On Derivative Estimation of the Mean Time to Failure 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

The mean time to failure (MTTF) of a Markovian system can be expressed as a ratio of two expectations. For highly reliable Markovian systems, the resulting ratio formula consists of one expectation that cannot be estimated with bounded relative error when using standard simulation, while the other, which we call a non-rare expectation, can be estimated with bounded relative error. We show that some derivatives of the non-rare expectation cannot be estimated with bounded relative error when using standard simulation, which in turn may lead to an estimator of the derivative of the MTTF that has unbounded relative error. However, if particular importance-sampling methods (e.g., balanced failure biasing) are used, then the estimator of the derivative of the non-rare expectation will have bounded relative error, which (under certain conditions) will yield an estimator of the derivative of the MTTF with bounded relative error.

Subject classifications: Probability, stochastic model applications: highly dependable systems. Simulation: statistical analysis of derivative estimators. Simulation, efficiency: importance sampling.

1 Introduction

The mean time to failure (MTTF) μ of a highly reliable Markovian system satisfies a ratio formula $\mu = \xi/\gamma$, where ξ and γ are expectations of random quantities defined over regenerative cycles; e.g., see Goyal et al. (1992). Shahabuddin (1994) showed that ξ can be estimated with bounded relative error when using standard simulation (i.e., no importance sampling), and so we call ξ a non-rare expectation. (A simulation estimator has bounded relative error if the ratio of the standard deviation to the mean remains bounded as the failure rates of the components vanish. In practice, this means that one can obtain good estimates of the mean independently of how rarely the system fails.) He also proved that the standard-simulation estimator of γ has unbounded relative error, and so we call γ a rare expectation; however, if certain importance-sampling methods such as balanced failure biasing (see Shahabuddin 1994 and Goyal et al. 1992) are used to estimate γ, the corresponding estimator has bounded relative error (see Nakayama 1996 for generalizations and Hammersley and Handscomb 1964 and Glynn and Iglehart 1989 for details on importance sampling). Moreover, Shahabuddin (1994) showed that if both the numerator and denominator are estimated with bounded relative error, the resulting estimator of μ has bounded relative error.

In this paper, we consider estimators of derivatives of μ with respect to the failure rate λ_i of any component type i obtained using the likelihood-ratio derivative method (e.g., see Glynn 1990). Letting ∂_i denote the derivative operator with respect to λ_i, we have that

$$\partial_i \mu = \frac{(\partial_i \xi)\ \gamma\ -\ \xi\ (\partial_i \gamma)}{\gamma^2}. \tag{1}$$

Since the standard-simulation estimator of γ does not have bounded relative error whereas the one for ξ does, the previous research on derivative estimation in reliability systems focused on $\partial_i \gamma$. Nakayama (1995) showed that for any component type i, the standard-simulation estimator of $\partial_i \gamma$ has unbounded relative error and the balanced-failure-biasing estimator of $\partial_i \gamma$ has bounded relative error; see Nakayama (1996) for generalizations.

However, we now prove that when standard simulation is applied, the estimator of $\partial_i \xi$ may not have bounded relative error, even though the estimator of ξ always does. We show by example that this can result in an estimator of $\partial_i \mu$ that has unbounded relative error, even if ξ, γ, and $\partial_i \gamma$ are estimated (mutually independently) with bounded relative error (i.e., ξ is estimated using standard simulation and γ and $\partial_i \gamma$ are estimated with, for example, balanced failure biasing). Hence, we apply particular importance-sampling schemes (e.g., balanced failure biasing) to obtain an estimator of $\partial_i \xi$ having bounded relative error, which then results (under certain conditions) in an estimator of $\partial_i \mu$ having bounded relative error.

The rest of the paper is organized as follows. Section 2 contains a description of the mathe-

matical model. In Section 3 we first review the basics of derivative estimation and importance sampling. Then we present results on the asymptotics of $\partial_i \xi$, which are subsequently used in an example showing that applying standard simulation to estimate $\partial_i \xi$ can result in an estimator of $\partial_i \mu$ having unbounded relative error. Section 3 concludes with our theorem on the bounded relative error of an estimator of $\partial_i \mu$. The proofs are collected in the appendix. For closely related empirical results on the estimation of $\partial_i \mu$, see Nakayama, Goyal, and Glynn (1994). (In that paper all four terms in (1) are estimated using the same simulation runs with importance sampling, whereas in our analysis here, we do not apply importance sampling to estimate ξ and the four quantities are estimated independently.)

2 Model

Shahabuddin (1994) developed a model of a highly reliable Markovian system to analyze some performance measure estimators, and Nakayama (1995) later modified it to study derivative estimators. We will work with the latter model, which we now describe briefly.

The system consists of $K < \infty$ different types of components, and there are n_i components of type i, with $n_i < \infty$. As time evolves, the components fail at random times and are repaired by some repairpersons. We model the evolution of the system as a continuous-time Markov chain (CTMC) $Y = \{Y(s) : s \geq 0\}$ on a finite state space S. We decompose S as $S = U \cup F$, where U (resp., F) is the set of operational (resp., failed) states. We assume that the system starts in state 0, the state with all components operational, and that $0 \in U$. Also, we assume that the system is coherent; i.e., if $x \in U$ and $y \in S$ with $n_i(y) \geq n_i(x)$ for all $i = 1, 2, \ldots, K$, then $y \in U$, where $n_i(x)$ denotes the number of components of type i operational in state x.

