Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

EFFICIENT ESTIMATION OF THE MEAN HITTING TIME TO A SET OF A REGENERATIVE SYSTEM

Marvin K. Nakayama

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

Bruno Tuffin

Inria, Univ Rennes, CNRS, IRISA
Campus de Beaulieu, 263 Avenue Général Leclerc
35042 Rennes, FRANCE

ABSTRACT

We consider using simulation to estimate the mean hitting time to a set of states in a regenerative process. A classical simulation estimator is based on a ratio representation of the mean hitting time, using crude simulation to estimate the numerator and importance sampling to handle the denominator, which corresponds to a rare event. But the estimator of the numerator can be inefficient when paths to the set are very long. We thus introduce a new estimator that expresses the numerator as a sum of two terms to be estimated separately. We provide theoretical analysis of a simple example showing that the new estimator can have much better behavior than the classical estimator. Numerical results further illustrate this.

1 INTRODUCTION

Consider a (nondelayed) regenerative process living on a state space $\mathscr{S}$, and let $\mathscr{A}$ be some subset of states. We are interested in estimating the mean μ of the hitting time to $\mathscr{A}$. In many practical settings, the mean hitting time is large, making crude simulation inefficient. For example, this is the case for the *mean time to failure* (MTTF) in a reliability context (e.g., see Goyal et al. 1992), or when examining excessive backlogs in queuing systems (e.g., see Parekh and Walrand 1989, among others). Goyal et al. (1992) and Glynn et al. (2017) consider a classical estimator of μ based on a ratio representation of μ, where the numerator in the ratio is estimated using crude simulation, and the denominator is independently handled via *importance sampling* (IS), which is a variance-reduction technique (VRT) that is well-suited to analyze rare events; see Glynn and Iglehart (1989) and Rubino and Tuffin (2009). Shahabuddin et al. (1988) call this approach *measure-specific importance sampling* (MSIS).

While the classical MSIS estimator can perform well on many models, it can do poorly in some contexts. This can occur when paths in which the set $\mathscr{A}$ is hit before a regeneration are very long (and of not-too-small probability) compared to cycles in which $\mathscr{A}$ is not hit. We thus introduce a new estimator of the mean hitting time that expresses the numerator of μ as a sum of two terms estimated separately using MSIS. Theoretical analysis on a simple model demonstrates that the new estimator can have strictly better performance than the classical estimator. Numerical results further illustrate this.

The rest of the paper proceeds as follows. Section 2 defines the problem. We review the classical estimator of the mean hitting time in Section 3, and Section 4 introduces the new estimator. We provide a theoretical analysis of a simple model in Section 5 to show that the new estimator can have strictly better performance than the classical estimator. Section 6 provides numerical results illustrating the better performance of the new estimator but also that it can be as efficient as the classical one for important classes of models. Some concluding remarks appear in Section 7.

2 PROBLEM DESCRIPTION AND NOTATION

Consider a continuous-time stochastic process $X = [X(t) : t \geq 0]$ evolving on a state space $\mathscr{S}$. We can also handle a discrete-time process $[X_m : m = 0,1,2,\ldots]$ by letting $X(t) = X_{\lfloor t \rfloor}$ for each $t \geq 0$, where $\lfloor \cdot \rfloor$ denotes the floor function. We assume that X is (classically) regenerative, with regeneration times $0 = \Gamma_0 < \Gamma_1 < \Gamma_2 < \cdots$, so the process "probabilistically restarts" at each Γ_i (p. 19 of Kalashnikov 1994). For example, an irreducible discrete-time or continuous-time Markov chain (DTMC or CTMC) on a finite state space is regenerative, with successive hits to a fixed state forming a sequence of regeneration times.

Let $\mathscr{A} \subset \mathscr{S}$ be a subset of states (e.g., "failed states" of a reliability system), and define $T = \inf\{t \geq 0 : X(t) \in \mathscr{A}\}$ as the *hitting time* (or first passage time) to $\mathscr{A}$. Our goal is to estimate

$$\mu = E[T], \tag{1}$$

which is the expected hitting time to $\mathscr{A}$ and is often called the MTTF in reliability settings.

For $i \geq 1$, let $\tau_i = \Gamma_i - \Gamma_{i-1}$, and we call the path segment $[X(\Gamma_{i-1} + s) : 0 \leq s < \tau_i]$ the ith (regenerative) *cycle* of X, which has length τ_i. The regenerative property ensures that $(\tau_i, [X(\Gamma_{i-1} + s) : 0 \leq s < \tau_i])$, $i \geq 1$, is a sequence of independent and identically distributed (i.i.d.) pairs of cycle lengths and cycles. Let τ be a generic copy of τ_i. For $i \geq 1$, let $T_i = \inf\{t \geq 0 : X(\Gamma_{i-1} + t) \in \mathscr{A}\}$ be the time elapsing after Γ_{i-1} until the next hit to $\mathscr{A}$. For $x, y \in \Re$, let $x \wedge y = \min(x, y)$, and define $\mathscr{I}(\cdot)$ as the indicator function, which equals 1 (resp., 0) when its argument is true (resp., false). Because X is regenerative, we have that $(T_i \wedge \tau_i, \mathscr{I}(T_i < \tau_i))$, $i \geq 1$, form an i.i.d. sequence of pairs.

3 CLASSICAL ESTIMATOR OF THE MEAN HITTING TIME

We now review a classical simulation estimator of μ, which has been previously studied in Goyal et al. (1992) and Glynn et al. (2017). Define

$$p = P(T < \tau), \tag{2}$$

and for a random variable Y and event B, let $E[Y;B] = E[Y\,\mathscr{I}(B)]$. Then μ in (1) satisfies

$$\begin{aligned}\mu &= E[T;T<\tau] + E[\tau + T - \tau; T > \tau] = E[T;T<\tau] + E[\tau;T>\tau] + E[T-\tau;T>\tau] \\ &= E[T\wedge\tau;T<\tau] + E[T\wedge\tau;T>\tau] + E[T-\tau \mid T>\tau]P(T>\tau) \;= E[T\wedge\tau] + \mu(1-p),\end{aligned} \tag{3}$$

where $E[T - \tau \mid T > \tau] = \mu$ by the regenerative property. Rearranging leads to representing μ as a ratio

$$\mu = \frac{E[T \wedge \tau]}{p} \equiv \frac{\zeta}{p} \tag{4}$$

where, because $p = E[\mathscr{I}(T < \tau)]$, both the numerator and denominator in (4) are expectations of cycle-based quantities, i.e., expectations of quantities that are measurable with respect to a single cycle.

