complete the repair of more than one component at a time. For example, this may happen if some component consists of a number of subcomponents, and the repairperson replaces the entire unit when enough of the subcomponents have failed. However, we do not allow for a single transition to consist of a number of components failing and others completing repair. This may occur, for example, if a repairperson fixes some component but breaks another one when replacing the repaired component.

We define a transition (x, y) to be a *failure transition*, which we denote by $y \succ x$, if $n_j(y) \leq n_j(x)$ for all $1 \leq j \leq C$ with $n_i(y) < n_i(x)$ for some type i. Similarly, we define (x, y) to be a *repair transition*, which we denote by $y \prec x$, if $n_j(y) \geq n_j(x)$ for all $1 \leq j \leq C$ with $n_i(y) > n_i(x)$ for some type i.

We let the behavior of a component depend on the state of the system. Thus, when the system is in state $x \in S$, we assume that the failure rate of components of type i is $\lambda_i(x) \geq 0$ and the rate of some repair transition (x, y) is $\mu(x, y) \geq 0$. Using this approach, we can allow for the operation of one component depend on other components being operational. For example, this may occur when a processor in a computer has a power supply, and the processor

is inoperational if the power supply is failed. Similarly, the repairperson may not be able to fix the processor until the power supply is repaired.

The infinitesimal generator matrix $Q = \{q(x,y) : x, y \in S\}$ of Y is given by

$$q(x,y) = \begin{cases} \sum_{k=1}^{C} n_k(x)\lambda_k(x)p(y;x,k) & \text{if } y \succ x \\ \mu(x,y) & \text{if } y \prec x \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

for $x \neq y$, and $q(x,x) = -\sum_{y \neq x} q(x,y)$. We let

$$q(x) = -q(x,x)$$

be the total transition rate out of state x.

We let $X = \{X_n : n \geq 0\}$ denote the embedded discrete time Markov chain (DTMC) of Y. The transition matrix of X is given by $\mathbf{P} = \{\mathbf{P}(x,y) : x, y \in S\}$, where $\mathbf{P}(x,y) = q(x,y)/q(x)$ for $x \neq y$, and $\mathbf{P}(x,x) = 0$. We define $\Gamma = \{(x,y) : x, y \in S, \mathbf{P}(x,y) > 0\}$, which is the set of possible transitions the system can make.

We assume that the system is composed of highly reliable components (i.e., the component failure rates are much smaller than the repair rates). (High reliability for the system can also be achieved by having high redundancies.) We model this by introducing a parameter ϵ and assume that the failure rate of the components of type i, $1 \leq i \leq C$, is

$$\lambda_i(x,\epsilon) = \tilde{\lambda}_i(x)\epsilon^{b_i(x)},$$

where $\tilde{\lambda}_i(x) \geq 0$ and $b_i(x) \geq 1$ are independent of ϵ, and $b_i(x)$ is integer-valued. We also let $p(\,\cdot\,; x, i)$ depend on ϵ; i.e., for all $(x,y) \in \Gamma$ such that $y \succ x$,

$$p(y;x,i) = p_\epsilon(y;x,i) = c_i(x,y)\epsilon^{d_i(x,y)}$$

where $d_i(x,y) \geq 0$ is integer-valued, $c_i(x,y) \geq 0$, and $\sum_{y \in S_i(x)} p_\epsilon(y;x,i) = 1$. We assume that the repair rates $\mu(x,y)$ are independent of ϵ. We will examine the behavior of the system as $\epsilon \to 0$.

For some constant d, a function f is said to be $o(\epsilon^d)$ if $f(\epsilon)/\epsilon^d \to 0$ as $\epsilon \to 0$. Similarly, $f(\epsilon) = O(\epsilon^d)$ if $|f(\epsilon)| \leq c_1\epsilon^d$ for some constant $c_1 > 0$ for all ϵ sufficiently small. Also, $f(\epsilon) = \underline{O}(\epsilon^d)$ if $|f(\epsilon)| \geq c_2\epsilon^d$ for some constant $c_2 > 0$ for all ϵ sufficiently small. Finally, $f(\epsilon) = \Theta(\epsilon^d)$ if $f(\epsilon) = O(\epsilon^d)$ and $f(\epsilon) = \underline{O}(\epsilon^d)$.

We define b_0 to be

$$b_0 \equiv \min_{1 \leq i \leq C} b_i(0),$$

and so $q(0) = \Theta(\epsilon^{b_0})$. For any $(x, y) \in \Gamma$, we define

$$b(x,y) = \begin{cases} \min\{b_i(x) + d_i(x,y) : 1 \le i \le C,\ n_i(x)\tilde{\lambda}_i(x)p_\epsilon(y;x,i) > 0\} & \text{if } y \succ x \\ 0 & \text{if } y \prec x \end{cases}, \quad (2)$$

which is the exponent of the order of magnitude of the rate of a transition (x, y). Thus, for any $(x, y) \in \Gamma$, $b(x, y) = d$ if $q(x, y) = \Theta(\epsilon^d)$, and $b(x, y) \ge 1$ if $y \succ x$.

We say that the system is *balanced* if the transition rates of all of the failure transitions are of the same order of magnitude (i.e., if for all $(x, y) \in \Gamma$ with $y \succ x$, $b(x, y) = b$ for some $b \ge 1$) and all of the $p_\epsilon(y; x, i)$ are independent of ϵ. In this situation, we may assume without loss of generality that $b = 1$. If the system is not balanced, then it is said to be *unbalanced*.

We will assume the following:

A1 *The DTMC X is irreducible over the state space S.*

A2 *For each state $x \in S$ with $x \neq 0$, there exists a state $y \in S$ such that $(x, y) \in \Gamma$ and $y \prec x$.*

A3 *For each state $z \in F$ such that $(0, z) \in \Gamma$, $q(0, z) = o(\epsilon^{b_0})$.*

Assumption A2 states that there is at least one repair transition possible from every state $x \neq 0$. This will be satisfied as long as a repairperson is busy whenever there are any components failed. If this is the case, then for $x \neq 0$, $q(x) = c(x) + o(1)$, where $c(x) > 0$, which implies that all failure transitions (x, y) with $x \neq 0$ have transition probability $\mathbf{P}(x, y) = \Theta(\epsilon^{b(x,y)})$. The assumption does not hold if there are deferred repairs; i.e., a repairperson does not start fixing a failed component until some number (greater than one) of components are failed. In this last situation Juneja and Shahabuddin [15] showed that the standard failure biasing techniques will not yield estimators having bounded relative error.

Assumption A3 stipulates that all transitions which take the system from the original state 0 immediately to a failed state must have transition rates which are much smaller than the largest transition rates from state 0. This ensures that system failures are rare events for the embedded DTMC when ϵ is small.

Our assumptions imply that the elements of the transition matrix have the following form. For any $(x, y) \in \Gamma$,

$$\mathbf{P}(x,y) = \begin{cases} \Theta(\epsilon^{b(x,y)}) & \text{if } x \neq 0 \\ \Theta(\epsilon^{b(x,y)-b_0}) & \text{if } x = 0 \end{cases}, \quad (3)$$

as $\epsilon \to 0$, where $b(x, y)$ is defined in (2). Note that since $b(x, y) = 0$ whenever $y \prec x$, all repair transitions have transition probabilities which are $\Theta(1)$.

We concentrate on estimating $\gamma \equiv P\{\tau_F < \tau_0\}$, where τ_A denotes the hitting time of the DTMC X to some set of states A; i.e., $\tau_A = \inf\{n > 0 : X_n \in A\}$. This performance measure is of interest for several reasons. First, the mean time to failure can be expressed as

$$MTTF = \frac{\xi}{\gamma}, \tag{4}$$

where $\xi = E[\sum_{k=0}^{\tau_{\min}-1} 1/q(X_k)]$ and $\tau_{\min} = \min\{n > 0 : X_n \in \{0, F\}\}$; e.g., see Goyal et al. [11]. Also, consider the unreliability of the system at time t; i.e., $U(t) = P\{T_F < t\}$, where T_F is the hitting time of the CTMC Y to the set F; i.e., $T_F = \inf\{t > 0 : Y_t \in F\}$. Then, Shahabuddin and Nakayama [27] established that $(1 - e^{-q(0)\gamma t})/U(t) \to 1$ as $\epsilon \to 0$ when $t = \Theta(\epsilon^{-r_t})$ with $r_t > 0$.

Shahabuddin [25] showed that if Assumptions A1–A3 hold, then there exists some constant $r \geq 1$ (which depends on the model being considered) such that

$$\gamma = \Theta(\epsilon^r). \tag{5}$$

We define

$$\begin{aligned} \Delta \;=\; \{\; & (x_0, \ldots, x_n) \;:\; n \geq 1,\; x_0 = 0,\; x_n \in F,\; x_i \notin \{0, F\} \text{ for } 1 \leq i < n, \\ & (x_i, x_{i+1}) \in \Gamma \text{ for } 0 \leq i < n \;\}, \end{aligned}$$

which is the set of sample paths of the embedded DTMC for which $\tau_F < \tau_0$. Furthermore, let

$$\Delta_m = \{(x_0, \ldots, x_n) \in \Delta : n \geq 1,\; P\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} = \Theta(\epsilon^m)\}$$

be the set of sample paths for which $\tau_F < \tau_0$ and have probability (under the original measure) of the order ϵ^m. Note that $\Delta = \cup_{m=r}^{\infty} \Delta_m$, where r is defined in (5).

3 Estimating the Performance Measure Using Importance Sampling

3.1 A General Necessary Condition For Bounded Relative Error

Since we will only consider the performance measure $\gamma = P\{\tau_F < \tau_0\}$, we can concentrate solely on the embedded DTMC X and do not have to work directly with the CTMC Y. Thus, we can define our sample space Ω as

$$\Omega = \{\; (x_0, \ldots, x_n) \;:\; n \geq 1,\; x_0 = 0,\; x_n \in \{0, F\},\; x_i \notin \{0, F\} \text{ for } 1 \leq i < n \;\},$$

which is the set of state sequences which start in state 0 and end in either 0 or F. Let $(\Omega, \mathcal{F})$ denote the probability space on which X is defined, and let P be the probability measure on $(\Omega, \mathcal{F})$ induced by the Q-matrix given in (1).

Consider estimating $\gamma = E[1\{\tau_F < \tau_0\}]$ using standard simulation. We accomplish this by generating i.i.d. samples $\hat{I}_1, \ldots, \hat{I}_n$ of $1\{\tau_F < \tau_0\}$ using the original probability measure P. The point estimate of γ is given by

$$\hat{\gamma}(n) = \frac{1}{n} \sum_{k=1}^{n} \hat{I}_k,$$

and the variance of $1\{\tau_F < \tau_0\}$ under the measure P is

$$\sigma^2 = \gamma - \gamma^2 = \Theta(\epsilon^r) - \Theta(\epsilon^{2r}) = \Theta(\epsilon^r)$$

as $\epsilon \to 0$. We define the relative error of our estimator to be the expected relative half-width of the resulting confidence interval for a fixed number of samples n and a given confidence level $1 - \delta$. Letting z_δ denote the $1 - \delta/2$ quantile of a standard normal distribution, we have that the relative error is

$$RE = z_\delta \frac{\sqrt{\sigma^2/n}}{\gamma} = \frac{z_\delta}{\sqrt{n}} \frac{\Theta(\epsilon^{r/2})}{\Theta(\epsilon^r)} = \frac{z_\delta}{\sqrt{n}} \Theta(\epsilon^{-r/2}) \to \infty$$

as $\epsilon \to 0$. Thus, the difficulty of estimating γ using standard simulation increases as system failures become rarer.

Importance sampling, which we describe below, is a technique which can be used to obtain more efficient estimators (when used properly). Recalling that Δ is the set of sample paths for which $\tau_F < \tau_0$, we define the following class of probability measures.

Definition 1 $\mathcal{I}$ is the class of probability measures P' defined on $(\Omega, \mathcal{F})$ such that $P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} > 0$ for all $(x_0, \ldots, x_n) \in \Delta$.

The class $\mathcal{I}$ is the set of valid importance sampling probability measures for estimating γ. Hence, for any $P' \in \mathcal{I}$,

$$\begin{aligned}
\gamma &= E[1\{\tau_F < \tau_0\}] = \sum_{(x_0,\ldots,x_n)\in\Delta} P\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} \\
&= \sum_{(x_0,\ldots,x_n)\in\Delta} \frac{P\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\}}{P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\}} P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} \\
&= \sum_{(x_0,\ldots,x_n)\in\Delta} L(x_0, \ldots, x_n) P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} = E'[1\{\tau_F < \tau_0\} L],
\end{aligned}$$

where E' is the expectation operator induced by the measure P' and

$$L(x_0, \ldots, x_n) = \frac{P\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\}}{P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\}}$$

is the Radon-Nikodym derivative of P with respect to P', or simply the likelihood ratio. See Hammersley and Handscomb [12] or Glynn and Iglehart [8] for further details.

Actually, $\mathcal{I}$ is a generalization of the "standard" class of legitimate importance sampling distributions since there may exist some $(x_0, \ldots, x_n) \notin \Delta$, $(x_0, \ldots, x_n) \in \Omega$ such that $P\{(X_0, \ldots, X_{\tau_{\min}}) = (x_0, \ldots, x_n)\} > 0$ and $P'\{(X_0, \ldots, X_{\tau_{\min}}) = (x_0, \ldots, x_n)\} = 0$. However, Definition 1 ensures that the new probability measure is positive over the part of the sample space that matters (i.e., $\{\tau_F < \tau_0\}$), which is sufficient; see Glynn and Iglehart [8].

We apply importance sampling as follows. Generate i.i.d. samples $(\tilde{I}_1, \tilde{L}_1), \ldots, (\tilde{I}_n, \tilde{L}_n)$ of $(1\{\tau_F < \tau_0\}, L)$ using the probability measure P'. We form the new point estimate

$$\tilde{\gamma}(n) = \frac{1}{n} \sum_{k=1}^{n} \tilde{I}_k \tilde{L},$$

and the variance of $1\{\tau_F < \tau_0\}L$ under the measure P' is

$$\sigma'^2 = E'[1\{\tau_F < \tau_0\}L^2] - \gamma^2.$$

Note that

$$\begin{aligned} E'[1\{\tau_F < \tau_0\}L^2] &= \sum_{(x_0,\ldots,x_n)\in\Delta} L^2(x_0, \ldots, x_n) P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} \\ &= \sum_{(x_0,\ldots,x_n)\in\Delta} L(x_0, \ldots, x_n) P\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} \\ &= E[1\{\tau_F < \tau_0\}L], \end{aligned} \tag{6}$$

and so we can compute the second moment under importance sampling in terms of the original probability measure. We define the relative error of the importance sampling estimator to be

$$RE' = z_\delta \frac{\sqrt{\sigma'^2/n}}{\gamma}.$$

The goal of importance sampling is to choose a $P' \in \mathcal{I}$ such that $E[1\{\tau_F < \tau_0\}L] < \gamma$, thereby reducing the variance over standard simulation. In fact, if we can select a $P' \in \mathcal{I}$ such that $\sigma'^2 = O(\epsilon^{2r})$, then

$$RE' = \frac{z_\delta}{\sqrt{n}} \frac{\sqrt{O(\epsilon^{2r})}}{\Theta(\epsilon^r)} = \frac{z_\delta}{\sqrt{n}} O(1),$$

which remains bounded (and possibly goes to zero) as $\epsilon \to 0$. Thus, we can obtain a good estimate of γ independently of how rare system failures are. If an estimator satisfies this property, we say that it has *bounded relative error*, a notion introduced by Shahabuddin [25].