The lifetimes of each component of type i are exponentially distributed with rate $\lambda_i > 0$, and we let $\lambda \equiv (\lambda_1, \lambda_2, \ldots, \lambda_K)$. We will examine derivatives with respect to the λ_i for different component types i. We let $p(y; x, i)$ be the probability that the next state visited is y when the current state is x and a component of type i fails. This general form of the state transitions allows for component failure propagation (i.e., the failure of one component causes others to fail simultaneously with some probability). We denote a failure transition (x, y) (i.e., a transition of Y corresponding to the failure of some component(s)) by "$x \stackrel{f}{\longrightarrow} y$." A repair transition (x, y) (i.e., a transition of Y corresponding to the repair of some component(s)), which we denote by "$x \stackrel{r}{\longrightarrow} y$," occurs at exponential rate $\nu(x, y) \geq 0$. A single transition (x, y) cannot consist of some components failing and others completing repair.

Let $\mathbf{Q} \equiv \mathbf{Q}(\lambda) = \{q(\lambda, x, y) : x, y \in S\}$ be the generator matrix of Y. Let $P\{\cdot\}$, $E[\cdot]$, and $\mathrm{Var}[\cdot]$ be the probability measure and expectation and variance operators, respectively,

induced by the **Q**-matrix. The total transition rate out of state x is

$$q(\lambda, x) \equiv -q(\lambda, x, x) = \sum_{i=1}^{K} n_i(x)\lambda_i + \sum_{y \,:\, x \xrightarrow{r} y} \nu(x, y). \tag{2}$$

Let $X = \{X_n : n = 0, 1, 2, \ldots\}$ be the embedded discrete-time Markov chain (DTMC) of the CTMC Y. Let $\mathbf{P}(\lambda) = \{\mathbf{P}(\lambda, x, y) : x, y \in S\}$ denote the transition probability matrix of X. Then define $\Gamma = \{(x, y) : \mathbf{P}(\lambda, x, y) > 0\}$, which is the set of possible transitions of the DTMC and is independent of the parameter setting λ.

As in Shahabuddin (1994) and Nakayama (1995), the failure rate of each component type i is parameterized as $\lambda_i(\epsilon) = \tilde{\lambda}_i \epsilon^{b_i}$, where $b_i \geq 1$ is an integer, $\tilde{\lambda}_i > 0$, and $\epsilon > 0$. We define $b_0 = \min_{1 \leq i \leq K} b_i$. All other parameters in the model (including the repair rates) are independent of ϵ. In the literature on highly reliable systems, the behavior of the system is examined as $\epsilon \to 0$. In the following, we will sometimes parameterize quantities by ϵ rather than λ to emphasize when limits are being considered.

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 use the following assumptions from Shahabuddin (1994) and Nakayama (1995):

A1 The CTMC Y is irreducible over S.

A2 For each state $x \in S$ with $x \neq 0$, there exists a state $y \in S$ (which depends on x) such that $(x, y) \in \Gamma$ and $x \xrightarrow{r} y$.

A3 For all states $y \in F$ such that $(0, y) \in \Gamma$, $q(\epsilon, 0, y) = o(\epsilon^{b_0})$.

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 **A3** ensures that if the fully operational system can reach a failed state y in one transition, then the probability of this transition is much smaller than the largest transition probability from state 0. Nakayama (1995) introduced **A4** and **A5** as technical assumptions to simplify the analysis of certain derivative estimators.

3 Derivative Estimators of μ

Our goal is to analyze estimators of derivatives of the MTTF μ, where $\mu = E[\tau_F|Y(0) = 0]$ with $\tau_F = \inf\{s > 0 : Y(s) \in F\}$. Goyal et al. 1992 showed that $\mu = \xi/\gamma$, with $\xi = E[\min\{G_F, G_0\}]$ and $\gamma = E[1\{T_F < T_0\}]$, where for some set of states $J \subset S$, $T_J = \inf\{n > 0 : X_n \in J\}$, $G_J = \sum_{k=0}^{T_J-1} 1/q(\lambda, X_k)$, and $1\{\cdot\}$ denotes the indicator function of the event $\{\cdot\}$. To simplify notation, let $G = \min\{G_F, G_0\} = \sum_{k=0}^{T-1} 1/q(\lambda, X_k)$ and $I = 1\{T_F < T_0\}$, where $T = \min\{T_0, T_F\}$. Also, observe that since $X_0 = 0$ with probability 1, $G = 1/q(\lambda, 0) + H$ with probability 1 and $\xi = (1/q(\lambda, 0)) + E[H]$, where $H = \sum_{k=1}^{T-1} 1/q(\lambda, X_k)$.

Now recall our expression for the derivative of μ given in (1). Shahabuddin (1994) analyzed estimators of γ and ξ, and Nakayama (1995) studied estimators of $\partial_i\gamma$. Thus, to complete the analysis of $\partial_i\mu$, we now will study $\partial_i\xi$. Using the likelihood ratio method for computing derivatives, we obtain $\partial_i\xi = E[G_i]$, where

$$G_i \equiv \frac{-\partial_i q(\lambda, 0)}{q^2(\lambda, 0)} + H_i + HS_i = \frac{-n_i}{q^2(\lambda, 0)} + H_i + HS_i$$

by (2), and

$$H_i \equiv \partial_i H = -\sum_{k=1}^{T-1} \frac{\partial_i q(\lambda, X_k)}{q^2(\lambda, X_k)} = -\sum_{k=1}^{T-1} \frac{n_i(X_k)}{q^2(\lambda, X_k)} \tag{3}$$

with $S_i = \sum_{k=0}^{T-1} \mathbf{P}_i(\lambda, X_k, X_{k+1})/\mathbf{P}(\lambda, X_k, X_{k+1})$ and $\mathbf{P}_i(\lambda, x, y) = \partial_i \mathbf{P}(\lambda, x, y)$. (For further details on the likelihood ratio method for estimating derivatives, see Glynn 1990.)