In practice, it is often the case that hitting $\mathscr{A}$ rarely occurs before τ. Then the denominator p in (4) is a small probability, and estimating *at least* this quantity requires applying a VRT if we want to get an accurate answer in reasonable computational time. This fact motivated Shahabuddin et al. (1988) to exploit (4) to estimate μ via MSIS. The key idea of MSIS is to use crude simulation to estimate expectations that are easy to estimate, and independently apply IS to estimate expectations based on rare events, which make them difficult to estimate with crude simulation. MSIS then allocates a fraction $0 < \beta < 1$ (resp., $1 - \beta$) of the simulation budget to estimate expectations based on non-rare (resp., rare) events. If the budget is the total number n of observations of $(T \wedge \tau, \mathscr{I}(T < \tau))$ to simulate, then we use βn (resp., $(1-\beta)n$) observations to estimate expectations not based on (resp., based on) rare events. We assume that βn is an integer; otherwise, we simulate $\lfloor \beta n \rfloor$ (resp., $\lfloor (1-\beta)n \rfloor$) observations with crude simulation (resp., IS).

In many (but not all) situations, the numerator $\zeta = E[T \wedge \tau]$ in (4) is easy to estimate using crude simulation. To do this, we generate $(T_i \wedge \tau_i, \mathscr{I}(T_i < \tau_i))$, $i = 1, 2, \ldots, \beta n$, as βn i.i.d. copies of $(T \wedge \tau, \mathscr{I}(T <$

$\tau)$) sampled using crude simulation. We then estimate ζ by

$$\widehat{\zeta}_{n,\beta} = \frac{1}{\beta n} \sum_{i=1}^{\beta n} T_i \wedge \tau_i. \tag{5}$$

When the denominator $p = P(T < \tau) = E[\mathscr{I}(T < \tau)]$ of (4) is a rare-event probability, it can be difficult to estimate via crude simulation. To see why, let us examine what happens as $p \to 0$. Define the *relative error* (RE) of an estimator to be its standard deviation divided by the quantity being estimated. The crude estimator $\mathscr{I}(T < \tau)$ of p has variance $p(1-p)$, so the RE of $\mathscr{I}(T < \tau)$ is $\mathrm{RE}[\mathscr{I}(T < \tau)] = \sqrt{p(1-p)}/p = \sqrt{(1-p)/p} \to \infty$ as $p \to 0$. Thus, it becomes more difficult to estimate p as $p \to 0$.

This motivates applying IS to estimate p. To do this, let P denote the probability measure under the original system dynamics, and let P' be another probability measure such that P is absolutely continuous (e.g., p. 422 of Billingsley 1995) with respect to P'. We call P' the *IS measure*, and let E' denote its corresponding expectation operator. Then applying a *change of measure* leads to

$$p = E[\mathscr{I}(T < \tau)] = \int \mathscr{I}(T < \tau)\, \mathrm{d}P = \int \mathscr{I}(T < \tau)\, \frac{\mathrm{d}P}{\mathrm{d}P'}\, \mathrm{d}P' = E'[\mathscr{I}(T < \tau)L], \tag{6}$$

where $L = \mathrm{d}P/\mathrm{d}P'$ is called the *likelihood ratio*. This representation suggests estimating p as follows. Let $(T_i' \wedge \tau_i', \mathscr{I}(T_i' < \tau_i'), L_i')$, $i = 1, 2, \ldots, (1-\beta)n$, be i.i.d. copies of $(T \wedge \tau, \mathscr{I}(T < \tau), L)$ generated using IS, which are independent of the crude sample $(T_i \wedge \tau_i, \mathscr{I}(T_i < \tau_i))$, $i = 1, 2, \ldots, \beta n$. We then estimate p by

$$\widehat{p}_{n,\beta} = \frac{1}{(1-\beta)n} \sum_{i=1}^{(1-\beta)n} \mathscr{I}(T_i' < \tau_i') L_i'. \tag{7}$$

Taking the ratio of the estimators in (5) and (7) yields a classical MSIS estimator of μ as

$$\widehat{\mu}_{n,\beta} = \frac{\widehat{\zeta}_{n,\beta}}{\widehat{p}_{n,\beta}} \tag{8}$$

if $\widehat{p}_{n,\beta} > 0$, and $\widehat{\mu}_{n,\beta} = 0$ if $\widehat{p}_{n,\beta} = 0$. As seen in Goyal et al. (1992), the estimator $\widehat{\mu}_{n,\beta}$ satisfies a *central limit theorem* (CLT). Specifically, assume that $\mathbb{V}\mathrm{ar}[T \wedge \tau] < \infty$ and $\mathbb{V}\mathrm{ar}'[\mathscr{I}(T < \tau)L] < \infty$, where $\mathbb{V}\mathrm{ar}$ (resp., $\mathbb{V}\mathrm{ar}'$) denotes the variance operator under the original (resp., IS) measure. Then the delta method (e.g., Section 3.3 of Serfling 1980) implies that

$$\sqrt{n}[\widehat{\mu}_{n,\beta} - \mu] \Rightarrow N(0, \sigma^2_{\mathrm{C},\beta}) \text{ as } n \to \infty$$

for any fixed $0 < \beta < 1$, where $\Rightarrow$ denotes convergence in distribution (e.g., Section 25 of Billingsley 1995), $N(a, s^2)$ is a normal random variable with mean a and variance s^2, and

$$\sigma^2_{\mathrm{C},\beta} = \frac{1}{p^2} \left(\frac{1}{\beta} \mathbb{V}\mathrm{ar}[T \wedge \tau] + \frac{\mu^2}{(1-\beta)} \mathbb{V}\mathrm{ar}'[\mathscr{I}(T < \tau)L] \right). \tag{9}$$

Here the subscript C in $\sigma^2_{\mathrm{C},\beta}$ denotes that this is the *asymptotic variance* of the classical estimator $\widehat{\mu}_{n,\beta}$.