To determine if an importance sampling estimator of a performance measure has bounded relative error for a given system, it might seem plausible that it is sufficient to analyze the behavior of the estimator on only the most likely paths to system failure (i.e., $(x_0, \ldots, x_n) \in \Delta_r$). However, the following result shows that this is not the case.

Theorem 1 *Consider any system satisfying Assumptions A1–A3. Also, consider any $P' \in \mathcal{I}$, and let RE' denote the relative error of the estimator of γ obtained using P'. Suppose $\gamma = \Theta(\epsilon^r)$ for some $r \geq 1$. If RE' remains bounded as $\epsilon \to 0$, then for all $(x_0, \ldots, x_n) \in \Delta_m$, $m \geq r$, $P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} = \underline{O}(\epsilon^{2m-2r})$.*

Proof. Suppose there exists some path $(y_0, \ldots, y_k) \in \Delta_m$, $m \geq r$, such that $P'\{(X_0, \ldots, X_{\tau_F}) = (y_0, \ldots, y_k)\} = O(\epsilon^{2m-2r+1})$. Using (6), we obtain

$$\begin{aligned} E'[1\{\tau_F < \tau_0\}L^2] &= \sum_{\substack{(x_0,\ldots,x_n)\in\Delta \\ n>0}} L(x_0, \ldots, x_n) P\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} \\ &\geq L(y_0, \ldots, y_k) P\{(X_0, \ldots, X_{\tau_F}) = (y_0, \ldots, y_k)\} \\ &= \frac{\Theta(\epsilon^m)}{O(\epsilon^{2m-2r+1})}\Theta(\epsilon^m) \;=\; \underline{O}(\epsilon^{2r-1}). \end{aligned}$$

Hence, since $\gamma = \Theta(\epsilon^r)$,

$$RE' = z_\delta \frac{\sqrt{\sigma'^2/n}}{\gamma} \geq \frac{z_\delta}{\sqrt{n}} \frac{\sqrt{\underline{O}(\epsilon^{2r-1})}}{\Theta(\epsilon^r)} = \frac{z_\delta}{\sqrt{n}} \underline{O}(\epsilon^{-1/2}) \to \infty$$

as $\epsilon \to 0$. ■

Theorem 1 shows that we cannot solely concentrate on the most likely paths to failure when designing an importance sampling scheme for simulating highly reliable Markovian systems. In a study of the simple failure biasing method, Nakayama [18] first noted the need to examine the secondary paths to failure (i.e., $(x_0, \ldots, x_n) \in \Delta - \Delta_r$) to determine if the resulting estimator will have bounded relative error. However, Theorem 1 clearly illustrates in general how the behavior of every path to failure affects the relative error of the estimator. Also, it is interesting to note that Theorem 1 implies that if the performance measure estimator has bounded relative error, then each of the most likely paths to failure must have probability of $\Theta(1)$ under the new measure P'.

Furthermore, Theorem 1 made no assumptions about the structure of P' other than it must be a valid importance sampling measure. In particular, it was not assumed that the importance sampling scheme is Markovian, even though the original measure P is.

3.2 A Necessary and Sufficient Condition For Bounded Relative Error

To obtain a condition that is both necessary and sufficient, we need to assume that the importance sampling methodology has more structure. Hence, we make the following definition.

Definition 2 $\mathcal{J}$ is the class of probability measures P' defined on $(\Omega, \mathcal{F})$ which satisfy the following properties:

(i) P' is Markovian with some transition matrix $\mathbf{P}'$;

(ii) For any $(w, y) \in \Gamma$, if $\mathbf{P}(w, y) = \Theta(\epsilon^d)$, then $\mathbf{P}'(w, y) = \underline{O}(\epsilon^d)$.

It is easy to show that $\mathcal{J} \subset \mathcal{I}$, and so any $P' \in \mathcal{J}$ is a valid importance sampling measure. Part (ii) of Definition 2 states that the probability of a transition under the new measure is never significantly smaller than the probability under the original measure. Furthermore, as we shall see later, the class $\mathcal{J}$ contains all of the failure biasing methods currently in the literature. The following theorem establishes a necessary and sufficient condition for any probability measure in $\mathcal{J}$ to give bounded relative error for the performance measure estimate.

Theorem 2 *Consider any system satisfying Assumptions A1–A3. Also, consider any $P' \in \mathcal{J}$, and let RE' denote the relative error of the estimator of γ obtained using P'. Suppose $\gamma = \Theta(\epsilon^r)$ for some $r \geq 1$. Then, RE' remains bounded as $\epsilon \to 0$ if and only if for all $(x_0, \ldots, x_n) \in \Delta_m$, $r \leq m \leq 2r - 1$, $P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} = \underline{O}(\epsilon^{2m-2r})$.*

Before proving the theorem, we first make some remarks about it. Theorem 2 can be potentially difficult to apply in practice because of the number of sample paths that must be examined. However, our result clearly shows that secondary paths to failure play an important role in determining the variance of an estimator obtained using some importance sampling method. (Later, we will state a simple sufficient condition due to Shahabuddin [25] for the performance measure estimator to have bounded relative error.)

We now compare Theorems 1 and 2. Consider $(x_0, \ldots, x_n) \in \Delta_m$ for some $m \geq r$. To apply Theorem 1, we must consider *all* $m \geq r$, whereas in Theorem 2, we only need to examine $r \leq m \leq 2r - 1$. To see why this is true, note that from Definition 2(ii), if $P' \in \mathcal{J}$, then $P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} = \underline{O}(\epsilon^m)$ for all $(x_0, \ldots, x_n) \in \Delta_m$ for any $m \geq r$. Hence, the condition that $P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} = \underline{O}(\epsilon^{2m-2r})$ is automatically satisfied when $m \geq 2r$ and $P' \in \mathcal{J}$. On the other hand, we cannot make the same conclusion for $P' \in \mathcal{I}$ since not enough structure was imposed on the class $\mathcal{I}$.

Theorem 2 generalizes a result established in Nakayama [18] which provided a necessary and sufficient condition for the special case of when simple failure biasing yields bounded relative error for the estimator of γ. However, Theorem 2 is more general since it applies to a broad class of measures $\mathcal{J}$ and not only to a specific method.

Now we proceed with the proof of Theorem 2. To establish the result, we will use the following lemma.

Lemma 1 *Consider any system satisfying Assumptions A1–A3. Consider $(x_0, \ldots, x_n) \in \Delta_m$, where $n > 0$ and $m \geq r$. Then*

(i) $n \leq (m+1)N$;

(ii) $|\Delta_m| \leq |S|^{(m+1)N}$;

(iii) $P\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} = \Theta(\epsilon^m)$ *and* $P\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} \leq \alpha\beta^m\epsilon^m$ *for all* $\epsilon > 0$ *sufficiently small, where* α *and* β *are constants which are independent of* $(x_0, \ldots, x_n)$ *and* m.

Furthermore, suppose $P' \in \mathcal{J}$, *and let* L *denote the Radon-Nikodym derivative of* P *with respect to* P'. *Then*

(iv) $L(x_0, \ldots, x_n) \leq \eta^{m+1}$ *for all* $\epsilon > 0$ *sufficiently small, where* η *is a constant which is independent of* $(x_0, \ldots, x_n)$, m, *and* ϵ.

Proof. The first part of (iii) is immediate from the definition of Δ_m. Parts (i) and (ii) and the upper bound of part (iii) are slight generalizations of results previously established in the proof of Theorem 1 of Nakayama [19], and so we omit their proofs.

To prove the validity of part (iv), first note that for any $P' \in \mathcal{J}$ and any $(x_0, \ldots, x_n) \in \Delta_m$,

$$P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} = \prod_{k=0}^{n-1} \mathbf{P}'(x_k, x_{k+1})$$

by part (i) of Definition 2. Thus,

$$L(x_0, \ldots, x_n) = \prod_{k=0}^{n-1} \frac{\mathbf{P}(x_k, x_{k+1})}{\mathbf{P}'(x_k, x_{k+1})}.$$

By part (ii) of Definition 2, for any $(x, y) \in \Gamma$, there exists some $\zeta(x, y) > 0$ which is independent of ϵ such that

$$\mathbf{P}'(x, y) \geq \zeta(x, y)\mathbf{P}(x, y)$$

for all ϵ sufficiently small. Define $\zeta' = \min\{\zeta(x, y) : (x, y) \in \Gamma\}$ and $\zeta_* = \min\{1, \zeta'\}$. Note that $\zeta_* > 0$ since $|S| < \infty$. Thus, for all sufficiently small $\epsilon > 0$,

$$L(x_0, \ldots, x_n) \leq \prod_{k=0}^{n-1} \frac{1}{\zeta_*} \leq \frac{1}{\zeta_*^{(m+1)N}}$$

by part (i). The proof is completed by letting $\eta = 1/\zeta_*^N$. ∎

Now we establish Theorem 2.

Proof of Theorem 2. Suppose that $P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} = \underline{O}(\epsilon^{2m-2r})$ for all $(x_0, \ldots, x_n) \in \Delta_m$, $r \leq m \leq 2r$. Since $\gamma = \Theta(\epsilon^r)$, we need to establish that $E'[1\{\tau_F < \tau_0\}L^2] = O(\epsilon^{2r})$.

From (6), we have that

$$E'[1\{\tau_F < \tau_0\}L^2] = \sum_{m=r}^{\infty} \sum_{\substack{(x_0,\ldots,x_n)\in\Delta_m \\ n>0}} L(x_0,\ldots,x_n)P\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\}.$$

Now consider some $(x_0,\ldots,x_n) \in \Delta_m$ with $r \le m \le 2r-1$. By assumption,

$$L(x_0,\ldots,x_n)P\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = \frac{\Theta(\epsilon^m)}{\underline{O}(\epsilon^{2m-2r})}\Theta(\epsilon^m) = O(\epsilon^{2r}).$$

Thus, since $|\Delta_m| < \infty$ for all m by Lemma 1(ii),

$$\sum_{m=r}^{2r-1} \sum_{\substack{(x_0,\ldots,x_n)\in\Delta_m \\ n>0}} L(x_0,\ldots,x_n)P\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = O(\epsilon^{2r}). \tag{7}$$

Also, using Lemma 1, we obtain

$$\begin{aligned}
&\sum_{m=2r}^{\infty} \sum_{\substack{(x_0,\ldots,x_n)\in\Delta_m \\ n>0}} L(x_0,\ldots,x_n)P\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} \\
&\le \sum_{m=2r}^{\infty} \sum_{\substack{(x_0,\ldots,x_n)\in\Delta_m \\ n>0}} \eta^{m+1}\alpha\beta^m\epsilon^m \le \sum_{m=2r}^{\infty} |S|^{(m+1)N}\eta^{m+1}\alpha\beta^m\epsilon^m \\
&= \alpha\eta|S|^N \sum_{m=2r}^{\infty} \Big(\beta\eta|S|^N\epsilon\Big)^m = \Theta(\epsilon^{2r})
\end{aligned} \tag{8}$$

as $\epsilon \to 0$. Hence, it follows from (7) and (8) that $E'[1\{\tau_F < \tau_0\}L^2] = O(\epsilon^{2r})$, and so RE' remains bounded as $\epsilon \to 0$.

Also, since $\mathcal{J} \subset \mathcal{I}$, it follows from Theorem 1 that if there exists some path $(y_0,\ldots,y_k) \in \Delta_m$, $r \le m \le 2r-1$, such that $P'(y_0,\ldots,y_k) = O(\epsilon^{2m-2r+1})$, then $RE' \to \infty$ as $\epsilon \to 0$ ■

3.3 Sufficient Conditions For Bounded Relative Error

The conditions of Theorem 2 can be potentially difficult to verify in practice because of the large number of sample paths that must be examined. However, the following result due to Shahabuddin [25] is a simple sufficient condition for bounded relative error.

Proposition 1 *Consider any system satisfying Assumptions A1–A3. Also, consider any $P' \in \mathcal{J}$, and let RE' denote the relative error of the estimator of γ obtained using P'. If $\mathbf{P}'(x,y) = \Theta(1)$ for all $(x,y) \in \Gamma$ with $x \in U$, then RE' remains bounded as $\epsilon \to 0$.*

Using a matrix-analytic approach of analysis, Shahabuddin [25] established the previous result. We now provide a simple proof using Theorem 2.

Proof. It is easy to see that if $\mathbf{P}'(x,y) = \Theta(1)$ for all $(x,y) \in \Gamma$ with $x \in U$, then $P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = \Theta(1)$ for all $(x_0,\ldots,x_n) \in \Delta$. Thus, Theorem 2 implies that RE' remains bounded as $\epsilon \to 0$. ∎

Shahabuddin [25] used Proposition 1 to prove that balanced failure biasing always yields performance measure estimators having bounded relative error and simple failure biasing does so when the system is balanced. We will now show that a large variety of importance sampling schemes satisfy the sufficient condition established in Proposition 1 when the system is balanced. To see this, we will define some additional classes of importance sampling methods. Before doing so, we make some definitions. For any state $x \in S$, let

$$F(x) = \{(x,y) \in \Gamma : y \succ x\},$$

which is the set of failure transitions from x, and we let

$$R(x) = \{(x,y) \in \Gamma : y \prec x\},$$

which is the set of repair transitions from x. Finally, we define

$$H = \{x \in U : F(x) \neq \emptyset, R(x) \neq \emptyset\},$$

which is the set of operational states from which there are both failure and repair transitions possible. Then we define the following class of probability measures.