We now briefly review the basics of standard simulation and importance sampling. Consider a random variable Z having a density f, and we want to estimate $\alpha = E[Z]$, where E is the expectation operator under f. We apply standard simulation by collecting i.i.d. samples $Z^{(j)}$, $j = 1, 2, \ldots, n$, of Z generated using density f and constructing the estimator $\hat{\alpha} = (1/n)\sum_{j=1}^{n} Z^{(j)}$. Let g be a density that is absolutely continuous with respect to f (i.e., $g(z) = 0$ implies $f(z) = 0$ for $z \in \Re$), and let $\tilde{E}$ denote its expectation operator. Define $L(z) = f(z)/g(z)$ to be the likelihood ratio evaluated at the point $z \in \Re$, and define the random variable $L \equiv L(Z)$. Then, $\alpha = \tilde{E}[ZL]$. We implement importance sampling by collecting i.i.d. samples $(Z^{(j)}, L^{(j)})$, $j = 1, 2, \ldots, n$, of (Z, L) generated using the density g, and constructing the estimator $\tilde{\alpha} = (1/n)\sum_{j=1}^{n} Z^{(j)}L^{(j)}$. Properly choosing the new density g can result in a (substantial) variance reduction. As we shall soon see, importance sampling can be generalized beyond the realm of single random variables having a density to consider complex stochastic systems; for more details on importance sampling, see Glynn and Iglehart (1989).

We now show how these ideas apply in the estimation of $\partial_i\xi$. We use standard simulation to estimate $\partial_i\xi$ by collecting i.i.d. observations $G_i^{(j)}$, $j = 1, 2, \ldots, n$, of G_i. Each sample $G_i^{(j)}$ is generated by using the transition matrix $\mathbf{P}$ to simulate the DTMC X starting in state 0 until

$F \cup \{0\}$ is hit. Then our standard-simulation estimator of $\partial_i \xi$ is

$$\hat{\partial}_i \xi = \frac{1}{n} \sum_{j=1}^{n} G_i^{(j)}.$$

Turning now to importance sampling, consider a distribution $\tilde{P}$ (which may depend on the particular failure and repair rates) that is absolutely continuous with respect to P. We define $L(x_0, \ldots, x_n) = P\{(X_0, \ldots, X_T) = (x_0, \ldots, x_n)\}/\tilde{P}\{(X_0, \ldots, X_T) = (x_0, \ldots, x_n)\}$ to be the likelihood ratio evaluated at the sample path $(x_0, \ldots, x_n)$, and let $L = L(X_0, \ldots, X_T)$. Then $\partial_i \xi = -n_i/q^2(\epsilon, 0) + \tilde{E}[(H_i + HS_i)L]$, where $\tilde{E}$ is the expectation operator under $\tilde{P}$. We assume

A6 The distribution $\tilde{P}$ is Markovian with transition matrix $\tilde{\mathbf{P}}(\epsilon) = (\tilde{\mathbf{P}}(\epsilon, x, y) : x, y \in S)$ such that $\tilde{\mathbf{P}}(\epsilon, x, y) = 0$ implies that $\mathbf{P}(\epsilon, x, y) = 0$ (i.e., $\tilde{\mathbf{P}}$ is absolutely continuous with respect to $\mathbf{P}$). Thus, $L(x_0, \ldots, x_n) = \prod_{k=0}^{n-1} \mathbf{P}(\epsilon, x_k, x_{k+1})/\tilde{\mathbf{P}}(\epsilon, x_k, x_{k+1})$. Also, for all $(x, y) \in \Gamma$, $\tilde{\mathbf{P}}(\epsilon, x, y) = \Theta(1)$ as $\epsilon \to 0$.

We apply importance sampling by collecting i.i.d. samples $(H_i^{(j)}, H^{(j)}, S_i^{(j)}, L^{(j)})$, $j = 1, 2, \ldots, n$, of (H_i, H, S_i, L) generated by simulating the DTMC using the transition matrix $\tilde{\mathbf{P}}$. Then

$$\tilde{\partial}_i \xi = -n_i/q^2(\epsilon, 0) + \frac{1}{n} \sum_{j=1}^{n} \left(H_i^{(j)} + H^{(j)} S_i^{(j)} \right) L^{(j)}$$

is the importance-sampling estimator.

Balanced failure biasing is an importance-sampling method that satisfies Assumption **A6**. The basic idea of the technique is as follows. From any state x having both failure and repair transitions, we increase (resp., decrease) the total probability of a failure (resp., repair) transition to ρ (resp., $1-\rho$), where ρ is independent of ϵ. We allocate the ρ equally to the individual failure transitions from x. The $1-\rho$ is allotted to the individual repair transitions in proportion to their original transition probabilities. From state 0, we change the transition probability of any possible (failure) transition to $1/m$, where m is the number of failure transitions possible from state 0. See Shahabuddin (1994) and Goyal et al. (1992) for more details.