3.1 Estimators Based on a CPU Budget

Rather than defining the computation budget in terms of the number n of observations of $(T \wedge \tau, \mathscr{I}(T < \tau))$ to simulate, we can instead specify the budget as the total amount b of CPU time available for simulating. For MSIS, fix a constant $0 < \gamma < 1$, and let γb (resp., $(1-\gamma)b$) be the amount of the total CPU budget b allocated to simulating with crude simulation (resp., IS). For each $i = 1, 2, \ldots$, let W_i be the (random) amount

of CPU time to generate $(T_i \wedge \tau_i, \mathscr{I}(T_i < \tau_i))$ using crude simulation. We assume that W_i, $i = 1,2,\ldots$, are i.i.d., which is reasonable because X is regenerative, and assume that $0 < \nu \equiv E[W_i] < \infty$. Let

$$M_\gamma(b) = \sup\left\{k \geq 0 : \sum_{i=1}^{k} W_i \leq \gamma b\right\}, \tag{10}$$

which is the number of crude observations obtained in γb units of CPU time. Then define the crude estimator of the numerator ζ in (4) based on overall CPU budget b and MSIS allocation parameter γ as

$$\widehat{\zeta}_\gamma(b) = \frac{1}{M_\gamma(b)} \sum_{i=1}^{M_\gamma(b)} T_i \wedge \tau_i \tag{11}$$

if $M_\gamma(b) \geq 1$, and let $\widehat{\zeta}_\gamma(b) = 0$ if $M_\gamma(b) = 0$.

Similarly, for each $i = 1,2,\ldots$, let W_i' be the (random) amount of CPU time to generate $(T_i' \wedge \tau_i', \mathscr{I}(T_i' < \tau_i'), L_i')$ using IS. We assume that W_i', $i = 1,2,\ldots$, are i.i.d., and assume that $0 < \nu' \equiv E[W_i'] < \infty$. Let

$$M_\gamma'(b) = \sup\left\{k \geq 0 : \sum_{i=1}^{k} W_i \leq (1-\gamma)b\right\}, \tag{12}$$

which is the number of IS observations obtained in $(1-\gamma)b$ units of CPU time. Then we define the IS estimator of the denominator p in (4) based on overall CPU budget b and allocation parameter γ as

$$\widehat{p}_\gamma(b) = \frac{1}{M_\gamma'(b)} \sum_{i=1}^{M_\gamma'(b)} \mathscr{I}(T_i' < \tau_i') \tag{13}$$

if $M_\gamma'(b) \geq 1$, and let $\widehat{p}_\gamma(b) = 0$ if $M_\gamma'(b) = 0$. Taking the ratio of (11) and (13) leads to the estimator

$$\widehat{\mu}_\gamma(b) = \frac{\widehat{\zeta}_\gamma(b)}{\widehat{p}_\gamma(b)} \tag{14}$$

if $\widehat{p}_\gamma(b) > 0$, and $\widehat{\mu}_\gamma(b) = 0$ if $\widehat{p}_\gamma(b) = 0$. Applying a random-time-change argument (Theorem 14.4 of Billingsley 1999) and the delta method, we can then show that $\sqrt{b}[\widehat{\mu}_\gamma(b) - \mu] \Rightarrow N(0, \bar{\sigma}^2_{\mathrm{C},\gamma})$ as $b \to \infty$ for any fixed $0 < \gamma < 1$, where

$$\bar{\sigma}^2_{\mathrm{C},\gamma} = \frac{1}{p^2}\left(\frac{\nu}{\gamma}\mathbb{V}\mathrm{ar}[T \wedge \tau] + \frac{\nu'\mu^2}{(1-\gamma)}\mathbb{V}\mathrm{ar}'[\mathscr{I}(T < \tau)L]\right). \tag{15}$$

Section III.D of Goyal et al. (1992) derives the optimal value of γ to minimize $\bar{\sigma}^2_{\mathrm{C},\gamma}$ as $\gamma_* = \theta/(1+\theta)$ with $\theta = \sqrt{(\nu\mathbb{V}\mathrm{ar}[T \wedge \tau])/(\mu^2\nu'\mathbb{V}\mathrm{ar}'[\mathscr{I}(T < \tau)L])}$.

4 A NEW ESTIMATOR FOR THE MEAN HITTING TIME

In certain situations, estimating the numerator $\zeta = E[T \wedge \tau]$ in (4) using crude simulation may be inefficient. For example, this can occur when $E[T \wedge \tau; T < \tau]$ is not insignificant compared to $E[T \wedge \tau]$, which Section 5 will demonstrate through a simple example. In this case, we can obtain another representation for μ by replacing $E[T \wedge \tau]$ in (4) with $E[T \wedge \tau; T > \tau] + E[T \wedge \tau; T < \tau]$, as in (3), which leads to

$$\mu = \frac{E[(T \wedge \tau)\mathscr{I}(T > \tau)] + E[(T \wedge \tau)\mathscr{I}(T < \tau)]}{p} \equiv \frac{\phi + \kappa}{p}, \tag{16}$$

where we note that ϕ, κ, and p are all cycle-based expectations.