Definition 3 $\mathcal{B}$ is the class of probability measures P' which satisfy the following properties:

(i) P' is Markovian with some transition matrix $\mathbf{P}'$;

(ii) For any state $x \in H$,

- **(a)** $\sum_{(x,y)\in F(x)} \mathbf{P}'(x,y) = \rho_0(x)$, where $\rho_0(x) = \Theta(1)$;
- **(b)** $\sum_{(x,y)\in R(x)} \mathbf{P}'(x,y) = 1 - \rho_0(x)$;

(iii) For any $(w,y) \in \Gamma$, if $\mathbf{P}(w,y) = \Theta(\epsilon^d)$, then $\mathbf{P}'(w,y) = \underline{O}(\epsilon^d)$.

Note that $\mathcal{B} \subset \mathcal{J}$, and we call any $P' \in \mathcal{B}$ a *failure biasing method*. Our definition captures the underlying principle of all of the existing failure biasing techniques, which we now describe. Under the original measure P, the probability of any failure transition from some state $x \in U$, $x \neq 0$, is $O(\epsilon)$ and the probability of a repair transition is $\Theta(1)$, as was shown in (2) and (3). The fundamental idea behind each failure biasing method is to increase the total probability to $\rho_0(x)$ of a failure transition from x, where $\rho_0(x) = \Theta(1)$ with $\rho_0(x) < 1$, Also, we decrease the

total probability of a repair transition from state x to $1-\rho_0(x)$. The various methods differ in the way they allocate the $\rho_0(x)$ and $1-\rho_0(x)$ to the individual failure and repair transitions from a state x, but they all do so in such a way that the probability of a transition under the new measure is never significantly smaller than its probability under the original measure. (This is part (iii) of Definition 3.) Also, the probabilities of transitions from states $z \in F$ are unaltered. Extensive empirical work shows that good results can be obtained by taking $\rho_0(x) = \rho_0$ for all x with $0.5 \leq \rho_0 \leq 0.9$; see Lewis and Böhm [17] and Goyal et al. [11] for further details. We describe each failure biasing method later.

In Definition 3 we have assumed that all failure biasing methods are Markovian. However, when estimating the steady-state unavailability, Conway and Goyal [4] suggest that the failure biasing should be turned off once a failed state is hit. From this point, the original probability measure is used until the system returns to the regenerative state, at which time the sample ends and we begin a new sample with the failure biasing enabled again. This technique is known as dynamic importance sampling since the importance sampling scheme depends on the sample path (and is therefore not Markovian). Because we are estimating γ in this paper, a sample ends once either a failed state or state 0 is hit. Thus, we do not use dynamic failure biasing.

We now examine what happens when certain failure biasing methods are used to estimate our performance measure when the system is balanced; i.e., $b(x,y) = 1$ for all $(x,y) \in \Gamma$ with $y \succ x$. To do this, we define the following class of importance sampling methods.

Definition 4 $\mathcal{P}$ is the class of probability measures $P' \in \mathcal{B}$ which satisfy the following properties:

(i) For all $x \in U$, there exist sets $F_k(x) \subset F(x)$, $F_k \neq \emptyset$, $k = 1, \ldots, m(x)$, such that $F(x) = \cup_{k=1}^{m(x)} F_k(x)$ and $F_j(x) \cap F_k(x) = \emptyset$ for all $j \neq k$.

(ii) If $x \in U$ and $(x,y) \in F_k(x)$, then

$$\mathbf{P}'(x,y) = \xi_k(x) \frac{\mathbf{P}(x,y)}{\sum_{(x,z)\in F_k(x)} \mathbf{P}(x,z)},$$

where $\xi_k(x) = \Theta(1)$ with $\sum_{k=1}^{m(x)} \xi_k(x) = \rho_0(x)$ if $x \in H$ and $\sum_{k=1}^{m(x)} \xi_k(x) = 1$ if $x \in U - H$.

In part (i) of Definition 4, the set of failure transitions from some opertional state x is decomposed into a number of subsets, where the number of subsets depends on the state x and the importance sampling method used. Then in part (ii), the transition probability of any failure transition under importance sampling is assigned proportionally to its original transition probability relative to some set. Thus, we call any $P' \in \mathcal{P}$ a *proportional failure*

biasing method. As we shall see in Section 3.4, the class $\mathcal{P}$ includes all of the exisiting failure biasing methods.

The following result shows that if we apply any proportional failure biasing method to a balanced system, then the resulting performance measure estimator will have bounded relative error. This result demonstrates that a large variety of importance sampling schemes satisfy the sufficient condition established in Proposition 1 when the system is balanced.

Theorem 3 *Consider any system satisfying Assumptions A1–A3. Also, consider any* $P' \in \mathcal{P}$ *and let* RE' *denote the relative error of the estimator of* γ *obtained using* P'*. If the system is balanced, then* $\mathbf{P}'(x,y) = \Theta(1)$ *for all* $(x,y) \in \Gamma$ *with* $x \in U$*, and so* RE' *remains bounded as* $\epsilon \to 0$.

Proof. Suppose $\gamma = \Theta(\epsilon^r)$ for some $r \geq 1$. Since the system is balanced, $b(x,y) = 1$ for all $(x,y) \in \Gamma$ with $y \succ x$. Thus, $\mathbf{P}(0,y) = \Theta(1)$ for all $(0,y) \in \Gamma$ by (2) and (3). Furthermore, $\mathbf{P}(x,y) = \Theta(\epsilon)$ for all $(x,y) \in F(x)$ with $x \neq 0$ by (3). Now consider any $(x,y) \in F(x)$ with $x \in U$, and suppose $(x,y) \in F_k(x)$. Because all transitions $(x,z) \in F_k(x)$ must have probabilities of the same ϵ-order under the original measure,

$$\frac{\mathbf{P}(x,y)}{\sum_{(x,z)\in F_k(x)} \mathbf{P}(x,z)} = \Theta(1),$$

which implies that $\mathbf{P}'(x,y) = \Theta(1)$ since $\xi_k(x) = \Theta(1)$. Also, all transitions $(x,y) \in R(x)$, $x \in U$, satisfy $\mathbf{P}'(x,y) = \Theta(1)$ by part (iii) of Definition 3. Hence, Proposition 1 implies that RE' remains bounded as $\epsilon \to 0$. ∎

3.4 Examples of Failure Biasing Methods

To demonstrate the usefulness of our results, we will apply them to study a number of importance sampling schemes currently in the literature that were developed to simulate highly reliable Markovian systems. Specifically, we will examine examples of failure biasing methods. For all of these techniques, $\rho_0(x) = \rho_0$ for all states x.

3.4.1 Balanced Failure Biasing

Shahabuddin [25] developed the balanced failure biasing method, which we now describe. Consider any $x \in U$. From any state $x \in H$, balanced failure biasing gives probability $\rho_0/|F(x)|$ to the individual failure transitions $(x,y) \in F(x)$ and allocates the $1-\rho_0$ to the individual repair transitions $(x,y) \in R(x)$ in proportion to their original transition probabilities. Also, for any state x for which $R(x) = \emptyset$, balanced failure biasing assigns probability $1/|F(x)|$ to the

individual failure transitions $(x, y) \in F(x)$. It does not alter the transition probabilities from any states $x \in F$. A more precise description is given by Shahabuddin [25].

It is easy to see that balanced failure biasing is a member of the class $\mathcal{B}$. Moreover, it is an element of $\mathcal{P}$ since we can decompose $F(x)$ by taking each of the $F_k(x)$ to consist of exactly one failure transition from x and then setting $\xi_k(x) = \rho_0/|F(x)|$ for all k. Thus, Theorem 3 implies that balanced failure biasing will yield performance measure estimators having bounded relative error when the system is balanced.

Shahabuddin [25] proved that balanced failure biasing always gives rise to bounded relative error for the performance measure estimator. However, let us now observe how we can use Theorem 2 to analyze this method. From the description of the transition matrix $\mathbf{P}'$ under balanced failure biasing, we see that $\mathbf{P}'(x, y) = \Theta(1)$ for all transitions $(x, y) \in \Gamma$ with $x \in U$. Thus, for any $(x_0, \ldots, x_n) \in \Delta$,

$$P'\{(X_0, \ldots, X_\tau) = (x_0, \ldots, x_n)\} = \Theta(1) \tag{9}$$

as $\epsilon \to 0$, where the exact probability depends on the path $(x_0, \ldots, x_n)$. Note that (9) holds no matter what the probability of the path is under the original measure P. Hence, balanced failure biasing satisfies the necessary and sufficient condition for bounded relative error established in Theorem 2.

To illustrate how balanced failure biasing works, we consider the following example. (We will return to this example when examining the other failure biasing methods.)

Example 1 Consider a system which has three types of components (i.e., $C = 3$), where the first two component types have a redundancy of two (i.e., $n_1 = n_2 = 2$), and the third type of component has a redundancy of one (i.e., $n_3 = 1$). Also, the components of type 1 and 2 have failure rate ϵ (i.e., $b_1 = b_2 = 1$), and the component of type 3 has failure rate ϵ^2 (i.e., $b_3 = 2$). Thus, $b_0 = 1$ and the system is unbalanced. There is a single repairperson who repairs components at rate 1 using a processor sharing discipline. For this problem, it is sufficient to define the state of the system to be $x = \langle x_1, x_2, x_3 \rangle$, where x_i is the number of failed components of type i. Initially, all components are operational, and the system is considered to be failed if and only if there is at least one component of each type failed. Thus, $F = \{\langle 1, 1, 1 \rangle, \langle 1, 2, 1 \rangle, \langle 2, 1, 1 \rangle, \langle 2, 2, 1 \rangle\}$. We assume there is no failure propagation. Figure 1 is a state diagram of this model with the arcs having the original transition probabilities, and Figure 2 is the same when using balanced failure biasing.

It is easy to show that

$$\gamma = 6\epsilon^3 + o(\epsilon^3),$$

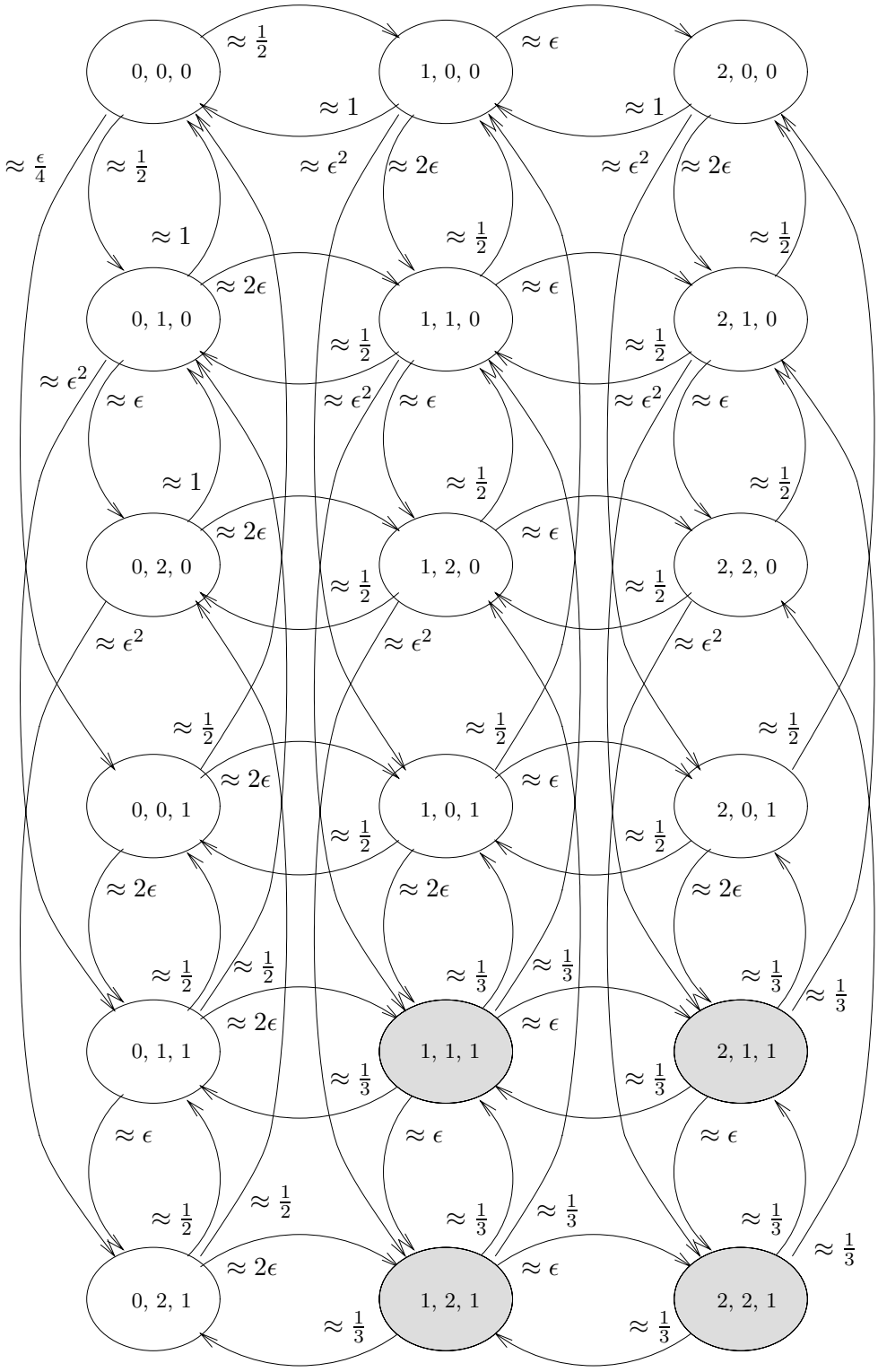

Figure 1: Transition diagram for Example 1 with original transition probabilities