As previously noted, we need estimators of γ, $\partial_i \gamma$, ξ, and $\partial_i \xi$ in (1) to estimate $\partial_i \mu$. In a manner analogous to how $\hat{\partial}_i \xi$ was developed earlier, we can construct $\hat{\xi}$ and $\hat{\gamma}$, which are the standard-simulation estimators of ξ and γ, respectively. Shahabuddin (1994) showed that $\xi = \Theta(\epsilon^{-b_0})$, and that $\mathrm{Var}[G] = O(1)$, where Var represents the variance operator under the measure P. Thus, the relative error of $\hat{\xi}$, defined as $RE(\hat{\xi}) \equiv \sqrt{\mathrm{Var}[G]}/\xi$, remains bounded (and actually vanishes) as $\epsilon \to 0$. On the other hand, Shahabuddin proved that $\gamma = \Theta(\epsilon^r)$ for some constant $r \geq 1$ which depends on the model and that $\mathrm{Var}[I] = \gamma - \gamma^2 = \Theta(\epsilon^r)$, so $RE(\hat{\gamma}) \equiv \sqrt{\mathrm{Var}[I]}/\gamma \to \infty$ as $\epsilon \to 0$. Also, $\tilde{\mathrm{V}}\mathrm{ar}[IL] = \Theta(\epsilon^{2r})$, where $\tilde{\mathrm{V}}\mathrm{ar}$ is the variance

operator under a probability measure $\tilde{P}$ satisfying Assumption **A6**. Hence, the corresponding importance-sampling estimator $\tilde{\gamma}$ of γ under the measure $\tilde{P}$ has bounded relative error. Nakayama (1995,1996) proved that $\partial_i\gamma = \Theta(\epsilon^{d_i})$, where $d_i = \min(r_i - b_i, \bar{r}_i - b_0)$ and $r_i \geq r$ and $\bar{r}_i \geq r$ are constants depending on the model. It was also shown that $\mathrm{Var}[IS_i] = \Theta(\epsilon^{v_i})$, where $v_i = \min(r_i - 2b_i, \bar{r}_i - 2b_0)$, and $\tilde{\mathrm{V}}\mathrm{ar}[IS_iL] = \Theta(\epsilon^{2d_i})$ under any measure $\tilde{P}$ satisfying Assumption **A6**. Therefore, the standard-simulation estimator of $\partial_i\gamma$ has unbounded relative error, whereas the importance-sampling estimator has bounded relative error. In this paper, we analyze various estimators of $\partial_i\xi$ and $\partial_i\mu$.

We start with the analysis of $\partial_i\xi$. The following result shows that standard simulation is not always an efficient way of estimating $\partial_i\xi$, and so importance sampling needs to be applied.

Lemma 1 *Consider any system as described in Section 2. Then,*

(i) $\partial_i\xi = -n_i/q^2(\epsilon, 0) + E[H_i + HS_i]$, *where* $n_i/q^2(\epsilon, 0) = \Theta(\epsilon^{-2b_0})$ *and* $E[H_i + HS_i] = O(\epsilon^{-b_0})$*; hence,* $\partial_i\xi = \Theta(\epsilon^{-2b_0})$*;*

(ii) *When applying standard simulation,* $\sigma_i^2 \equiv Var[G_i] = \Theta(\epsilon^{-b_0-b_i})$*. Thus, when standard simulation is used,* $RE(\hat{\partial}_i\xi) \equiv \sigma_i/(\partial_i\xi)$ *remains bounded as* $\epsilon \to 0$ *if and only if* $b_i \leq 3b_0$*.*

If an importance-sampling distribution $\tilde{P}$ *satisfying Assumption* **A6** *is applied, then*

(iii) $\tilde{\sigma}_i^2 \equiv \tilde{Var}[-n_i/q^2(\epsilon, 0) + (H_i + HS_i)L] = O(\epsilon^{-2b_0})$ *as* $\epsilon \to 0$*;*

(iv) *For any component type* i*, the relative error of the importance-sampling estimator of* $\partial_i\xi$*,* $RE(\tilde{\partial}_i\xi) \equiv \tilde{\sigma}_i/(\partial_i\xi) \to 0$ *as* $\epsilon \to 0$*.*

Now let us turn to the estimation of $\partial_i\mu$, which is what we are really interested in. For our estimator of $\partial_i\mu$, we assume that $\partial_i\xi$, γ, ξ, and $\partial_i\gamma$ are estimated mutually independently. Suppose we have four (importance-sampling) probability measures P_A, P_B, P_C, and P_D having corresponding expectation operators E_A, E_B, E_C, and E_D. Any of these probability measures may be the original probability measure P. Also, suppose we have random variables A, B, C, and D having measures P_A, P_B, P_C, and P_D, respectively, for which $E_A[A] = \partial_i\xi$, $E_B[B] = \gamma$, $E_C[C] = \xi$, and $E_D[D] = \partial_i\gamma$. We assume there exist constants $c_1 \neq 0$, $c_2 \neq 0$, $c_3 \neq 0$, $c_4 \neq 0$ and a, b, c, d that are independent of ϵ such that $\partial_i\xi = c_1\epsilon^a + o(\epsilon^a)$, $\gamma = c_2\epsilon^b + o(\epsilon^b)$, $\xi = c_3\epsilon^c + o(\epsilon^c)$, and $\partial_i\gamma = c_4\epsilon^d + o(\epsilon^d)$. As we saw in Lemma 1, $a = -2b_0$. Shahabuddin (1994) showed that $b = r \geq 1$ and $c = -b_0$, and Nakayama (1995) established that $d = d_i$.