Now we apply MSIS, with crude simulation (resp., IS) to estimate ϕ (resp., κ and p). Thus,

$$\widehat{\phi}_{n,\beta} = \frac{1}{\beta n} \sum_{i=1}^{\beta n} (T_i \wedge \tau_i)\, \mathscr{I}(\tau_i < T_i) \tag{17}$$

is a crude estimator of ϕ, where $(T_i \wedge \tau_i, \mathscr{I}(T_i < \tau_i))$, $i = 1,2,\ldots,\beta n$, is the same crude sample used to construct $\widehat{\zeta}_{n,\beta}$ in (5). (We could choose $0 < \beta < 1$ here to be different than what was used in Section 3.) We estimate p using $\widehat{p}_{n,\beta}$ from (7). Finally, for κ in (16), we apply a change of measure, as was done in (6), to express $\kappa = E[(T \wedge \tau)\mathscr{I}(T < \tau)] = E'[(T \wedge \tau)\mathscr{I}(T < \tau)L]$. This leads to an IS estimator of κ as

$$\widehat{\kappa}_{n,\beta} = \frac{1}{(1-\beta)n} \sum_{i=1}^{(1-\beta)n} (T_i' \wedge \tau_i')\mathscr{I}(T_i' < \tau_i')L_i' \tag{18}$$

where $(T_i' \wedge \tau_i', \mathscr{I}(T_i' < \tau_i'), L_i')$, $i = 1,2,\ldots,(1-\beta)n$, is the same IS sample used in computing (7). Replacing ϕ, κ, and p in (16) with their respective estimators in (17), (18), and (7) results in a new estimator of μ as

$$\widetilde{\mu}_{n,\beta} = \frac{\widehat{\phi}_{n,\beta} + \widehat{\kappa}_{n,\beta}}{\widehat{p}_{n,\beta}} \tag{19}$$

if $\widehat{p}_{n,\beta} \neq 0$, and $\widetilde{\mu}_{n,\beta} = 0$ otherwise.

Moreover, as in Section 3.1, we can also define another new estimator of μ analogous to that in (19) but instead based on a fixed amount b of CPU time. To do this, define a crude estimator of ϕ as

$$\widehat{\phi}_\gamma(b) = \frac{1}{M_\gamma(b)} \sum_{i=1}^{M_\gamma(b)} (T_i \wedge \tau_i)\mathscr{I}(\tau_i < T_i)$$

if $M_\gamma(b) \geq 1$, and let $\widehat{\phi}_\gamma(b) = 0$ if $M_\gamma(b) = 0$, where $M_\gamma(b)$ is defined in (10) and γ is the MSIS CPU allocation parameter. Also, for IS estimators of p and κ, we use $\widehat{p}_\gamma(b)$ as in (13), and we estimate κ by

$$\widehat{\kappa}_\gamma(b) = \frac{1}{M_\gamma'(b)} \sum_{i=1}^{M_\gamma'(b)} (T_i' \wedge \tau_i')\mathscr{I}(T_i < \tau_i)L_i'$$

if $M_\gamma'(b) \geq 1$, and let $\widehat{\kappa}_\gamma(b) = 0$ if $M_\gamma'(b) = 0$, where $M_\gamma'(b)$ is defined in (12). Then we define

$$\widetilde{\mu}_\gamma(b) = \frac{\widehat{\phi}_\gamma(b) + \widehat{\kappa}_\gamma(b)}{\widehat{p}_\gamma(b)} \tag{20}$$

as another new estimator of μ if $\widehat{p}_\gamma(b) \neq 0$, and let $\widetilde{\mu}_\gamma(b) = 0$ if $\widehat{p}_\gamma(b) = 0$.

We can apply the delta method and a random-time-change argument to establish the following CLT, where the subscript N in $\sigma^2_{\mathrm{N},\beta}$ and $\bar{\sigma}^2_{\mathrm{N},\gamma}$ below denotes that these are the asymptotic variances for the new estimators $\widetilde{\mu}_{n,\beta}$ and $\widetilde{\mu}_\gamma(b)$, respectively.

Proposition 1 If $\eta^2 \equiv \mathbb{V}\mathrm{ar}[(T \wedge \tau)\mathscr{I}(\tau < T)] < \infty$ and $\psi^2 \equiv \mathbb{V}\mathrm{ar}'[((T \wedge \tau) - \mu)\mathscr{I}(T < \tau)L] < \infty$, then

$$\sqrt{n}[\widetilde{\mu}_{n,\beta} - \mu] \Rightarrow N(0, \sigma^2_{\mathrm{N},\beta}) \quad \text{as } n \to \infty$$

for any fixed $0 < \beta < 1$, where

$$\sigma^2_{\mathrm{N},\beta} = \frac{1}{p^2}\left(\frac{\eta^2}{\beta} + \frac{\psi^2}{1-\beta}\right). \tag{21}$$

If in addition $0 < \nu < \infty$ and $0 < \nu' < \infty$, then

$$\sqrt{b}[\widetilde{\mu}_\gamma(b) - \mu] \Rightarrow N(0, \bar{\sigma}^2_{\mathrm{N},\gamma}) \text{ as } b \to \infty$$

for any fixed $0 < \gamma < 1$, where

$$\bar{\sigma}^2_{\mathrm{N},\gamma} = \frac{1}{p^2}\left(\frac{\nu\eta^2}{\gamma} + \frac{\nu'\psi^2}{1-\gamma}\right). \tag{22}$$

Note that $\psi^2 = \mathbb{V}\mathrm{ar}'[(T \wedge \tau)\mathscr{I}(T < \tau)L] + \mu^2\mathbb{V}\mathrm{ar}'[\mathscr{I}(T < \tau)L] - 2\mu\mathbb{C}\mathrm{ov}'[(T \wedge \tau)\mathscr{I}(T < \tau)L, \mathscr{I}(T < \tau)L]$, where $\mathbb{C}\mathrm{ov}'$ denotes covariance under IS, so (21) and (22) capture the dependence between the IS estimators of κ and p. Also, the value of $\bar{\sigma}^2_{\mathrm{N},\gamma}$ in (22) depends on γ, and we can minimize $\bar{\sigma}^2_{\mathrm{N},\gamma}$ using

$$\gamma^* = \frac{\delta}{1+\delta}, \text{ with } \delta = \left(\frac{\nu\mathbb{V}\mathrm{ar}[(T \wedge \tau)\mathscr{I}(\tau < T)]}{\nu'\mathbb{V}\mathrm{ar}'[(T \wedge \tau)\mathscr{I}(T < \tau)L - \mu\mathscr{I}(T < \tau)L]}\right)^{1/2}. \tag{23}$$