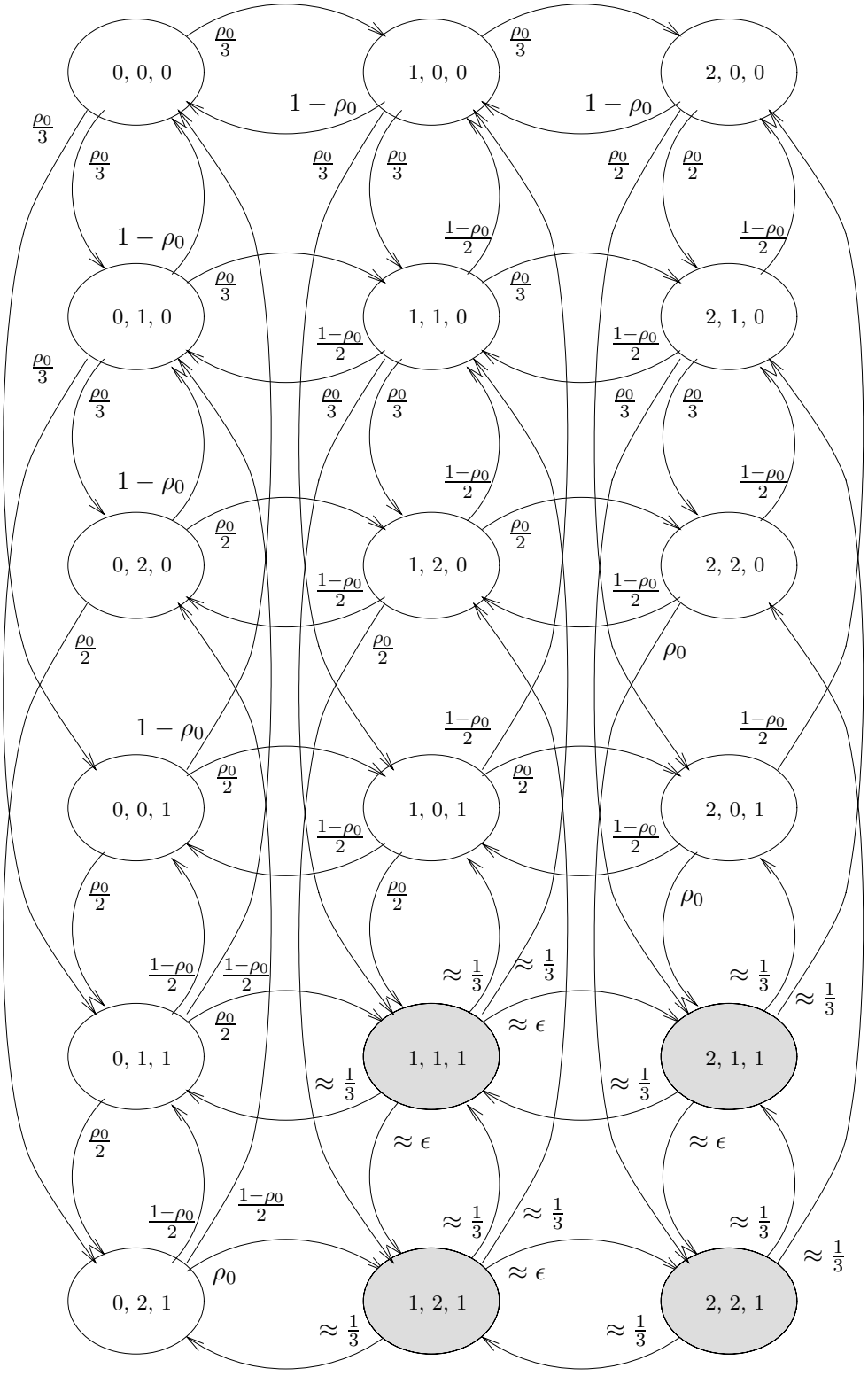

Figure 2: Transition diagram for Example 1 under balanced failure biasing

and so $r = 3$. By Theorem 2 we only need to check the paths in Δ_m, $3 \leq m \leq 5$, to determine if γ can be estimated with bounded relative error using balanced failure biasing. As we noted above in (9), all paths to failure have probability $\Theta(1)$ under balanced failure biasing. Hence, we can estimate γ with bounded relative error. In fact, we can show (after a lot of algebra) that when using balanced failure biasing,

$$\sigma'^2 = \left(\frac{114}{\rho_0^2} - 36\right)\epsilon^6 + o(\epsilon^6)$$

and so

$$RE' = \frac{\sqrt{114/\rho_0^2 - 36}}{6} + o(1),$$

which remains bounded as $\epsilon \to 0$. ●

3.4.2 Simple Failure Biasing

Lewis and Böhm [17] originally developed the simple failure biasing method, and Goyal et al. [9] and Shahabuddin et al. [28] later modified it. We now describe the method. From any state $x \in H$, we allocate the ρ_0 and $1 - \rho_0$ to the individual failure and repair transitions, respectively, in proportion to their original probabilities. We do not alter the transition probabilities from state 0 or from any state $x \in F$. In some sense the simple failure biasing method is a natural way of implementing importance sampling since it preserves the underlying structure of the system. A more precise description of simple failure biasing is given in Shahabuddin [25] and Nakayama [18].

Simple failure biasing is a member of the class $\mathcal{B}$. Moreover, it is an element of $\mathcal{P}$ since for each $x \in U$, we can let $m(x) = 1$ and $F_1(x) = F(x)$. Thus, Theorem 3 implies that simple failure biasing yields performance measure estimators having bounded relative error when the system is balanced. This result was previously established by Shahabuddin [25].

Shahabuddin [25] also showed by example that simple failure biasing may not give bounded relative error for unbalanced systems. However, Nakayama [18] demonstrated that this is not always the case by constructing an example of an unbalanced system for which simple failure biasing gives bounded relative error. Furthermore, Nakayama established a necessary and sufficient condition that characterizes when simple failure biasing will give bounded relative error in the estimation of γ. This condition is equivalent to the one given in Theorem 2 specialized to the case of simple failure biasing.

As previously mentioned, Shahabuddin [25] constructed an example showing that simple failure biasing may not give bounded relative error. This is also shown in the following example.

Example 1 (continued) Figure 3 is a state diagram of this model when using simple fail-

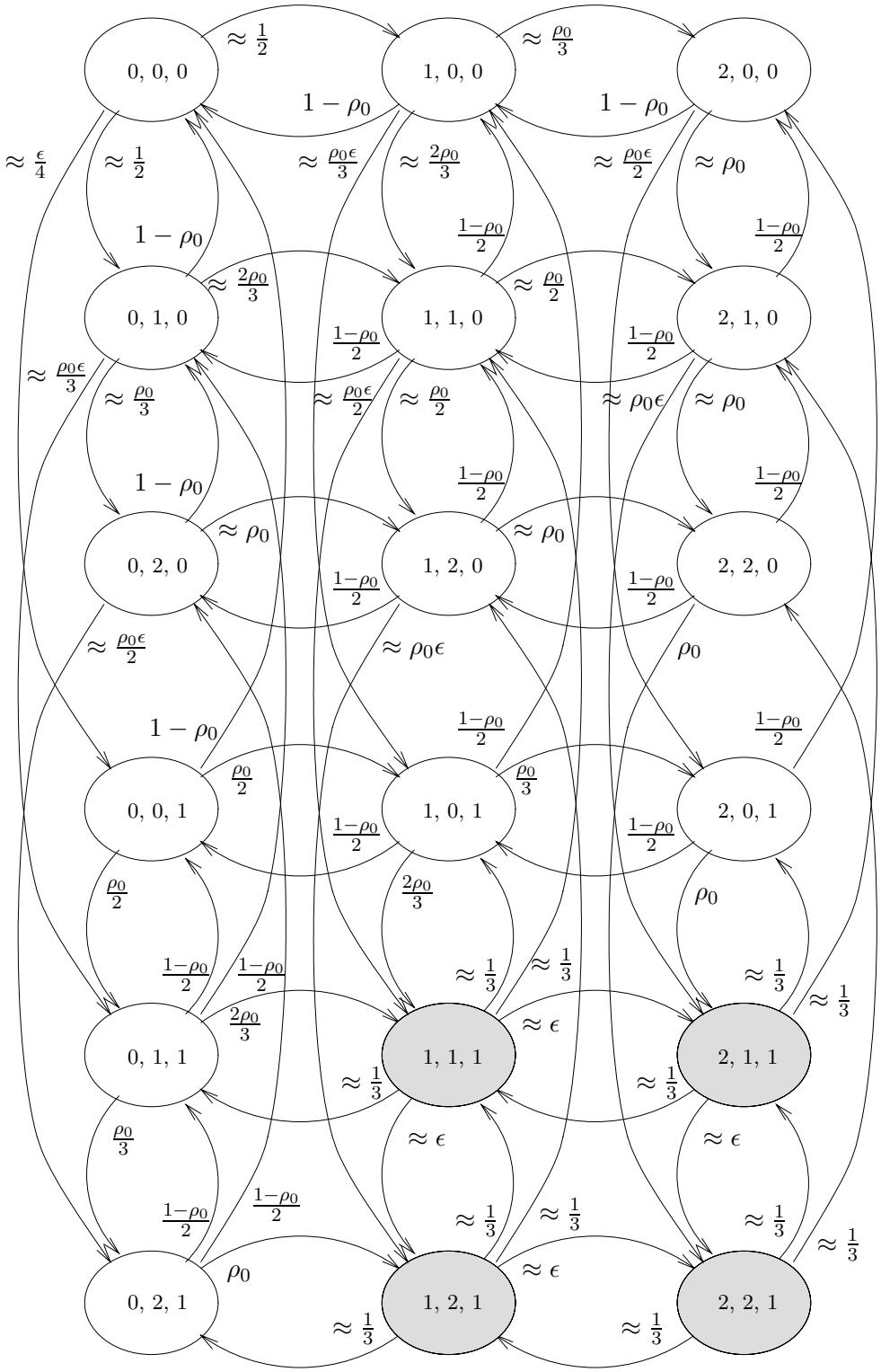

Figure 3: Transition diagram for Example 1 under simple failure biasing

ure biasing. Consider the path $(\langle 0,0,0\rangle, \langle 0,0,1\rangle, \langle 1,0,1\rangle, \langle 1,1,1\rangle) \in \Delta_3$. Since simple failure biasing does not alter the transition probabilities from the initial state, $\mathbf{P}'(\langle 0,0,0\rangle, \langle 0,0,1\rangle) = \epsilon/4 + o(\epsilon)$ under simple failure biasing. All of the other transitions in the path have probability $\Theta(1)$ under simple failure biasing, and so the the entire path has probability $\Theta(\epsilon)$ under simple failure biasing. However, Theorem 2 requires this path to have new probability $\underline{O}(1)$ for there to be bounded relative error. Therefore, the performance measure estimator in this example will not have bounded relative error if we use simple failure biasing. In fact we can show (after a lot of algebra) that under simple failure biasing,

$$\sigma'^2 = \frac{54}{\rho_0^2}\epsilon^5 + o(\epsilon^5),$$

and so

$$RE' = \frac{\sqrt{3/2}}{\rho_0}\epsilon^{-1/2} + o(\epsilon^{-1/2}) \to \infty$$

as $\epsilon \to 0$. ●

In Section 3.4.1 we saw that balanced failure biasing always results in bounded relative error. Thus, balanced failure biasing is more robust than simple failure biasing. However, as we previously mentioned, Shahabuddin [25] proved that simple failure biasing will always result in bounded relative error when the system is balanced, and Nakayama [18] showed by example that when simple failure biasing gives bounded relative error, the coefficient of the leading term in the asymptotic expansion for the variance resulting from simple failure biasing can be smaller than that from balanced failure biasing. These coefficients are important since they largely determine the expected half-width of the resulting confidence intervals when simulating real systems with small but fixed failure rates. Thus, simple failure biasing may be more appropriate than balanced failure biasing in certain contexts.

3.4.3 Bias2 Failure Biasing

Now we describe the bias2 failure biasing method of importance sampling, which was developed by Goyal et al. [11]. The basic idea of this approach is as follows. Bias2 failure biasing does not alter the transition probabilities from state 0 or from any state $x \in F$. From any state $x \in H$, the technique gives a higher combined probability ρ_1, where $\rho_1 = \Theta(1)$, to those failure transitions corresponding to component types which have at least one of their type already failed. More precisely, for each state $x \in H$, define

$$F_2(x) = \{(x,y) \in F(x) : n_i(y) < n_i(x) < n_i(0) \text{ for some component type } i\},$$

which is the set of failure transitions (x,y) that have at least one component of some type i failed in state x and at least one component of that type fails on the transition (x,y). Also,

define

$$F_1(x) = F(x) - F_2(x),$$

which is the set of the other failure transitions from x. Furthermore, for any state $x \in S$, define

$$\begin{aligned} p_{F_1}(x) &= \sum_{(x,y)\in F_1(x)} \mathbf{P}(x,z), \\ p_{F_2}(x) &= \sum_{(x,y)\in F_2(x)} \mathbf{P}(x,z), \\ p_F(x) &= \sum_{(x,y)\in F(x)} \mathbf{P}(x,z), \\ p_R(x) &= \sum_{(x,y)\in R(x)} \mathbf{P}(x,z). \end{aligned}$$

Thus, $p_{F_1}(x)$ is the total probability of taking a failure transition in $F_1(x)$ from x, $p_{F_2}(x)$ is the same for the set $F_2(x)$, $p_F(x)$ is the total probability of taking any failure transition from x, and $p_R(x)$ is the total probability of taking a repair transition from x. We construct the new transition matrix $\mathbf{P'}$ from the original transition matrix $\mathbf{P}$ using the ensuing algorithm.

(i) For $(x, y) \notin \Gamma$,

$$\mathbf{P'}(x, y) = 0.$$

(ii) For $(x, y) \in \Gamma$ and $x \in U$,

(a) With $x = 0$,

$$\mathbf{P'}(x, y) = \mathbf{P}(x, y);$$

(b) With $x \neq 0$, $F_1(x) \neq \emptyset$, and $F_2(x) \neq \emptyset$,

$$\mathbf{P'}(x,y) = \begin{cases} \rho_0\rho_1\mathbf{P}(x,y)/p_{F_2}(x) & \text{if } (x,y) \in F_2(x) \\ \rho_0(1-\rho_1)\mathbf{P}(x,y)/p_{F_1}(x) & \text{if } (x,y) \in F_1(x) \\ (1-\rho_0)\mathbf{P}(x,y)/p_R(x) & \text{if } (x,y) \in R(x) \\ 0 & \text{otherwise} \end{cases}.$$

(c) With $x \neq 0$ and either $F_1(x) = \emptyset$ or $F_2(x) = \emptyset$,

$$\mathbf{P'}(x,y) = \begin{cases} \rho_0\mathbf{P}(x,y)/p_F(x) & \text{if } (x,y) \in F(x) \\ (1-\rho_0)\mathbf{P}(x,y)/p_R(x) & \text{if } (x,y) \in R(x) \\ 0 & \text{otherwise} \end{cases}.$$

(iii) For $(x, y) \in \Gamma$ and $x \in F$,

$$\mathbf{P'}(x, y) = \mathbf{P}(x, y).$$

Extensive empirical work suggests that ρ_1 should be chosen such that $0.5 \leq \rho_1 \leq 0.9$; see Goyal et al. [11] for further details.