We construct an estimator of $\partial_i\mu$ as follows. Collect n_a (resp., n_b, n_c, and n_d) i.i.d. samples of A (resp., B, C, and D) generated using measure P_A (resp., P_B, P_C, and P_D), where the observations of A, B, C, and D are mutually independent. This yields $A^{(i)}$, $i = 1, 2, \ldots, n_a$;

$B^{(j)}$, $j = 1, 2, \ldots, n_b$; $C^{(k)}$, $k = 1, 2, \ldots, n_c$; and $D^{(l)}$, $l = 1, 2, \ldots, n_d$. Let $\bar{A} = \sum_{i=1}^{n_a} A^{(i)}/n_a$, and similarly define $\bar{B}$, $\bar{C}$ and $\bar{D}$. Then our estimator of $\partial_i \mu$ is $\tilde{\partial}_i \mu = (\bar{A}\bar{B} - \bar{C}\bar{D})/\bar{B}^2$. This approach is known as "measure-specific importance sampling"; see Goyal et al. (1992). Now let σ_A^2, σ_B^2, σ_C^2, and σ_D^2 be the variances of A, B, C, and D, respectively, under their corresponding measures. Nakayama, Goyal, and Glynn (1994) showed that the asymptotic variance of the resulting estimator of $\partial_i \mu$ is

$$\sigma^2 = \frac{1}{\gamma^2}\sigma_A^2 + \left[\frac{2\xi(\partial_i\gamma) - (\partial_i\xi)\gamma}{\gamma^3}\right]^2 \sigma_B^2 + \frac{(\partial_i\gamma)^2}{\gamma^4}\sigma_C^2 + \frac{\xi^2}{\gamma^4}\sigma_D^2. \tag{4}$$

There are no covariance terms since A, B, C, and D are mutually independent.

Now let us examine a particular estimator of $\partial_i \mu$ for a specific model.

Example 1: Consider a system with three types of components, where all three component types have a redundancy of 1 (i.e., $n_1 = n_2 = n_3 = 1$). The first two component types have failure rate ϵ (i.e., $b_1 = b_2 = 1$), and components of the third type have failure rate ϵ^4 (i.e., $b_3 = 4$); therefore, $b_0 = 1$. Failed components are fixed by a single repairperson at rate 1 using a processor-sharing discipline, and the system is operational if and only if at least two components (of any types) are operational. We consider derivatives with respect to λ_3. It can be shown that $\partial_3 \xi = -\epsilon^{-2}/4 + o(\epsilon^{-2})$, $\gamma = \epsilon + o(\epsilon)$, $\xi = \epsilon^{-1}/2 + o(\epsilon^{-1})$, $\partial_3 \gamma = 3/2 + o(1)$, $\mu = \epsilon^{-2}/2 + o(\epsilon^{-2})$, and $\partial_3 \mu = -\epsilon^{-3} + o(\epsilon^{-3})$. Now suppose that we use measure-specific importance sampling in which γ and $\partial_3 \gamma$ are estimated using a probability measure $\tilde{P}$ satisfying Assumption **A6** and ξ and $\partial_3 \xi$ are estimated using standard simulation; i.e., $A = G_i$, $B = IL$, $C = G$, $D = IS_iL$ and $P_A = P_C = P$ and $P_B = P_D = \tilde{P}$. We can show that $\sigma_A^2 = \Theta(\epsilon^{-5})$, $\sigma_B^2 = \Theta(\epsilon^2)$, $\sigma_C^2 = \Theta(1)$, and $\sigma_D^2 = \Theta(1)$. Hence, the estimators of γ, ξ, and $\partial_3 \gamma$ have bounded relative error, but the estimator of $\partial_3 \xi$ does not. Moreover, the variance of the resulting estimator of $\partial_3 \mu$ is $\Theta(\epsilon^{-7})$, and so the estimator of $\partial_3 \mu$ has unbounded relative error. ■

Thus, if the estimator of $\partial_i \xi$ has unbounded relative error, then the resulting estimator of $\partial_i \mu$ may also, even though the other terms in (1) are estimated with bounded relative error. On the other hand, we have the following theorem, which shows that under certain conditions, if each of the four terms is estimated with bounded relative error, then the corresponding estimator of $\partial_i \mu$ has bounded relative error.

Theorem 1 *Consider any system as described in Section 2. Using the notation above, assume that $c_1c_2 - c_3c_4 \neq 0$ whenever $a + b = c + d$. Consider any importance-sampling distribution $\tilde{P}$ satisfying Assumption* **A6**, *and let $P_C = P$ and $P_A = P_B = P_D = \tilde{P}$. Let $A = -n_i/q^2(\epsilon, 0) + (H_i + HS_i)L$, $B = IL$, $C = G$, $D = IS_iL$, and assume that A, B, C, and D are mutually independent. Then, $\sigma/(\partial_i \mu)$ remains bounded as $\epsilon \to 0$.*

Remarks:

(i) In Example 1, $c_1 = -1/4$, $c_2 = 1$, $c_3 = 1/2$, $c_4 = 3/2$, and $a = -2$, $b = 1$, $c = -1$, $d = 0$. Thus, $a + b = c + d$ and $c_1c_2 - c_3c_4 \neq 0$, and so the estimator of $\partial_3\mu$ based on Theorem 1 will have bounded relative error.