5 THEORETICAL ANALYSIS OF A SIMPLE MODEL

Through a theoretical analysis of a simple example, we next show that the new estimator $\widetilde{\mu}_{n,\beta}$ (resp., $\widetilde{\mu}_\gamma(b)$) can strictly outperform its classical counterpart $\widehat{\mu}_{n,\beta}$ (resp., $\widehat{\mu}_\gamma(b)$). We will consider a sequence of models indexed by a rarity parameter ε, and letting $\varepsilon \to 0$ will lead to $p \equiv p_\varepsilon$ from (2) satisfying $p_\varepsilon \to 0$. We should index all variables by a subscript ε, but we leave out the subscript ε to simplify the notation.

Consider a reliability system with $Q+1$ identical components in total. The system starts with all components operational. After a first component fails, it can be repaired, which occurs with probability $1-\varepsilon$. But if a second component fails before the first is repaired, which occurs with probability ε, then the rest of the components deterministically fail one by one until all are failed, bringing the entire system down. Once the system is down, the system resets with all components again operational.

We can model the evolution of the system as a DTMC $X = [X_m : m = 0, 1, 2, \ldots]$ on state space $\mathscr{S} = \{0, 1, 2, \ldots, Q+1\}$, with $\mathscr{A} = \{Q+1\}$, where in state $s \geq 0$, exactly s components are failed. The transition probability matrix $P = [P(s;s') : s, s' \in \mathscr{S}]$ has $P(0;1) = 1$, $P(1;2) = 1 - P(1;0) = \varepsilon$, $P(s;s+1) = 1$ for $2 \leq s \leq Q$, and $P(Q+1;0) = 1$; all other $P(s;s') = 0$. Let regenerations be returns to state 0.

We let $Q = 1/\varepsilon^r$ for some constant $r \geq 0$, and will examine the behavior of our estimators of μ for different r. Setting $r = 0$ leads to a 3-state model, but $r > 0$ results in a large state space when ε is small.

Figure 1 gives the transition diagram of the model, which has only two possibilities for paths of $T \wedge \tau$:

- $0 \to 1 \to 0$, which occurs with probability $1-\varepsilon$; and
- $0 \to 1 \to 2 \to \cdots \to Q+1$, which occurs with probability ε.

For IS, we replace ε in the transition probabilities with a parameter α, which may depend on ε, where we assume that $\alpha \equiv \alpha_\varepsilon \to a \in (0,1)$ as $\varepsilon \to 0$. Specifically, IS uses a transition probability matrix $P' = [P'(s;s') : s, s' \in \mathscr{S}]$, where $P'(1;2) = \alpha$, $P'(1;0) = 1-\alpha$, and $P'(s;s') = P(s;s')$ for all other transitions. Thus, for a path up to $T \wedge \tau$ in which $T < \tau$, the likelihood ratio is $L = \prod_{s=0}^{Q} P(s;s+1)/P'(s;s+1) = \varepsilon/\alpha$.

We will examine the RE of our estimators $\widehat{\mu}_{n,\beta}$ from (8) and $\widetilde{\mu}_{n,\beta}$ from (19), where the RE here uses the square root of the asymptotic variance from its CLT. Thus, the relative error of $\widehat{\mu}_{n,\beta}$ is $\mathrm{RE}[\widehat{\mu}_{n,\beta}] = \sigma_{\mathrm{C},\beta}/\mu$ for $\sigma^2_{\mathrm{C},\beta}$ from (9), and $\mathrm{RE}[\widetilde{\mu}_{n,\beta}] = \sigma_{\mathrm{N},\beta}/\mu$ for $\sigma^2_{\mathrm{N},\beta}$ from (21). We say that an estimator has *bounded* (resp., *vanishing*) RE (BRE, resp., VRE) if the RE remains bounded (resp., vanishes) as $\varepsilon \to 0$.

The numerator of μ in (4) satisfies

$$\zeta = E[T \wedge \tau] = 2(1-\varepsilon) + (Q+1)\varepsilon = 2 + (Q-1)\varepsilon = \Theta(\varepsilon^{\min(0,1-r)}) \tag{24}$$

as $\varepsilon \to 0$ because $Q = \varepsilon^{-r}$, where for functions $f(\varepsilon)$ and $g(\varepsilon)$, the notation $f(\varepsilon) = \Theta(g(\varepsilon))$ denotes that $f(\varepsilon)/g(\varepsilon) \to c$ as $\varepsilon \to 0$ for some constant $c \neq 0$. Note that for $r > 1$, we have that $\zeta \to \infty$ as $\varepsilon \to 0$. The

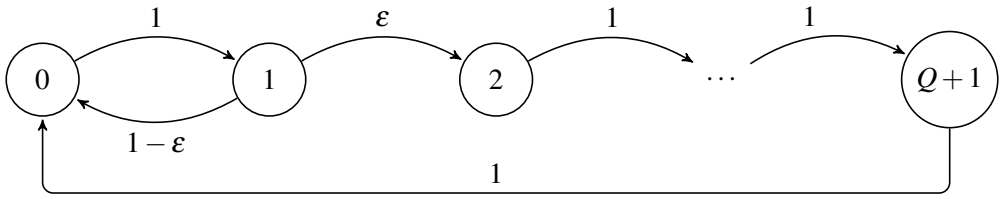

Figure 1: A simple example with $Q+2$ states and only two possibilities for paths of $T \wedge \tau$.