Bias2 failure biasing is a member of the class $\mathcal{B}$. Moreover, it is an element of $\mathcal{P}$, which can be seen as follows. For each $x \in U$, define $F_1(x)$ and $F_2(x)$ as above. If $F_1(x) \neq \emptyset$ and $F_2(x) \neq \emptyset$, then let $\xi_1(x) = \rho_0(1-\rho_1)$ and $\xi_2(x) = \rho_0\rho_1$. If $F_2(x) = \emptyset$, then let $\xi_1(x) = \rho_0$, and $\xi_2(x) = \rho_0$ if $F_1(x) = \emptyset$. Thus, Theorem 3 implies that bias2 failure biasing yields performance measure estimators having bounded relative error when the system is balanced. Now let us examine what happens when we apply bias2 failure biasing to our previous example.

Example 1 (continued) Figure 4 is a state diagram of this model when using bias2 failure biasing. Consider the path $(\langle 0,0,0\rangle, \langle 0,0,1\rangle, \langle 1,0,1\rangle, \langle 1,1,1\rangle) \in \Delta_3$. Since bias2 failure biasing does not alter the transition probabilities from the initial state, $\mathbf{P'}(\langle 0,0,0\rangle, \langle 0,0,1\rangle) = \epsilon/4 + o(\epsilon)$ under bias2 failure biasing. All of the other transitions in the path have probability $\Theta(1)$ under bias2 failure biasing, and so the the entire path has probability $\Theta(\epsilon)$ under bias2 failure biasing. However, Theorem 2 require this path to have new probability $\underline{O}(1)$ for there to be bounded relative error. Therefore, the estimator of γ in this example will not have bounded relative error if we use bias2 failure biasing. In fact we can show (after a lot of algebra) that under bias2 failure biasing,

$$\sigma'^2 = \frac{24}{\rho_0^2(1-\rho_1)^2}\epsilon^5 + o(\epsilon^5),$$

and so

$$RE' = \frac{\sqrt{2/3}}{\rho_0(1-\rho_1)}\epsilon^{-1/2} + o(\epsilon^{-1/2}) \to \infty$$

as $\epsilon \to 0$. ●

Since bias2 failure biasing may not result in bounded relative error for unbalanced systems whereas balanced failure biasing always does, balanced failure biasing is more robust. However, Goyal et al. [11] showed empirically that bias2 failure biasing can give better results than balanced failure biasing when simulating balanced systems.

We can modify the bias2 failure biasing method so that it will always yield bounded relative error as follows. Change step (ii) of the above algorithm to

(ii') For $(x,y) \in \Gamma$ and $x \in U$,

(a) With $x = 0$,

$$\mathbf{P'}(x,y) = \mathbf{P}(x,y);$$

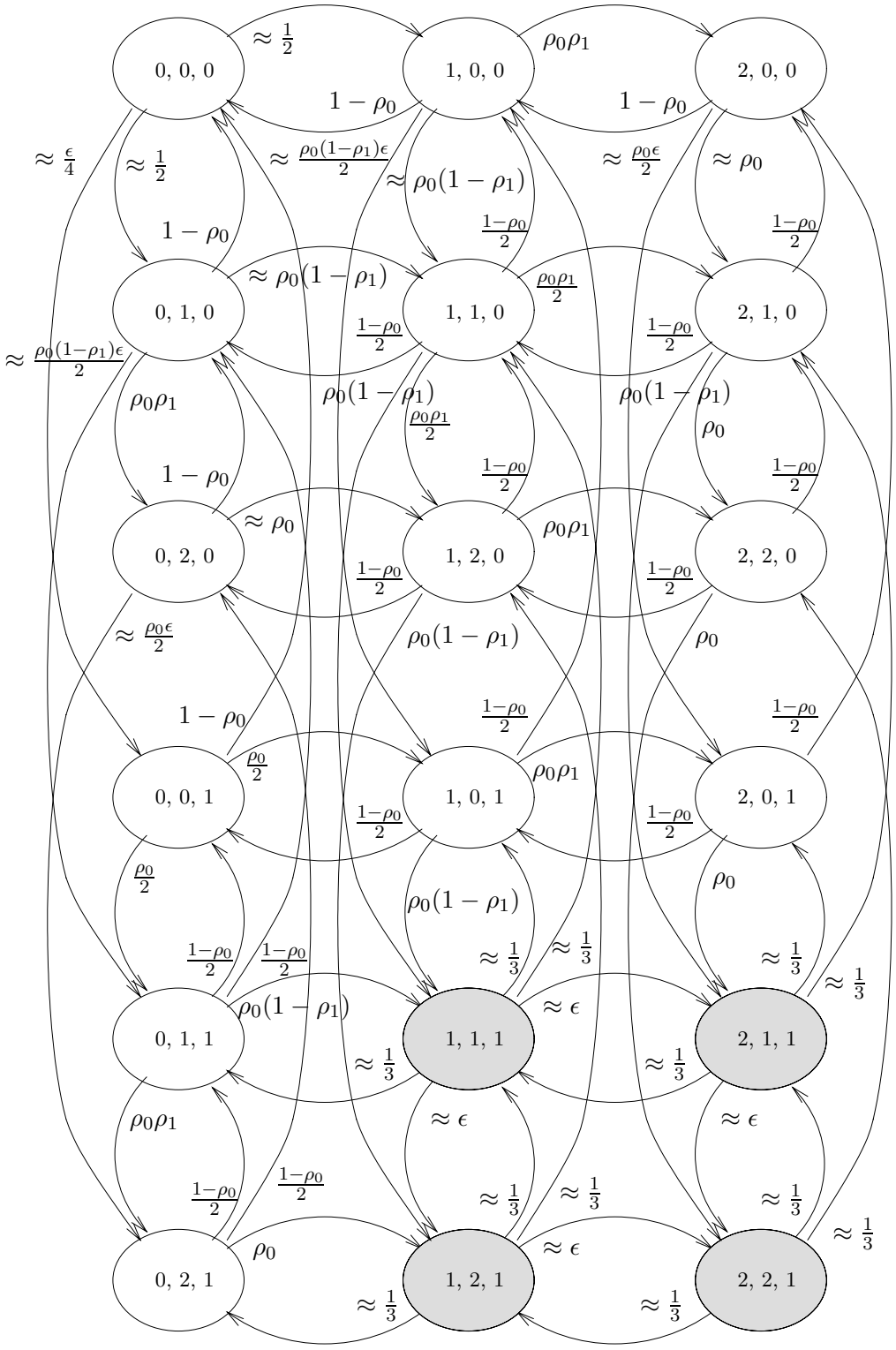

Figure 4: Transition diagram for Example 1 under bias2 failure biasing

(b) With $x \neq 0$, $F_1(x) \neq \emptyset$, and $F_2(x) \neq \emptyset$,

$$\mathbf{P}'(x,y) = \begin{cases} \rho_0\rho_1/|F_2(x)| & \text{if } (x,y) \in F_2(x) \\ \rho_0(1-\rho_1)/|F_1(x)| & \text{if } (x,y) \in F_1(x) \\ (1-\rho_0)\mathbf{P}(x,y)/p_R(x) & \text{if } (x,y) \in R(x) \\ 0 & \text{otherwise} \end{cases}.$$

(c) With $x \neq 0$ and either $F_1(x) = \emptyset$ or $F_2(x) = \emptyset$,

$$\mathbf{P}'(x,y) = \begin{cases} \rho_0/|F_0(x)| & \text{if } (x,y) \in F(x) \\ (1-\rho_0)\mathbf{P}(x,y)/p_R(x) & \text{if } (x,y) \in R(x) \\ 0 & \text{otherwise} \end{cases}.$$

We call the new resulting method bias2 balanced failure biasing, and it is easy to show that Theorem 2 implies that it will always give rise to bounded relative error.

3.4.4 Failure Distance Biasing

Failure distance biasing, developed by Carrasco [1, 2], is an importance sampling scheme which falls into the class of failure biasing methods. To describe the technique, we need some definitions. For any state $x \in U$, define the *failure distance* as

$$d(x) = \min_{\substack{y \succ x, \\ y \in F}} \left(\sum_{i=1}^{C} n_i(x) - n_i(y) \right),$$

which is the minimum number of failing components whose failure in x would take the system down. Also, define $d(x) = 0$ for all $x \in F$. Now consider any failure transition (x, y) with $x \in U$. Then we say that (x, y) is *dominant* if $d(y) < d(x)$, and *non-dominant* otherwise. Also, (x, y) is *critical* if $d(y) < d(x) - 1$. The *criticality* of (x, y) is defined to be $c(x, y) = d(x) - d(y)$. Now fix ρ_d and ρ_c, where $\rho_d = \Theta(1)$ and $\rho_c = \Theta(1)$. (Carrasco [1] suggests setting $\rho_0 = 0.8$, $\rho_d = 0.7$, and $\rho_c = 0.2$.) Then we construct the transition matrix $\mathbf{P}'$ for failure distance biasing as follows.

We do not alter any of the transition probabilities from states $x \in F$. Also, we give probability 0 to any transition $(x, y) \notin \Gamma$ under failure distance biasing. Now consider any state $x \in U$. The algorithm contains a number of steps in which some set of feasible transitions $(x, y) \in \Gamma$ is divided into two sets, where we allocate some probability to the transitions in one set and pass the other set on to the next step. We skip a step if one of the two sets is empty. In the first step, the set of feasible transitions $(x, y) \in \Gamma$ is divided into the set of failure transitions, $F(x)$, and the set of repair transitions, $R(x)$. If $R(x) \neq \emptyset$, then we give the set $R(x)$ a total probability of $1 - \rho_0$, where the $1 - \rho_0$ is allocated to the various repair transitions

in proportion to the original transition probabilities. If $R(x) \neq \emptyset$, then we give the set $F(x)$ a total (conditional) probability of ρ_0 and pass it on to the next step. We then divide the set $F(x)$ into the set of dominant transitions and the set of non-dominant transitions. We assign a (conditional) probability of $(1-\rho_d)$ to the set of non-dominant transitions (given that we are in the set $F(x)$), which we allocate to the individual transitions in proportion to their original transition probabilities. We give the set of dominant transitions a (conditional) probability of ρ_d (given that we are in the set $F(x)$) and pass this set on to the next step. If the set of dominant transitions is composed of transitions having different criticalities, then we divide the set of dominant transitions from x into two sets: one containing all of the transitions with the smallest criticality and the other having all of the other transitions. The set consisting of the transitions with the smallest criticalities is assigned a (conditional) probability of $(1-\rho_c)$ (given that we are in the set of dominant transitions), where the individual transitions in the set are allocated probabilities in proportion to their original transition probabilities. The other set is allotted a (conditional) probability of ρ_c (given that we are in the set of dominant transitions) and is passed on to the next step. We repeat the last step as long as the remaining set contains transitions having different criticalities. The process ends once all of the transitions in the remaining set have the same criticality. To illustrate this, consider the following example from Carrasco [1]. Suppose from state x there are a number of repair transitions and some failure transitions having criticalities 1, 2, and 3. Then distance failure biasing assigns probabilities $1-\rho_0$, $\rho_0(1-\rho_c)$, $\rho_0\rho_c(1-\rho_c)$, and $\rho_0\rho_c^2$ to the respective sets.

We can easily show that failure distance biasing is a member of both class $\mathcal{B}$ and class $\mathcal{P}$. Thus, Theorem 3 implies that it will yield performance measure estimators having bounded relative error when the system is balanced. Now we apply failure distance biasing to our previous example.

Example 1 (continued) Figure 5 is a state diagram of this model when using failure distance biasing. Note that $d(\langle 0,0,0\rangle) = 3$. Also, all states with exactly one component failed (i.e., states $\langle 1,0,0\rangle$, $\langle 0,1,0\rangle$, and $\langle 0,0,1\rangle$) have a failure distance of 2. Thus, all of the failure transitions from state $\langle 0,0,0\rangle$ have a criticality of 1. Since there are no repair transtions from state $\langle 0,0,0\rangle$, the probabilities of all failure transitions from state $\langle 0,0,0\rangle$ under failure distance biasing are the same as under the original measure.

Now consider the path $(\langle 0,0,0\rangle, \langle 0,0,1\rangle, \langle 1,0,1\rangle, \langle 1,1,1\rangle) \in \Delta_3$. Under failure distance biasing, the first transition of the path has probability $\mathbf{P'}(\langle 0,0,0\rangle, \langle 0,0,1\rangle) = \mathbf{P}(\langle 0,0,0\rangle, \langle 0,0,1\rangle) = \epsilon/4 + o(\epsilon)$. Since each other transition has probability at most $\Theta(1)$, the entire path has probability

$$P'\{(X_0, \ldots, X_{\tau_F}) = (\langle 0,0,0\rangle, \langle 0,0,1\rangle, \langle 1,0,1\rangle, \langle 1,1,1\rangle)\} = O(\epsilon)$$