(ii) We can show that when the technical condition of Theorem 1 is not satisfied, the resulting derivative corresponds to a small sensitivity (i.e., there are other sensitivities that are asymptotically at least an order of magnitude larger). (The sensitivity of the MTTF with respect to λ_i is defined to be $\lambda_i \cdot \partial_i\mu$, which measures the effect of relative changes in the parameter value on the overall MTTF.) Therefore, even when we may not be able to estimate the derivative with bounded relative error, it is typically not that important since there are other derivatives with respect to other failure rates that have a much larger impact on the MTTF.

(iii) Theorem 1 easily generalizes to an estimator of the derivative of *any* ratio formula, not only for the MTTF. Specifically, suppose $(\bar{A}\bar{B} - \bar{C}\bar{D})/\bar{B}^2$ is any derivative estimator (not just for $\partial_i\mu$) with $\bar{A}$, $\bar{B}$, $\bar{C}$, and $\bar{D}$ mutually independent and each having bounded relative error under their respective probability measures. Then, $(\bar{A}\bar{B} - \bar{C}\bar{D})/\bar{B}^2$ has bounded relative error if the technical condition of Theorem 1 holds. In particular, the steady-state unavailability satisfies a ratio formula consisting of a rare expectation and a non-rare expectation, and so we can similarly analyze this performance measure.

4 Appendix

We will prove Lemma 1 by using some order of magnitude and bounding arguments as in Nakayama (1995). To do this, we define Ω to be the set of state sequences that start in state 0 and end in either 0 or F; i.e.,

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

For $m \geq 0$, let

$$\Omega_m = \{ \ (x_0, \ldots, x_n) \in \Omega \ : \ n \geq 1, \ P\{(X_0, \ldots, X_n) = (x_0, \ldots, x_n)\} = \Theta(\epsilon^m) \ \},$$

the set of state sequences in Ω that have probability of order ϵ^m (under the original measure).

For each component type i, let $T_i = \inf\{k > 0 : X_{k-1} \xrightarrow{f} X_k, \ n_i(X_{k-1})p(X_k; X_{k-1}, i) > 0\}$, which is the first failure transition of the DTMC X that may have been triggered by a failure of a component of type i. We use the notation $T_i(x_0, \ldots, x_n)$ to denote the random

variable T_i evaluated at the sample path $(x_0, \ldots, x_n) \in \Omega$. (We do the same for the random variables T, S_i and H.) Also, for each component type i, we define

$$\begin{aligned} \Omega^i &= \{(x_0, \ldots, x_n) \in \Omega : n \geq 1,\ T_i(x_0, \ldots, x_n) \leq T(x_0, \ldots, x_n)\}, \\ \bar{\Omega}^i &= \{(x_0, \ldots, x_n) \in \Omega : n \geq 1,\ T_i(x_0, \ldots, x_n) > T(x_0, \ldots, x_n)\}. \end{aligned}$$

Furthermore, let $\Omega^i_m = \Omega^i \cap \Omega_m$ and $\bar{\Omega}^i_m = \bar{\Omega}^i \cap \Omega_m$, and note that $\Omega^i = \cup_{m=0}^{\infty} \Omega^i_m$ and $\bar{\Omega}^i = \cup_{m=0}^{\infty} \bar{\Omega}^i_m$. Let $N = \sum_{i=1}^{K} n_i$, which is the total number of components in the system.

Lemma 2 *Under the assumptions of Lemma 1, the following hold:*

(i) $\Omega_0 \neq \emptyset$ *and* $|\Omega_m| \leq |S|^{(m+1)(N+1)}$ *for all* $m \geq 0$.

(ii) *For any path* $(x_0, \ldots, x_n) \in \Omega_m$ *with* $m \geq 0$, $n \leq (m+1)(N+1)$ *and* $P\{(X_0, \ldots, X_n) = (x_0, \ldots, x_n)\} \leq \rho \upsilon^m \epsilon^m$ *for all* $\epsilon > 0$ *sufficiently small, where* ρ *and* υ *are constants which are independent of* $(x_0, \ldots, x_n)$, m, *and* $\epsilon > 0$.

Also, for each component type i,

(iii) $\Omega^i_m = \emptyset$ *for* $m < b_i - b_0$, *and* $\Omega^i_{b_i - b_0} \neq \emptyset$. *For all* $(x_0, \ldots, x_n) \in \Omega^i_{b_i - b_0}$, $T_i(x_0, \ldots, x_n) = 1$, $x_{l-1} \xrightarrow{r} x_l$ *for all* $2 \leq l \leq n$, *and* $S_i(x_0, \ldots, x_n) = \Theta(\epsilon^{-b_i})$. *For any path* $(x_0, \ldots, x_n) \in \Omega^i_m$ *with* $m \geq b_i - b_0$, $|S_i(x_0, \ldots, x_n)| \leq (m+1)\phi\epsilon^{-b_i}$, *where* ϕ *is some constant which is independent of* $(x_0, \ldots, x_n)$, m, *and* $\epsilon > 0$.