denominator of (4) is $p = \varepsilon$, so

$$\mu = \frac{2+(Q-1)\varepsilon}{\varepsilon} = \Theta(\varepsilon^{-\max(1,r)}). \tag{25}$$

5.1 Asymptotic Behavior of the Classical Estimators $\widehat{\mu}_{n,\beta}$ and $\widehat{\mu}_\gamma(b)$

We first analyze the behavior of the classical estimator $\widehat{\mu}_{n,\beta}$ from (8). For the crude estimator in (5) of the numerator ζ, simple calculations yield $\mathbb{V}\mathrm{ar}[T \wedge \tau] = \varepsilon - \varepsilon^2 - 2Q\varepsilon + Q^2\varepsilon + 2Q\varepsilon^2 - Q^2\varepsilon^2$. As a consequence,

$$\mathbb{V}\mathrm{ar}[T \wedge \tau] = 0 \quad \text{when } r = 0 \tag{26}$$

because then $T \wedge \tau$ is always 2, so the crude estimator of ζ has zero relative error when $r = 0$. But

$$\mathbb{V}\mathrm{ar}[T \wedge \tau] = \varepsilon - \varepsilon^2 - 2\varepsilon^{1-r} + \varepsilon^{1-2r} + 2\varepsilon^{2-r} - \varepsilon^{2-2r} = \Theta(\varepsilon^{1-2r}) \quad \text{when } r > 0 \tag{27}$$

as $\varepsilon \to 0$. Hence, (24) implies that the relative error of the crude estimator of ζ when $r > 0$ is

$$\mathrm{RE}[T \wedge \tau] = \frac{\sqrt{\mathbb{V}\mathrm{ar}[T \wedge \tau]}}{\zeta} = \frac{\Theta(\varepsilon^{(1-2r)/2})}{\Theta(\varepsilon^{\min(0,1-r)})} = \Theta(\varepsilon^{\max(1/2-r,-1/2)}),$$

so as $\varepsilon \to 0$, the RE vanishes for $0 < r < 1/2$, remains bounded for $r = 1/2$, and is unbounded for $r > 1/2$.

The denominator of (4) is $p = \varepsilon = E[\mathscr{I}(T < \tau)] = E'[\mathscr{I}(T < \tau)L]$. For its IS estimator $\widehat{p}_{n,\beta}$, we have

$$\mathbb{V}\mathrm{ar}'[\mathscr{I}(T < \tau)L] = E'[(\mathscr{I}(T < \tau)L)^2] - p^2 = \frac{\varepsilon^2}{\alpha} - \varepsilon^2 = \varepsilon^2\left(\frac{1}{\alpha} - 1\right), \tag{28}$$

so the RE of the IS estimator $\widehat{p}_{n,\beta}$ is $\Theta(1)$ as $\varepsilon \to 0$; i.e., the estimator $\widehat{p}_{n,\beta}$ has BRE as $\varepsilon \to 0$.

We now analyze the asymptotic variance $\sigma^2_{\mathrm{C},\beta}$ in (9) of $\widehat{\mu}_{n,\beta}$. When $r = 0$, we see that

$$\sigma^2_{\mathrm{C},\beta} = \frac{1}{\varepsilon^2}\left[\frac{1}{\beta}0 + \Theta(\varepsilon^{-2\max(1,r)})\frac{\varepsilon^2}{1-\beta}\left(\frac{1}{\alpha} - 1\right)\right] = \Theta(\varepsilon^{-2})$$

by (26) and (28); i.e., $\sigma_{\mathrm{C},\beta} = \Theta(\varepsilon^{-1})$. Hence, (25) implies that

$$\mathrm{RE}[\widehat{\mu}_{n,\beta}] = \Theta(1) \quad \text{when } r = 0 \tag{29}$$

as $\varepsilon \to 0$, so $\widehat{\mu}_{n,\beta}$ has BRE when $r = 0$.

But when $r > 0$, the RE of $\widehat{\mu}_{n,\beta}$ may not be bounded. Putting (27) and (28) into (9) for $r > 0$ yields

$$\sigma^2_{\mathrm{C},\beta} = \frac{1}{\varepsilon^2}\left[\frac{1}{\beta}\Theta(\varepsilon^{1-2r}) + \Theta(\varepsilon^{-2\max(1,r)})\frac{\varepsilon^2}{1-\beta}\left(\frac{1}{\alpha} - 1\right)\right] = \Theta(\varepsilon^{\min[-1-2r,-2\max(1,r)]}),$$

as $\varepsilon \to 0$, so $\sigma_{\mathrm{C},\beta} = \Theta(\varepsilon^{\min[-1-2r,-2\max(1,r)]/2})$. Hence, (25) implies that

$$\mathrm{RE}[\widehat{\mu}_{n,\beta}] = \Theta(\varepsilon^{\min[-1-2r,-2\max(1,r)]/2+\max(1,r)}) = \Theta(\varepsilon^{\min[\max(1/2-r,-1/2),0]}) \quad \text{when } r > 0 \tag{30}$$

as $\varepsilon \to 0$. By (29), we conclude that $\widehat{\mu}_{n,\beta}$ has BRE when $0 \le r \le 1/2$, but its RE $\to \infty$ when $r > 1/2$.