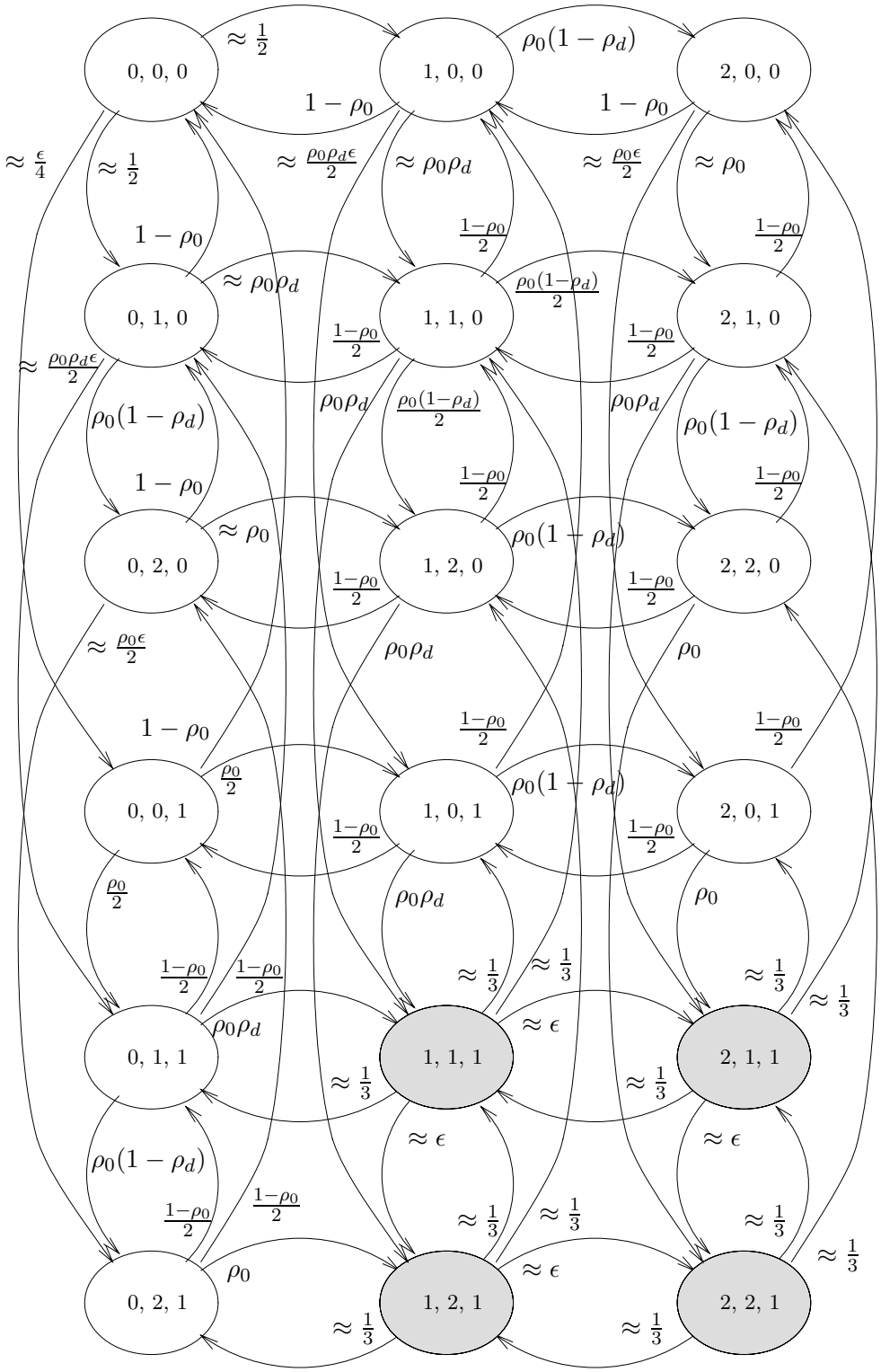

Figure 5: Transition diagram for Example 1 under failure distance biasing

under failure distance biasing. However, Theorem 2 requires this path to have probability $\Theta(1)$ under failure distance biasing to achieve bounded relative error. Therefore, the performance measure estimator in this example will not have bounded relative error if we use failure distance biasing. In fact we can show (after a lot of algebra) that under failure distance biasing,

$$\sigma'^2 = \frac{24}{\rho_0^2 \rho_d^2}\epsilon^5 + o(\epsilon^5)$$

and

$$RE' = \frac{\sqrt{2/3}}{\rho_0 \rho_d}\epsilon^{-1/2} + o(\epsilon^{-1/2}) \to \infty$$

as $\epsilon \to 0$. ●

Using the same type of modifications done to the bias2 failure biasing method to obtain the bias2 balanced failure biasing method, we can modify the failure distance biasing method so that it will always yield bounded relative error. We call the resulting algorithm balanced failure distance biasing and do not present it explicitly.

4 Estimating Derivatives Using Importance Sampling

We now examine the behavior of estimates of derivatives of γ with respect to component failure rates obtained using importance sampling. Our goal is to prove results similar to those established in the previous section for likelihood ratio derivative estimators.

4.1 Additional System Structure

To obtain results on likelihood ratio derivative estimators, we will assume that our system has more structure. In particular, we will limit the generality of failure propagation and disallow state dependent failure rates for the components. The following modifications were developed by Nakayama [18] mainly to simplify the calculations.

First, we no longer allow components to have state dependent failure rates. Thus, the failure rate of components of type i is λ_i, regardless of the state of the system. When parameterizing by ϵ, we have that $\lambda_i = \lambda_i(\epsilon) = \tilde{\lambda}_i \epsilon^{b_i}$, where $b_i \geq 1$ is integer-valued, $\tilde{\lambda}_i > 0$, and b_i and $\tilde{\lambda}_i$ do not depend on the state of the system. This implies that

$$b_0 = \min_{1 \leq i \leq C} b_i.$$

We restrict the generality of failure propagation as follows. First, we no longer permit the $p(\,\cdot\,; x, i)$ to depend on ϵ. Also, we assume that the following hold:

A4 *If* $p(y; x, i) > 0$ *and* $p(y; x, j) > 0$, *then* $b_i = b_j$.

A5 *If there exists a component type i such that $b_i = b_0$ and $p(y;0,i) > 0$, then there exists another component type $j \neq i$ such that $b_j = b_0$ and $p(y;0,j) \neq p(y;0,i)$.*

Assumption A4 stipulates that if there is a failure transition which can be triggered by the failure of two different types of components, then the failure rates of the two component types must be of the same order of magnitude. This assumption enables us to determine the order of magnitude of the transition rate of any failure transition. More specifically, Assumption A4 implies that the transition rate of any $(x,y) \in \Gamma$ satisfies

$$q(x,y) = \begin{cases} c(x,y)\epsilon^{d(x,y)} & \text{if } y \succ x \\ \mu(x,y) & \text{if } y \prec x \\ 0 & \text{otherwise} \end{cases},$$

where $c(x,y) > 0$, $d(x,y) \geq 1$ are integer-valued, $\epsilon > 0$, and $\mu(x,y) > 0$. Note that the transition rates for failure transitions consist of a single term rather than a sum as before. We use Assumption A4 to determine the order of magnitude of the derivatives; see the proof of Lemma 11 of Nakayama [19] for further details.

Assumption A5 requires that if there is some component type i which has one of the largest failure rates and a failure of a component of this type can trigger a transition from state 0 to some state y, then there must exist some other component type j which can trigger the same transition but with a different probability. Note that A5 holds when the component type j in the assumption satisfies $b_j = b_0$ and $p(y;0,j) = 0$. This condition may not be unreasonable when considering large systems. We use Assumption A5 to ensure that there is no concellation when calculating certain quantities associated with the derivatives; see the proof of Lemma 11 of Nakayama [19] for further details.

If there is no failure propagation, then Assumption A5 is automatically satisfied and A4 reduces to requiring there to be at least two different component types having failure rates of the largest order ϵ^{b_0}; i.e., there exists i and j such that $i \neq j$ and $b_i = b_j = b_0$. Also, a similar situation occurs if we limit the generality of failure propagation by restricting that each failure transition can only be triggered by the failure of a single type of component; i.e., for each $(x,y) \in \Gamma$ with $y \succ x$, there exists only one component type i for which $p(y;x,i) > 0$.

We now define some more notation. For each component type i, let

$$\tau_i = \inf\{k > 0 : n_i(X_{k-1})p(X_k; X_{k-1}, i) > 0,\ X_k \succ X_{k-1}\},$$

which is the first failure transition of the DTMC X that may have been triggered by a failure of a component of type i. Note that we can decompose the event $\{\tau_F < \tau_0\}$ as

$$\{\tau_F < \tau_0\} = \{\tau_i \leq \tau_F < \tau_0\} \cup \{\tau_F < \min\{\tau_0, \tau_i\}\}, \tag{10}$$

where the two subsets are disjoint. Nakayama [19] showed that there exist $r_i \geq r$ and $\bar{r}_i \geq r$ such that

$$P\{\tau_i \leq \tau_F < \tau_0\} = \Theta(\epsilon^{r_i}) \tag{11}$$

and

$$P\{\tau_F < \min\{\tau_i, \tau_0\}\} = \Theta(\epsilon^{\bar{r}_i}), \tag{12}$$

where r is defined in (5). If $P\{\tau_F < \min\{\tau_0, \tau_i\}\} = 0$, then we define $\bar{r}_i = \infty$. Also, it is easy to see that (10) implies that

$$\min\{r_i, \bar{r}_i\} = r. \tag{13}$$

For each component type i, we let

$$\begin{aligned} \Delta^i &= \{(x_0, \ldots, x_n) \in \Delta : n \geq 1,\ n_i(x_k)p(x_{k+1}; x_k, i) > 0 \\ &\qquad \text{for some } 0 \leq k < n \text{ such that } x_{k+1} \succ x_k\} \end{aligned}$$

be the set of paths to system failure in which at least one of the failure transitions along the path could have been triggered by a failure of a component of type i; i.e., the set of paths for which $\tau_i \leq \tau_F < \tau_0$. Similarly, we define

$$\begin{aligned} \bar{\Delta}^i &= \{(x_0, \ldots, x_n) \in \Delta : n \geq 1,\ n_i(x_k)p(x_{k+1}; x_k, i) = 0 \\ &\qquad \text{for all } 0 \leq k < n \text{ such that } x_{k+1} \succ x_k\}, \end{aligned}$$

which is the set of paths to system failure in which none of the failure transitions along the path could have been triggered by a failure of a component of type i; i.e., the set of paths for which $\tau_F < \min\{\tau_i, \tau_0\}$. Furthermore, let $\Delta^i_m = \Delta^i \cap \Delta_m$ and $\bar{\Delta}^i_m = \bar{\Delta}^i \cap \Delta_m$, and note that $\Delta^i = \cup_{m=r_i}^{\infty} \Delta^i_m$ and $\bar{\Delta}^i = \cup_{m=\bar{r}_i}^{\infty} \bar{\Delta}^i_m$, where r_i and $\bar{r}_i$ are as defined in (11) and (12), respectively.

4.2 A General Necessary Condition For Bounded Relative Error For Derivative Estimators

We now analyze likelihood ratio derivative estimators of the partial derivative of γ with respect to the failure rate of component type i obtained using importance sampling. Throughout the rest of the paper, we will employ the notation $\partial_i A(\lambda_1, \ldots, \lambda_C) = \frac{\partial}{\partial \lambda_i} A(\lambda_1, \ldots, \lambda_C)$ for some function $A(\lambda_1, \ldots, \lambda_C)$.

Using the likelihood ratio method of estimating derivatives, we obtain

$$\partial_i \gamma = E[1\{\tau_F < \tau_0\} D_i],$$

where

$$D_i = \sum_{k=0}^{\tau_{\min}-1} \frac{\partial_i P(X_k, X_{k+1})}{P(X_k, X_{k+1})}.$$

See Glynn [7], Reiman and Weiss [24], and Nakayama et al. [20] for further details. The following result gives an expression in terms of ϵ for the partial derivative with respect to the failure rate of component type i.

Proposition 2 *Consider any system satisfying Assumptions A1–A5. For all ϵ sufficiently small,*

$$\partial_i \gamma = \Theta(\epsilon^{\min\{r_i - b_i,\ \bar{r}_i - b_0\}}),$$

where r_i and $\bar{r}_i$ are as defined in (11) and (12), respectively.

See Nakayama [19] for the proof. In general, we cannot say whether $r_i - b_i \leq \bar{r}_i - b_0$ or $r_i - b_i > \bar{r}_i - b_0$.

We can apply importance sampling to estimate the derivatives with respect to λ_i. To do so, we first define the following class of probability measures.

Definition 5 $\mathcal{I}'$ is the class of probability measures P' defined on $(\Omega, \mathcal{F})$ such that $P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} > 0$ for all $(x_0, \ldots, x_n) \in \Delta$ with $D_i(x_0, \ldots, x_n) \neq 0$.

The class $\mathcal{I}'$ is the set of valid importance sampling measures for estimating $\partial_i \gamma$. Note that Definition 5 places minimal restrictions on the structure of any $P' \in \mathcal{I}'$. In particular, it is not assumed that $P' \in \mathcal{I}'$ is Markovian.

For any $P' \in \mathcal{I}'$,

$$\begin{aligned}
\partial_i \gamma &= E[1\{\tau_F < \tau_0\} D_i] = \sum_{(x_0,\ldots,x_n)\in\Delta} P\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} \\
&= \sum_{(x_0,\ldots,x_n)\in\Delta} D_i(x_0, \ldots, x_n) \frac{P\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\}}{P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\}} P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} \\
&= \sum_{(x_0,\ldots,x_n)\in\Delta} D_i(x_0, \ldots, x_n) L(x_0, \ldots, x_n) P'\{(X_0, \ldots, X_{\tau_F}) = (x_0, \ldots, x_n)\} \\
&= E'[1\{\tau_F < \tau_0\} D_i L].
\end{aligned}$$

We obtain an importance sampling estimator of the derivative as follows. Generate i.i.d. samples $(\tilde{I}_1, \tilde{D}_1, \tilde{L}_1), \ldots, (\tilde{I}_n, \tilde{D}_n, \tilde{L}_n)$ of $(1\{\tau_F < \tau_0\}, D_i, L)$ using the probability measure P'. We form the point estimate

$$\tilde{\partial}_i \gamma(n) = \frac{1}{n} \sum_{k=1}^{n} \tilde{I}_k \tilde{D}_i \tilde{L},$$

and the variance of $1\{\tau_F < \tau_0\}D_i L$ under the measure P' is

$$\sigma_i'^2 = E'[1\{\tau_F < \tau_0\}D_i^2 L^2] - (\partial_i \gamma)^2.$$

Note that

$$\begin{aligned}
& E'[1\{\tau_F < \tau_0\}D_i^2 L^2] \\
&= \sum_{(x_0,\ldots,x_n)\in\Delta} D_i^2(x_0,\ldots,x_n)L^2(x_0,\ldots,x_n)P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} \\
&= \sum_{(x_0,\ldots,x_n)\in\Delta} D_i^2(x_0,\ldots,x_n)L(x_0,\ldots,x_n)P\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} \\
&= E[1\{\tau_F < \tau_0\}D_i^2 L], \qquad (14)
\end{aligned}$$

and so we can compute the second moment of the derivative estimator under importance sampling in terms of the original probability measure.

Now we establish a necessary condition analogous to Theorem 1 for the derivative estimators to have bounded relative error.

Theorem 4 *Consider any system satisfying Assumptions A1–A5. Also, consider any $P' \in \mathcal{I}'$, and let RE_i' denote the relative error of the estimator of $\partial_i\gamma$ obtained using P'. Suppose $\partial_i\gamma = \Theta(\epsilon^{\min\{r_i-b_i,\ \bar{r}_i-b_0\}})$ for some $r_i \geq r$ and $\bar{r}_i \geq r$. If RE_i' remains bounded as $\epsilon \to 0$, then the following hold:*

- **(i)** *If $r_i - b_i \leq \bar{r}_i - b_0$, then*
  - **(a)** *$P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = \underline{O}(\epsilon^{2m-2r_i})$ for all $(x_0,\ldots,x_n) \in \Delta_m^i$, $m \geq r_i$; and*
  - **(b)** *$P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = \underline{O}(\epsilon^{2m-2r_i-2b_0+2b_i})$ for all $(x_0,\ldots,x_n) \in \bar{\Delta}_m^i$, $m \geq \bar{r}_i$;*
- **(ii)** *If $r_i - b_i > \bar{r}_i - b_0$, then*
  - **(a)** *$P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = \underline{O}(\epsilon^{2m-2\bar{r}_i+2b_0-2b_i})$ for all $(x_0,\ldots,x_n) \in \Delta_m^i$, $m \geq r_i$; and*
  - **(b)** *$P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = \underline{O}(\epsilon^{2m-2\bar{r}_i})$ for all $(x_0,\ldots,x_n) \in \bar{\Delta}_m^i$, $m \geq \bar{r}_i$.*

To prove this result, we will need the following lemma, which establishes the order of magnitude of the derivative of the likelihood ratio. See Nakayama [18] for the proof.