(iv) $\bar{\Omega}^i_0 \neq \emptyset$. *For all* $(x_0, \ldots, x_n) \in \bar{\Omega}^i_0$, $x_{l-1} \xrightarrow{r} x_l$ *for all* $2 \leq l \leq n$, *and* $S_i(x_0, \ldots, x_n) = \Theta(\epsilon^{-b_0})$. *For any path* $(x_0, \ldots, x_n) \in \bar{\Omega}^i_m$ *with* $m \geq 0$, $|S_i(x_0, \ldots, x_n)| \leq (m+1)\phi\epsilon^{-b_0}$.

We can prove Lemma 2(i) by constructing a path $(0, x_1, x_2, \ldots, x_n) \in \Omega_0$ whose first transition is a failure transition with transition probability $\Theta(1)$ and the remaining transitions are repair transitions (there may be more than one due to failure propagation). The proof is straightforward and is omitted. Also, the proofs of the other parts are not included since they can be shown with arguments like those used to establish part (i) and Theorems 2 and 3 of Nakayama (1995). Parts (iii)–(iv) of Lemma 2 imply that $\Omega^i = \cup_{m=b_i-b_0}^{\infty} \Omega^i_m$ and $\bar{\Omega}^i = \cup_{m=0}^{\infty} \bar{\Omega}^i_m$.

Proof of Lemma 1. First we prove part (i). Note that

$$\partial_i \xi = \frac{-n_i}{q^2(\epsilon, 0)} + E[H_i] + E[HS_i], \tag{5}$$

where we recall the definition of H_i in (3). Observe that

$$\frac{-n_i}{q^2(\epsilon, 0)} = \Theta(\epsilon^{-2b_0}) \tag{6}$$

since $q(\epsilon, 0) = \Theta(\epsilon^{b_0})$ by (2) and the definition of b_0.

We now analyze the second term on the right-hand side of (5). First, define $\bar{\kappa}(\epsilon) = \max\{n_i(x)/q^2(\epsilon, x) : x \in S,\ x \neq 0\}$, which is $\Theta(1)$ by Assumption **A2** since $|S| < \infty$. Thus, there exists some constant $0 < \kappa < \infty$ such that $\bar{\kappa}(\epsilon) \leq \kappa$ for all ϵ sufficiently small. As a consequence, $\sum_{k=1}^{T-1} n_i(X_k)/q^2(\epsilon, X_k) \leq (T-1)\kappa$ for all sufficiently small $\epsilon > 0$, and so $E\left[\sum_{k=1}^{T-1} n_i(X_k)/q^2(\epsilon, X_k)\right] \leq \kappa E[T-1]$ for all sufficiently small $\epsilon > 0$. Using Lemma 2 and bounding arguments similar to those use in the proof of Theorem 1 of Nakayama (1995), we can show that $E[T] = \Theta(1)$ for all sufficiently small $\epsilon > 0$, which implies that

$$E[H_i] = O(1). \tag{7}$$

We now analyze the third term on the right-hand side of (5). Note that

$$E[HS_i] = E[HS_i;\ T_i \leq T] + E[HS_i;\ T_i > T], \tag{8}$$

where $E[Z;\ J] \equiv E[Z \cdot 1\{J\}]$ for some set of events J. We analyze the two terms on the right-hand side of (8) separately. For the first term, observe that $\{T_i \leq T\} = \Omega^i$, and

$$\begin{aligned}
&E[HS_i;\ T_i \leq T] \\
&= \sum_{\substack{(x_0,\ldots,x_n)\in\Omega^i_{b_i-b_0} \\ n>0}} H(x_0,\ldots,x_n)\ S_i(x_0,\ldots,x_n)\ P\{(X_0,\ldots,X_n) = (x_0,\ldots,x_n)\} \\
&\quad + \sum_{m=b_i-b_0+1}^{\infty} \sum_{\substack{(x_0,\ldots,x_n)\in\Omega^i_m \\ n>0}} H(x_0,\ldots,x_n)\ S_i(x_0,\ldots,x_n)\ P\{(X_0,\ldots,X_n) = (x_0,\ldots,x_n)\} \\
&\equiv M_1 + M_2.
\end{aligned}$$

To analyze M_1, consider any path $(x_0,\ldots,x_n) \in \Omega^i_{b_i-b_0}$, $n > 0$, and note that $x_k \neq 0$ for $0 < k < n$. Define $\bar{\upsilon}(\epsilon) = \max\{1/q(\epsilon, x) : x \in S,\ x \neq 0\}$, which is $\Theta(1)$ by Assumption **A2** since $|S| < \infty$. Thus, there exists some constant $0 < \upsilon < \infty$ such that $\bar{\upsilon}(\epsilon) \leq \upsilon$ for all $\epsilon > 0$ sufficiently small. It then follows that for any path $(x_0,\ldots,x_n) \in \Omega_m$, $m \geq 0$, $H(x_0,\ldots,x_n) = \sum_{k=1}^{n-1}(1/q(\epsilon, x_k)) \leq n\upsilon \leq (m+1)(N+1)\upsilon$ for all sufficiently small $\epsilon > 0$ by Lemma 2(ii), which implies that $H(x_0,\ldots,x_n) = O(1)$ for any path $(x_0,\ldots,x_n) \in \Omega_m$. Lemma 2(iii) states that $S_i(x_0,\ldots,x_n) = \Theta(\epsilon^{-b_i})$ for each $(x_0,\ldots,x_n) \in \Omega^i_{b_i-b_0}$, and $P\{(X_0,\ldots,X_n) = (x_0,\ldots,x_n)\} = \Theta(\epsilon^{b_i-b_0})$ for each $(x_0,\ldots,x_n) \in \Omega^i_{b_i-b_0}$. Hence, since the number of sample paths in $\Omega^i_{b_i-b_0}$ is finite by Lemma 2(i), we have that $M_1 = O(\epsilon^{-b_0})$. Also, using bounding arguments similar to those used in the proof of Theorem 1 of Nakayama (1995), we can use Lemma 2(ii)–(iv) to show that $M_2 = o(\epsilon^{-b_0})$, and so $E[HS_i;\ T_i \leq T] = O(\epsilon^{-b_0})$. We can similarly apply Lemma 2(iii)–(iv) to prove that the second term in (8) satisfies $E[HS_i;\ T_i > T] = O(\epsilon^{-b_0})$, and so $E[HS_i] = O(\epsilon^{-b_0})$. This, along with (6) and (7), establishes part (i).