We can also examine the relative error of the estimator $\widehat{\mu}_\gamma(b)$ in (14) based on a CPU time b as the computing budget; i.e., define $\overline{\mathrm{RE}}_b[\widehat{\mu}_\gamma(b)] = \bar{\sigma}_{\mathrm{C},\gamma}/\mu$, where $\bar{\sigma}_{\mathrm{C},\gamma}$ is as in (15). In this case, we assume that the mean CPU time ν (resp., ν') to generate an observation with crude simulation (resp., IS) is

$$\nu = E[T \wedge \tau] = 2(1-\varepsilon) + (Q+1)\varepsilon = 2 - \varepsilon + \varepsilon^{1-r} = \Theta(\varepsilon^{\min(0,1-r)}), \tag{31}$$

$$\nu' = E'[T \wedge \tau] = 2(1-\alpha) + (Q+1)\alpha = 2 - \alpha + \alpha\varepsilon^{-r} = \Theta(\varepsilon^{-r}). \tag{32}$$

For $r = 0$, putting (25), (26), (28), (31), and (32) into (15) and ignoring the $\Theta(1)$ terms yield

$$\bar{\sigma}^2_{\mathrm{C},\gamma} = \frac{1}{\varepsilon^2}\left[\frac{\Theta(\varepsilon^{\min(0,1-r)})}{\gamma}\cdot 0 + \Theta(\varepsilon^{-r})\Theta(\varepsilon^{-2\max(1,r)})\frac{\varepsilon^2}{1-\gamma}\right] = \Theta(\varepsilon^{-2}),$$

so $\overline{\mathrm{RE}}_b[\widehat{\mu}_\gamma(b)] = \Theta(1)$ as $\varepsilon \to 0$ when $r = 0$. If instead $r > 0$, then using (27) rather than (26) leads to

$$\begin{aligned}\bar{\sigma}^2_{\mathrm{C},\gamma} &= \frac{1}{\varepsilon^2}\left[\frac{\Theta(\varepsilon^{\min(0,1-r)})}{\gamma}\Theta(\varepsilon^{1-2r}) + \Theta(\varepsilon^{-r})\Theta(\varepsilon^{-2\max(1,r)})\frac{\varepsilon^2}{1-\gamma}\right] \\ &= \Theta(\varepsilon^{-\max(1+2r,3r)}) + \Theta(\varepsilon^{-\max(2+r,3r)}) = \Theta(\varepsilon^{-\max(2+r,3r)})\end{aligned}$$

because in the penultimate step, the first option in each max is larger if and only if $r < 1$, in which case $1+2r < 2+r$. Hence, we see that $\bar{\sigma}_{\mathrm{C},\gamma} = \Theta(\varepsilon^{-\max(2+r,3r)/2})$. Recall that $\mu = \Theta(\varepsilon^{-\max(1,r)})$ by (25), and note that $\max(1,r) = r$ if and only if $\max(2+r,3r) = 3r$. Thus, for all $r > 0$, we get

$$\overline{\mathrm{RE}}_b[\widehat{\mu}_\gamma(b)] = \Theta(\varepsilon^{-r/2}), \tag{33}$$

which is unbounded as $\varepsilon \to 0$.

5.2 Asymptotic Behavior of the New Estimators $\widetilde{\mu}_{n,\beta}$ and $\widetilde{\mu}_\gamma(b)$

We next derive the RE of the new estimator $\widetilde{\mu}_{n,\beta}$ in (19) for each $r \ge 0$. To calculate the asymptotic variance $\sigma^2_{\mathrm{N},\beta}$ in (21), note that $\eta^2 = \mathbb{V}\mathrm{ar}[(T \wedge \tau)\mathscr{I}(\tau < T)] = 4(1-\varepsilon) - [2(1-\varepsilon)]^2 = 4\varepsilon(1-\varepsilon) = \Theta(\varepsilon)$. Also, by (25), the variance ψ^2 in the second term of (21) satisfies

$$\begin{aligned}\psi^2 &= E'[((T \wedge \tau) - \mu)^2 \mathscr{I}(T < \tau)L^2] - (E[((T \wedge \tau) - \mu)\mathscr{I}(T < \tau)])^2 \\ &= \left(Q + 1 - 2\varepsilon^{-1} - (Q-1)\right)^2 \varepsilon^2/\alpha - \left([Q + 1 - 2\varepsilon^{-1} - (Q-1)]\varepsilon\right)^2 = 4(\alpha^{-1}-1)(1-\varepsilon)^2.\end{aligned}$$

As a consequence, (21) becomes

$$\sigma^2_{\mathrm{N},\beta} = \frac{1}{\varepsilon^2}\left(\frac{1}{\beta}[4\varepsilon(1-\varepsilon)] + 4\frac{\alpha^{-1}-1}{1-\beta}(1-\varepsilon)^2\right) = \Theta(\varepsilon^{-2}) \tag{34}$$

as $\varepsilon \to 0$, giving $\sigma_{\mathrm{N},\beta} = \Theta(\varepsilon^{-1})$. Thus, (25) implies that

$$\mathrm{RE}(\widetilde{\mu}_{n,\beta}) = \Theta(\varepsilon^{\max(0,r-1)}), \tag{35}$$

so as $\varepsilon \to 0$, the new estimator $\widetilde{\mu}_{n,\beta}$ has BRE for $0 \le r \le 1$, and VRE for $r > 1$. In contrast, (29) and (30) show that the RE of the classical estimator $\widehat{\mu}_{n,\beta}$ is bounded for $0 \le r \le 1/2$ and unbounded for $r > 1/2$.

Now consider the relative error of the new estimator $\widetilde{\mu}_\gamma(b)$ in (20) having CPU time b as the computing budget; i.e., $\overline{\mathrm{RE}}_b[\widetilde{\mu}_\gamma(b)] = \bar{\sigma}_{\mathrm{N},\gamma}/\mu$ with $\bar{\sigma}_{\mathrm{N},\gamma}$ as in (22). Modifying (34) to incorporate ν and ν' yields

$$\bar{\sigma}^2_{\mathrm{N},\gamma} = \frac{1}{\varepsilon^2}\left(\frac{\nu}{\gamma}[4\varepsilon(1-\varepsilon)] + 4\nu'\frac{\alpha^{-1}-1}{1-\gamma}(1-\varepsilon)^2\right) = \Theta(\varepsilon^{\min(-1,-r)}) + \Theta(\varepsilon^{-2-r}) = \Theta(\varepsilon^{-2-r})$$

as $\varepsilon \to 0$, by (31) and (32). Hence, we get

$$\overline{\mathrm{RE}}_b[\widetilde{\mu}_\gamma(b)] = \Theta(\varepsilon^{\max(-r/2, r/2-1)}), \tag{36}$$

which is bounded for $r = 0$ and for $r = 2$, unbounded for $0 < r < 2$, and vanishing for $r > 2$. Comparing (36) with (33) for the classical estimator $\widehat{\mu}_\gamma(b)$ based on CPU time, we see that the new estimator outperforms the classical estimator for $r > 1$, and has the same asymptotic exponent when $0 \le r \le 1$.