Lemma 2 *Consider any system satisfying Assumptions A1–A5. Also, consider $(x_0,\ldots,x_n) \in \Delta_m$, where $n > 0$ and $m \geq r$. Then there exists a constant ϕ which is independent of $(x_0,\ldots,x_n)$ and ϵ such that for all $\epsilon > 0$ sufficiently small,*

(i) $D_i(x_0,\ldots,x_n) = \Theta(\epsilon^{-b_i})$ *and* $|D_i(x_0,\ldots,x_n)| \le (m+1)\phi\epsilon^{-b_i}$ *if* $(x_0,\ldots,x_n) \in \Delta_m^i$;

(ii) $D_i(x_0,\ldots,x_n) = \Theta(\epsilon^{-b_0})$ *and* $|D_i(x_0,\ldots,x_n)| \le (m+1)\phi\epsilon^{-b_0}$ *if* $(x_0,\ldots,x_n) \in \bar{\Delta}_m^i$.

Now we are in a position to establish Theorem 4.

Proof of Theorem 4. First assume $r_i - b_i \le \bar{r}_i - b_0$. Then Proposition 2 implies that $\partial_i\gamma = \Theta(\epsilon^{r_i-b_i})$. Now suppose there exists some path $(y_0,\ldots,y_k) \in \Delta_m^i$ with $m \ge r_i$ such that $P'\{(X_0,\ldots,X_{\tau_F}) = (y_0,\ldots,y_k)\} = O(\epsilon^{2m-2r_i+1})$. By Lemma 2, $D_i^2(y_0,\ldots,y_k) = \Theta(\epsilon^{-2b_i})$, and so (14) implies that

$$\begin{aligned}
&E'[1\{\tau_F < \tau_0\}D_i^2L^2] \\
&= \sum_{\substack{(x_0,\ldots,x_n)\in\Delta \\ n>0}} D_i^2(x_0,\ldots,x_n)L(x_0,\ldots,x_n)P\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} \\
&\ge D_i^2(y_0,\ldots,y_k)L(y_0,\ldots,y_k)P\{(X_0,\ldots,X_{\tau_F}) = (y_0,\ldots,y_k)\} \\
&= \Theta(\epsilon^{-2b_i})\frac{\Theta(\epsilon^m)}{O(\epsilon^{2m-2r_i+1})}\Theta(\epsilon^m) \;=\; \underline{O}(\epsilon^{2r_i-2b_i-1}).
\end{aligned}$$

Hence, $\sigma_i'^2 = \underline{O}(\epsilon^{2r_i-2b_i-1})$, and it follows that $RE_i' \to \infty$ as $\epsilon \to 0$ since $\partial_i\gamma = \Theta(\epsilon^{r_i-b_i})$. Similarly, we can show $RE_i' \to \infty$ as $\epsilon \to 0$ if any of the other conditions are violated. ∎

Theorem 4 clearly illustrates how the behavior of every path to failure affects the relative error of the derivative estimator. Furthermore, Theorem 4 made no assumptions about the structure of P' other than it must be a valid importance sampling measure. In particular, it was not assumed that the importance sampling scheme is Markovian, even though the original measure P is.

4.3 A Necessary and Sufficient Condition For Bounded Relative Error for Derivative Estimators

To establish a necessary and sufficient condition for likelihood ratio derivative estimators to have bounded relative error, we must add more structure to the importance sampling schemes considered. As in Section 3.4, we will be able to obtain stronger results by considering the class $\mathcal{J}$.

Theorem 5 *Consider any system satisfying Assumptions A1–A5. Also, consider any* $P' \in \mathcal{J}$, *and let* RE_i' *denote the relative error of the estimator of* $\partial_i\gamma$ *obtained using* P'. *Suppose* $\partial_i\gamma = \Theta(\epsilon^{\min\{r_i-b_i,\ \bar{r}_i-b_0\}})$ *for some* $r_i \ge r$ *and* $\bar{r}_i \ge r$. *Then,* RE_i' *remains bounded as* $\epsilon \to 0$ *if and only if the following hold:*

(i) *If* $r_i - b_i \le \bar{r}_i - b_0$, *then*

(a) $P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = \underline{O}(\epsilon^{2m-2r_i})$ *for all* $(x_0,\ldots,x_n) \in \Delta_m^i$, $r_i \leq m \leq 2r_i - 1$*; and*

(b) $P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = \underline{O}(\epsilon^{2m-2r_i-2b_0+2b_i})$ *for all* $(x_0,\ldots,x_n) \in \bar{\Delta}_m^i$, $\bar{r}_i \leq m \leq 2r_i - 2b_i + 2b_0 - 1$.

(ii) *If* $r_i - b_i > \bar{r}_i - b_0$*, then*

(a) $P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = \underline{O}(\epsilon^{2m-2\bar{r}_i+2b_0-2b_i})$ *for all* $(x_0,\ldots,x_n) \in \Delta_m^i$, $r_i \leq m \leq 2\bar{r}_i + 2b_i - 2b_0 - 1$*; and*

(b) $P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = \underline{O}(\epsilon^{2m-2\bar{r}_i})$ *for all* $(x_0,\ldots,x_n) \in \bar{\Delta}_m^i$, $\bar{r}_i \leq m \leq 2\bar{r}_i - 1$.

The proof of Theorem 5 is given in the appendix.

We now discuss Theorem 5. First, we compare it to Theorem 4. Suppose that $r_i - b_i \leq \bar{r}_i - b_0$, and consider $(x_0,\ldots,x_n) \in \Delta_m^i$ for some $m \geq r_i$. To apply Theorem 4, we must consider all $m \geq r_i$, whereas in Theorem 5, we only need to examine $r_i \leq m \leq 2r_i - 1$. The reason for the difference is exactly the same as that for the contrast between Theorems 1 and 2; see the discussion after the statement of Theorem 2. Also, Nakayama [18] established conditions analogous to those in Theorem 5 for the special case of simple failure biasing, and Nakayama [19] proved directly that balanced failure biasing always gives bounded relative error for the estimator of $\partial_i\gamma$. (We will also examine how Theorem 5 applies to these two methods in the following sections.) Furthermore, Nakayama [18] constructed examples demonstrating that in the case of simple failure biasing, the conditions in Theorem 2 do not imply the conditions in Theorem 5, and the converse does not hold as well. In particular, it was shown that when using simple failure biasing, it is possible to estimate a derivative more efficiently than the performance measure. This illustrates the need for developing the Theorem 5.

4.4 Sufficient Conditions For Bounded Relative Error for Derivative Estimators

The conditions of Theorem 5 can be potentially difficult to verify in practice because of the large number of sample paths that must be examined. However, the following result is a simple sufficient condition for bounded relative error for the derivative estimators.

Proposition 3 *Consider any system satisfying Assumptions A1–A3. Also, consider any* $P' \in \mathcal{J}$*, and let* RE_i' *denote the relative error of the estimator of* $\partial_i\gamma$ *obtained using* P'*. If* $\mathbf{P}'(x,y) = \Theta(1)$ *for all* $(x,y) \in \Gamma$*, then* RE_i' *remains bounded as* $\epsilon \to 0$.

Proof. It is easy to see that if $\mathbf{P}'(x,y) = \Theta(1)$ for all $(x,y) \in \Gamma$, then $P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = \Theta(1)$ for all $(x_0,\ldots,x_n) \in \Delta$. Thus, Theorem 5 implies that RE' remains bounded as $\epsilon \to 0$. ∎

The following result shows that if we apply any proportional failure biasing method to a balanced system, then the resulting derivative estimator will have bounded relative error. We omit the proof as the result can be established in exactly the same Theorem 3 was proved.

Theorem 6 *Consider any system satisfying Assumptions A1–A3. Also, consider any $P' \in \mathcal{P}$ and let RE'_i denote the relative error of the estimator of $\partial_i\gamma$ obtained using P'. If the system is balanced, then RE'_i remains bounded as $\epsilon \to 0$.*

4.5 Examples of Failure Biasing Methods

Now we examine how our results apply to the different failure biasing methods discussed in Sections 3.4.1–3.4.4.

4.5.1 Balanced Failure Biasing

Nakayama [19] proved directly that balanced failure biasing always gives rise to bounded relative error for the estimator of the derivative with respect to any component failure rate. However, let us now observe how Theorem 5 applies in this situation. As we previously established in (9), when using balanced failure biasing, $P'\{(X_0,\ldots,X_\tau) = (x_0,\ldots,x_n)\} = \Theta(1)$ as $\epsilon \to 0$ for any $(x_0,\ldots,x_n) \in \Delta$, where the exact probability depends on the path $(x_0,\ldots,x_n)$. This holds no matter what the probability of the path is under the original measure P. Hence, balanced failure biasing satisfies the necessary and sufficient condition for estimating $\partial_i\gamma$ with bounded relative error established in Theorem 5.

Let us now investigate what happens when balanced failure biasing is used to estimate derivatives in our previous example.

Example 1 (continued) Recall that Figure 2 is the transition probability diagram under balanced failure biasing. We can show (after a lot of algebra) that

$$\begin{aligned} \partial_1\gamma &= 3\epsilon^2 + o(\epsilon^2) \\ \partial_2\gamma &= 3\epsilon^2 + o(\epsilon^2) \\ \partial_3\gamma &= 6\epsilon + o(\epsilon) \end{aligned}$$

and

$$\sigma_1'^{\,2} = \left(\frac{57}{2\rho_0^2} - 9\right)\epsilon^4 + o(\epsilon^{-3})$$

$$
\begin{aligned}
{\sigma'_2}^2 &= \left(\frac{57}{2\rho_0^2} - 9\right)\epsilon^4 + o(\epsilon^{-3}) \\
{\sigma'_3}^2 &= \left(\frac{114}{\rho_0^2} - 36\right)\epsilon^2 + o(\epsilon^2)
\end{aligned}
$$

under balanced failure biasing, and so

$$
\begin{aligned}
RE'_1 &= \frac{1}{3}\sqrt{\frac{57}{2\rho_0^2} - 9} + o(1) \\
RE'_2 &= \frac{1}{3}\sqrt{\frac{57}{2\rho_0^2} - 9} + o(1) \\
RE'_3 &= \frac{1}{6}\sqrt{\frac{114}{\rho_0^2} - 36} + o(1),
\end{aligned}
$$

all of which remain bounded as $\epsilon \to 0$. ●

4.5.2 Simple Failure Biasing

Nakayama [18] established a necessary and sufficient condition for when simple failure biasing gives rise to derivative estimators having bounded relative error. This condition is equivalent to the one given in Theorem 5 specialized to the case of simple failure biasing. Furthermore, Nakayama [18] presented examples showing that the conditions in Theorems 2 and 5 are not equivalent and that neither implies the other. In particular, it was shown that it is possible to obtain better estimates for a derivative than for the performance measure when using simple failure biasing. This contrasts the situation that occurs when using balanced failure biasing, in which the performance measure and all of the derivatives can be estimated with the same relative error.

Theorem 6 implies that simple failure biasing yields derivative estimators having bounded relative error when the system is balanced. Let us now investigate how the method works with estimating derivatives in our previous example.

Example 1 (continued) Recall that Figure 3 is the transition probability diagram under simple failure biasing. Consider estimating the derivative with respect to λ_1. We can easily show that $r_1 = 3$ and $\bar{r}_1 = \infty$ since at least one component of each type must fail for the system to fail, and so $r_1 - b_1 = 2 < \bar{r}_1 - b_0 = \infty$. As we previously established, the path $(\langle 0,0,0\rangle, \langle 0,0,1\rangle, \langle 1,0,1\rangle, \langle 1,1,1\rangle) \in \Delta_3^1$ under simple failure biasing has probability $\Theta(\epsilon) \neq \underline{O}(1)$, and so Theorem 5 implies that the estimator of the derivative with respect to λ_1 will not have bounded relative error when simple failure biasing is used. In fact, we can show (after a lot of algebra) that

$$
{\sigma'_1}^2 = \frac{6}{\rho_0^2}\epsilon^3 + o(\epsilon^3)
$$

under simple failure biasing, which implies

$$RE_1' = \frac{\sqrt{2/3}}{\rho_0}\epsilon^{-1/2} + o(\epsilon^{-1/2}).$$

Similarly, we can show that the simple failure biasing estimators of $\partial_2\gamma$ and $\partial_3\gamma$ do not have bounded relative error. ●

4.5.3 Bias2 Failure Biasing

Theorem 6 implies that bias2 failure biasing yields derivative estimators having bounded relative error when the system is balanced. Let us now investigate how the method works with estimating derivatives in our previous example.

Example 1 (continued) Recall that Figure 4 is the transition probability diagram under bias2 failure biasing. Consider estimating the derivative with respect to λ_1. As we previously established, the path $(\langle 0,0,0\rangle, \langle 0,0,1\rangle, \langle 1,0,1\rangle, \langle 1,1,1\rangle) \in \Delta_3^1$ under bias2 failure biasing has probability $\Theta(\epsilon) \neq \underline{O}(1)$, and so Theorem 5 implies that the estimator of the derivative with respect to λ_1 will not have bounded relative error when bias2 failure biasing is used. In fact, we can show (after a lot of algebra) that

$$\sigma_1'^2 = \frac{4}{\rho_0^2(1-\rho_1)}\epsilon^3 + o(\epsilon^3)$$

under bias2 failure biasing, which implies

$$RE_1' = \frac{2}{3\rho_0\sqrt{1-\rho_1}}\epsilon^{-1/2} + o(\epsilon^{-1/2}) \to \infty$$

as $\epsilon \to 0$. Similarly, we can show that the bias2 failure biasing estimators of $\partial_2\gamma$ and $\partial_3\gamma$ do not have bounded relative error. ●

Since bias2 failure biasing may not result in bounded relative error when estimating derivatives in unbalanced systems whereas balanced failure biasing always does, balanced failure biasing is more robust. However, as Theorem 6 showed, bias2 failure biasing will yield derivative estimators with bounded relative error when simulating a balanced system. As in the case of simple failure biasing, it may turn out that for balanced systems, the coefficient in the leading term of the asymptotic expansion for the variance of a derivative estimator is better under bias2 failure biasing than under balanced failure biasing. Thus, bias2 failure biasing may be more appropriate than balanced failure biasing in certain situations.