To prove part (ii), first note that

$$\mathrm{Var}[G_i] = \mathrm{Var}\,[H_i + HS_i] = E[(H_i + HS_i)^2] - (E[H_i + HS_i])^2$$

and that $E[(H_i + HS_i)^2] = E[H_i^2] + E[(HS_i)^2] + 2E[HH_iS_i]$. Following the same line of reasoning used to establish part (i), we can prove that $E[H_i^2] = O(1)$ and $E[H_iHS_i] = O(\epsilon^{-b_0})$. Thus, we need to show that $E[(HS_i)^2] = \Theta(\epsilon^{-b_0-b_i})$, which we do by first noting that $E[(HS_i)^2] = E[(HS_i)^2;\ T_i \le T] + E[(HS_i)^2;\ T_i > T]$. Then by following arguments like those used to prove part (i), we can show that $E[(HS_i)^2;\ T_i \le T] = \Theta(\epsilon^{-b_0-b_i})$ and $E[(HS_i)^2;\ T_i > T] = \Theta(\epsilon^{-2b_0}) = O(\epsilon^{-b_0-b_i})$ since $b_i \ge b_0$. Therefore, $E[(HS_i)^2] = \Theta(\epsilon^{-b_0-b_i})$, and $E[(H_i + HS_i)^2] = \Theta(\epsilon^{-b_0-b_i})$. Combining this with part (i) establishes part (ii).

We omit the proofs of parts (iii) and (iv) since they can be established using the techniques above and in the proof of Theorem 9 of Nakayama (1995). ■

Proof of Theorem 1. We assumed that $\partial_i\xi = \Theta(\epsilon^a)$, $\gamma = \Theta(\epsilon^b)$, $\xi = \Theta(\epsilon^c)$, and $\partial_i\gamma = \Theta(\epsilon^d)$. Then, $\partial_i\mu = (\Theta(\epsilon^{a+b}) - \Theta(\epsilon^{c+d}))/\Theta(\epsilon^{2b}) = \Theta(\epsilon^{\min(a-b,\ c+d-2b)})$ since $c_1c_2 - c_3c_4 \ne 0$ whenever $a + b = c + d$. Because all of the estimators have bounded relative error (see Lemma 1, Shahabuddin 1994, and Nakayama 1995,1996), $\sigma_A^2 = O(\epsilon^{2a})$, $\sigma_B^2 = O(\epsilon^{2b})$, $\sigma_C^2 = O(\epsilon^{2c})$, $\sigma_D^2 = O(\epsilon^{2d})$. Then (4) implies that

$$\begin{aligned}\sigma^2 &= \Theta(\epsilon^{-2b})O(\epsilon^{2a}) + O(\epsilon^{2\min(c+d-3b,\ a-2b)})O(\epsilon^{2b}) + \Theta(\epsilon^{2d-4b})O(\epsilon^{2c}) + \Theta(\epsilon^{2c-4b})O(\epsilon^{2d}) \\ &= O(\epsilon^{\min(2c+2d-4b,\ 2a-2b)}) = O(\epsilon^{2\min(c+d-2b,\ a-b)}),\end{aligned}$$

and the result easily follows. ■

Acknowledgments

This research was partially supported by NJIT SBR Grant 421180. Also, the author would like to thank the Area Editor, Associate Editor, and two anonymous referees for the detailed comments, which improved the quality of the paper.

References

Glynn, P. W. 1990. Likelihood Ratio Derivative Estimators for Stochastic Systems. *Comm. ACM* **33**, 75–84.

Glynn, P. W. and D. L. Iglehart. 1989. Importance Sampling for Stochastic Simulations. *Mgmt. Sci.* **35**, 1367–1393.

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

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

Nakayama, M. K. 1995. Asymptotics of Likelihood Ratio Derivative Estimators in Simulations of Highly Reliable Markovian Systems. *Mgmt. Sci.* **41**, 524–554.

Nakayama, M. K. 1996. General Conditions for Bounded Relative Error in Simulations of Highly Reliable Markovian Systems. *Adv. Appl. Prob.* **28**, 687–727.

Nakayama, M. K., A. Goyal, and P. W. Glynn. 1994. Likelihood Ratio Sensitivity Analysis for Markovian Models of Highly Dependable Systems. *Opns. Res.* **42**, 137–157.

Shahabuddin, P. 1994. Importance Sampling for the Simulation of Highly Reliable Markovian Systems. *Mgmt. Sci.* **40**, 333–352.