4.5.4 Failure Distance Biasing

Theorem 6 implies that failure distance biasing yields derivative estimators having bounded relative error when the system is balanced. Let us now investigate how the method works with estimating derivatives in our previous example.

Example 1 (continued) Recall that Figure 5 is the transition probability diagram under failure distance biasing. Consider estimating the derivative with respect to the failure rate of component type 1. Now consider the path $(\langle 0,0,0\rangle, \langle 0,0,1\rangle, \langle 1,0,1\rangle, \langle 1,1,1\rangle)$, which is in Δ_3^1. We previously showed that the entire path has probability

$$P'\{(X_0,\ldots,X_{\tau_F}) = (\langle 0,0,0\rangle, \langle 0,0,1\rangle, \langle 1,0,1\rangle, \langle 1,1,1\rangle)\} = O(\epsilon)$$

under failure distance biasing. However, Theorem 5 requires this path to have probability $\Theta(1)$ under failure distance biasing to achieve bounded relative error. Therefore, the estimator of the derivative with respect to λ_1 in this example will not have bounded relative error if we use failure distance biasing. In fact, we can show (after a lot of algebra) that

$$\sigma_1'^{\,2} = \frac{4}{\rho_0^2\rho_d}\epsilon^3 + o(\epsilon^3)$$

under failure distance biasing, which implies

$$RE_1' = \frac{2}{3\rho_0\rho_d^{1/2}}\epsilon^{-1/2} + o(\epsilon^{-1/2}) \to \infty$$

as $\epsilon \to 0$. Similarly, we can show that the failure distance biasing estimators of $\partial_2\gamma$ and $\partial_3\gamma$ do not have bounded relative error. ●

Since failure distance biasing may not result in bounded relative error when estimating derivatives in unbalanced systems whereas balanced failure biasing always does, balanced failure biasing is more robust. However, as Theorem 6 showed, failure distance biasing will yield derivative estimators with bounded relative error when simulating a balanced system. As in the case of simple failure biasing, it may turn out that for balanced systems, the coefficient in the leading term of the asymptotic expansion for the variance of a derivative estimator is better under failure distance biasing than under balanced failure biasing. Thus, failure distance biasing may be more appropriate than balanced failure biasing in certain contexts.

5 Conclusion and Directions for Future Research

In this paper we have established general conditions determining when an importance sampling method will yield estimators of performance measures and their derivatives with bounded

relative error. In particular, we provided a necessary and sufficient conditon for a class of importance sampling measures which include all of the failure biasing methods currently in the literature. Using our condition, we analyzed the various failure biasing methods and showed that of these, only the balanced failure biasing method is guaranteed to always produce bounded relative error. However, for a given model, another failure biasing method may yield slightly better estimators.

One topic for future research is to develop a simple method of determining for a given model which failure biasing method works best. Theorems 2 and 5 provide insight into this problem, but applying these results in practice can be potentially difficult because of the number of sample paths which must be examined. Another area worth investigating is to use our new theory to devise more efficient importance sampling schemes for highly reliable Markovian systems. Also, as demonstrated by Juneja and Shahabuddin [15], failure biasing methods no longer produce estimators having bounded relative error when there are deferred repairs. Thus, it would be interesting to determine if necessary and sufficient conditions for bounded relative error similar to those established in this paper can also be proven for these types of systems. Finally, we would like to establish a general theory for broad classes of importance sampling schemes for highly reliable non-Markovian systems.

6 Acknowledgements

Much of this work was completed while the author was a post-doctoral fellow at the IBM Thomas J. Watson Research Laboratories in Yorktown Heights, NY. The author would like to thank Randy Nelson for suggesting to work on this problem and for giving helpful comments on an earlier draft of this paper. Also, the author extends his gratitude to Perwez Shahabuddin for providing suggestions on the paper.

7 Appendix

Proof of Theorem 5. First assume that $r_i - b_i \leq \bar{r}_i - b_0$, and suppose that conditions (i)(a) and (i)(b) hold. Then Proposition 2 implies that $\partial_i \gamma = \Theta(\epsilon^{r_i - b_i})$. Thus, we need to establish that $E'[1\{\tau_F < \tau_0\} D_i^2 L^2] = \Theta(\epsilon^{2r_i - 2b_i})$. Using the Schwarz inequality, we obtain

$$E'[1\{\tau_F < \tau_0\} D_i^2 L^2] \geq \left(E'[1\{\tau_F < \tau_0\} D_i L]\right)^2 = (\partial_i \gamma)^2 = \Theta(\epsilon^{2r_i - 2b_i}),$$

and so it suffices to show $E'[1\{\tau_F < \tau_0\} D_i^2 L^2] = O(\epsilon^{2r_i - 2b_i})$.

From (14), we have that

$$\begin{aligned}
&E'[1\{\tau_F < \tau_0\}D_i^2L^2] \\
&= E[1\{\tau_i \le \tau_F < \tau_0\}D_i^2L] + E[1\{\tau_F < \min\{\tau_i,\tau_0\}\}D_i^2L] \\
&= \sum_{m=r_i}^{\infty} \sum_{\substack{(x_0,\ldots,x_n)\in\Delta_m^i \\ n>0}} D_i^2(x_0,\ldots,x_n)L(x_0,\ldots,x_n)P\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} \\
&\quad + \sum_{m=\bar{r}_i}^{\infty} \sum_{\substack{(x_0,\ldots,x_n)\in\bar{\Delta}_m^i \\ n>0}} D_i^2(x_0,\ldots,x_n)L(x_0,\ldots,x_n)P\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\}.
\end{aligned}$$

Now consider some $(x_0,\ldots,x_n) \in \Delta_m^i$ with $r_i \le m \le 2r_i - 1$. By assumption,

$$P'\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = O(\epsilon^{2m-2r_i}),$$

and so from Lemma 2(i),

$$\begin{aligned}
&D_i^2(x_0,\ldots,x_n)L(x_0,\ldots,x_n)P\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} \\
&\quad = \Theta(\epsilon^{-2b_i})\frac{\Theta(\epsilon^m)}{\underline{O}(\epsilon^{2m-2r_i})}\Theta(\epsilon^m) \;=\; O(\epsilon^{2r_i-2b_i}).
\end{aligned}$$

Thus, since $|\Delta_m^i| < \infty$ for all m by Lemma 1(ii),

$$\sum_{m=r_i}^{2r_i-1} \sum_{\substack{(x_0,\ldots,x_n)\in\Delta_m^i \\ n>0}} D_i^2(x_0,\ldots,x_n)L(x_0,\ldots,x_n)P\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} = O(\epsilon^{2r_i-2b_i}). \tag{15}$$

Also, Lemmas 1 and 2 imply that

$$\begin{aligned}
&\sum_{m=2r_i}^{\infty} \sum_{\substack{(x_0,\ldots,x_n)\in\Delta_m^i \\ n>0}} D_i^2(x_0,\ldots,x_n)L(x_0,\ldots,x_n)P\{(X_0,\ldots,X_{\tau_F}) = (x_0,\ldots,x_n)\} \\
&\le \sum_{m=2r_i}^{\infty} \sum_{\substack{(x_0,\ldots,x_n)\in\Delta_m^i \\ n>0}} \left((m+1)\phi\epsilon^{-b_i}\right)^2 \eta^{m+1}\alpha\beta^m\epsilon^m \\
&\le \sum_{m=2r_i}^{\infty} |S|^{(m+1)N} \left((m+1)\phi\epsilon^{-b_i}\right)^2 \eta^{m+1}\alpha\beta^m\epsilon^m \\
&= \alpha\eta\phi^2|S|^N\epsilon^{-2b_i} \sum_{m=2r_i}^{\infty} (m+1)^2 \left(\beta\eta|S|^N\epsilon\right)^m \;=\; \Theta(\epsilon^{2r_i-2b_i})
\end{aligned}$$

as $\epsilon \to 0$. Hence, $E[1\{\tau_i \le \tau_F < \tau_0\}D_i^2L] = O(\epsilon^{2r_i-2b_i})$. Similarly, we can show that $E[1\{\tau_F < \min\{\tau_i,\tau_0\}\}D_i^2L] = O(\epsilon^{2r_i-2b_i})$, and it follows that $E'[1\{\tau_F < \tau_0\}D_i^2L^2] = O(\epsilon^{2r_i-2b_i})$, which is what we needed to establish. Moreover, using the same types of arguments, we can show that the result holds when $r_i - b_i > \bar{r}_i - b_0$ and the conditions (ii)(a) and (ii)(b) are in force.

On the other hand, since $\mathcal{J} \subset \mathcal{I}'$, it follows from Theorem 4 that $RE_i' \to \infty$ as $\epsilon \to 0$ if the conditions do not hold. ■

References

[1] Carrasco, J. A. Failure distance based simulation of repairable fault tolerant systems. *Proceedings of the 5th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*. Balbo, G. and Serazzi, G. (Ed.). Elsevier Science Publishers B.V., (1992), 351–365.

[2] Carrasco, J. A. Efficient transient simulation of failure/repair Markovian models. *Proceedings of the Tenth Symposium on Reliable Distributed Systems*, IEEE Press, (1991), 152–161.

[3] Chang, C. S., Heidelberger, P., Juneja, S., and Shahabuddin, P. Effective bandwidth and fast simulation of ATM intree networks. *Performance Evaluation* **20** (1994), 45–66.

[4] Conway, A. E. and Goyal, A. Monte Carlo simulation of computer system availability/reliability models. *Proc. 17th International Symposium on Fault Tolerant Computing*, Pittsburgh, PA, (1987), 230–235.

[5] Cottrell, M., Fort, J. C., and Malgouyres, G. Large deviations and rare events in the study of stochastic algorithms. *IEEE Trans. Automatic Control* **AC-28** (1983), 907–920.

[6] Glasserman, P. *Gradient Estimation Via Perturbation Analysis*. Kluwer Academic, Norwell, Massachusetts, (1990).

[7] Glynn, P. W. Likelihood ratio derivative estimators for stochastic systems. *Comm. ACM* **33** (1990), 75–84.

[8] Glynn, P. W. and Iglehart, D. L. Importance sampling for stochastic simulations. *Management Sci.* **35** (1989), 1367–1393.

[9] Goyal, A., Heidelberger, P. and Shahabuddin, P. Measure Specific Dynamic Importance Sampling for Availability Simulations. *Proceedings of the 1987 Winter Simulation Conference*, Thesen, A., Grant, H. and Kelton, W. D. (eds.), IEEE Press, (1987), 351–357.

[10] Goyal, A. and Lavenberg, S. S. Modeling and analysis of computer system availability. *IBM J. Res. Develop.* **31** (1987), 651–664.

[11] Goyal, A., Shahabuddin, P., Heidelberger, P., Nicola, V. F. and Glynn, P. W. A unified framework for simulating Markovian models of highly dependable systems. *IEEE Trans. Comput.* **C-41** (1992), 36–51.

[12] Hammersley, J. M. and Handscomb, D. C. *Monte Carlo Methods*. Metheun, London, (1964).

[13] Heidelberger, P., Nicola, V. F., and Shahabuddin, P. Simultaneous and efficient simulation of highly dependable systems with different underlying distributions. *Proceedings of the 1992 Winter Simulation Conference*, Swain, J. J., Goldsman, D., Crain, R. C., and Wilson, J. R. (eds.), 458–465.

[14] Heidelberger, P., Shahabuddin, P., and Nicola, V. F. Bounded relative error in estimating transient measures of highly dependable non-Markovian systems. *ACM Transactions on Modeling and Computer Simulation* **4** (1994), 137–164.

[15] Juneja, S. and Shahabuddin, P. Fast simulation of Markovian reliability/availability models with general repair policies. *Proc. 22nd International Symposium on Fault Tolerant Computing*, Boston, MA, (1992), IEEE Computer Society Press, Los Alamitos, California, 150–159.

[16] L'Ecuyer, P. A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Sci.* **36** (1990), 1364–1383.

[17] Lewis, E. E. and Böhm, F. Monte Carlo simulation of Markov unreliability models. *Nuclear Engineering and Design* **77** (1984), 49–62.

[18] Nakayama, M. K. A characterization of the simple failure biasing method for simulations of highly reliable Markovian systems. *ACM Transactions on Modeling and Computer Simulation* **4** (1994), 52–88.

[19] Nakayama, M. K. Asymptotics of likelihood ratio derivative estimators in simulations of highly reliable Markovian systems. *Management Science*, to appear.

[20] Nakayama, M. K., Goyal, A. and Glynn, P. W. Likelihood ratio sensitivity analysis for Markovian models of highly dependable systems. *Operations Research* **42** (1994), 137–157.

[21] Nicola, V. F., Nakayama, M. K., Heidelberger, P., and Goyal, A. Fast simulation of highly dependable systems with general failure and repair processes. *IEEE Trans. Comput.* **42** (1993), 1440–1452.

[22] Nicola, V. F., Heidelberger, P., and Shahabuddin, P. Uniformization and exponential transformation: Techniques for fast simulation of highly dependable non-Markovian systems. *Proceedings of the 22nd International Symposium on Fault-Tolerant Computing*, IEEE Computer Society Press (1992), 130–139.

[23] Nicola, V. F., Shahabuddin, P., Heidelberger, P., and Glynn, P. W. Fast simulation of steady-state availability in non-Markovian highly dependable systems. In *Proceedings of the 23rd International Symposium on Fault-Tolerant Computing*, Toulouse, France, (1993), IEEE Computer Society Press, Los Alamitos, California, 38–47.

[24] Reiman, M. I. and Weiss, A. Sensitivity analysis for simulations via likelihood ratios. *Oper. Res.* **37** (1989), 830–844.

[25] Shahabuddin, P. Importance sampling for the simulation of highly reliable Markovian systems. *Management Sci.* **40** (1994), 333–352.

[26] Shahabuddin, P. Fast transient simulation of Markovian models of highly dependable systems. *Performance Evaluation* **20** (1994), 267–286.

[27] Shahabuddin, P. and Nakayama, M. K. Estimation of reliability and its derivatives for large time horizons in Markovian systems. *Proceedings of the 1993 Winter Simulation Conference*, Evans, G. W., Mollaghasemi, M., Russell, E.C, and Biles, W. E. (eds.), IEEE Press, (1993), 422–429.

[28] Shahabuddin, P., Nicola, V. F., Heidelberger, P., Goyal, A. and Glynn, P. W. Variance reduction in mean time to failure simulations. *Proceedings of the 1988 Winter Simulation Conference*, Abrams, M. A., Haigh, P. L. and Comfort, J. C. (eds.), IEEE Press, (1988), 491–499.