6 NUMERICAL RESULTS

We next give numerical results for different models simulated for a fixed total number n of observations to obtain the estimators $\widehat{\mu}_{n,\beta}$ in (8) and $\widetilde{\mu}_{n,\beta}$ in (19). We used the following approach to determine the MSIS allocation parameter β for $\widetilde{\mu}_{n,\beta}$ to take into account the expected CPU times ν and ν', where as before, ν is the mean time to generate each of the βn observations of $(T \wedge \tau, \mathscr{I}(T < \tau))$ with crude simulation, and ν' is the mean time to generate each of the $(1-\beta)n$ observations of $(T \wedge \tau, \mathscr{I}(T < \tau), L)$ via IS. Recall that γ^* in (23) is the optimal MSIS CPU allocation parameter to minimize the asymptotic variance $\bar{\sigma}^2_{\mathrm{N},\gamma}$ in (22) of the estimator $\widetilde{\mu}_\gamma(b)$ based on a computation budget of CPU time b, and we want to determine the value of β that corresponds to γ^*. For any large b, we obtain approximately $\gamma^* b/\nu$ (resp., $(1-\gamma^*)b/\nu'$) observations with crude simulation (resp., IS), so we equate $\gamma^* b/\nu$ to βn, yielding $\beta = \gamma^* b/(\nu n)$. The total number of crude and IS observations obtained with CPU budget b and MSIS allocation γ^* is roughly $[\gamma^* b/\nu] + [(1-\gamma^*)b/\nu']$, which we set equal to n. This then leads to $\beta = \nu'\delta/(\nu'\delta + \nu)$, with δ from (23). For the classical estimator $\widehat{\mu}_{n,\beta}$, we applied a similar approach for determining β but using γ_* defined after (15) rather than γ^*. In both cases, we ran a presimulation of 10% of the total number n of observations for each of crude simulation and IS to estimate the unknown quantities in γ^* and γ_*. After computing the resulting value of β from the presimulation, we projected β to lie in $[0.1, 0.9]$ to ensure that the sample sizes are sufficiently large to obtain stable variance estimates.

6.1 The Simple $(Q+2)$-state Example

We first check numerically the gains with the new estimator on the example introduced in Section 5. Table 1 displays the results obtained with $n = 10^6$ observations of $(T \wedge \tau, \mathscr{I}(T < \tau))$ and $\alpha = 0.5$ for IS. The "Rel. Err." column contains estimates of $\sigma_{\mathrm{C},\beta}/\mu$ (resp., $\sigma_{\mathrm{N},\beta}/\mu$), which is the relative error of the estimator $\widehat{\mu}_{n,\beta}$ in (8) (resp., $\widetilde{\mu}_{n,\beta}$ in (19)) based on an observation budget n, where $\sigma^2_{\mathrm{C},\beta}$ and $\sigma^2_{\mathrm{N},\beta}$ are from (9) and (21), respectively. These results confirm the RE theory as $\varepsilon \to 0$ in (30) and (35), even if here γ^* and γ_* depend on ε due to the optimization for each set of parameters. Specifically, as ε shrinks, the relative error of the classical estimator remains bounded for $r = 1/2$, but it increases for $r > 1/2$, as in (30). But for the new estimator, the relative error remains bounded for $r \le 1$ and vanishes for $r > 1$, as in (35).

The behavior changes when we further account for the CPU time. To study this, the last column of Table 1 gives the *work-normalized relative variance* (WNRV), defined as the estimate of $\sigma^2_{\mathrm{C},\beta}/n$ (or $\sigma^2_{\mathrm{N},\beta}/n$) multiplied by CPU time divided by the squared estimator of μ. Hence, $\mathrm{WNRV}^{1/2}$ provides an estimate of the RE based on a CPU budget, as in (33) and (36). Examining WNRV makes sense to analyze the relative precision for a computational budget (L'Ecuyer et al. 2010): having BRE and even VRE may not be sufficient to have small error for a given budget if CPU time increases as $\varepsilon \to 0$. Even if we have

Table 1: Results for the mean time to reach to state $Q+1$ in the example of Section 5.

Estimator	ε	r	Est.	0.95 Confidence Interval	Rel. Err.	CPU (sec)	WNRV
Classical	0.5	3	1.1008e+01	(1.0972e+01, 1.1043e+01)	1.639e+00	1.11e-01	2.98e-07
New	0.5	3	1.1001e+01	(1.0993e+01, 1.1009e+01)	3.641e-01	8.95e-02	1.19e-08
Classical	0.1	3	1.0197e+03	(1.0114e+03, 1.0280e+03)	4.151e+00	3.26e+00	5.62e-05
New	0.1	3	1.0190e+03	(1.0189e+03, 1.0190e+03)	2.501e-02	4.91e+00	3.07e-09
Classical	0.01	3	1.0042e+06	(9.8265e+05, 1.0258e+06)	1.095e+01	1.63e+03	1.95e-01
New	0.01	3	1.0002e+06	(1.0002e+06, 1.0002e+06)	2.604e-04	4.49e+03	3.05e-10
Classical	0.1	2	1.1902e+02	(1.1817e+02, 1.1987e+02)	3.646e+00	3.81e-01	5.06e-06
New	0.1	2	1.1899e+02	(1.1894e+02, 1.1904e+02)	2.095e-01	5.46e-01	2.40e-08
Classical	0.01	2	1.0102e+04	(9.8883e+03, 1.0316e+04)	1.080e+01	1.64e+01	1.91e-03
New	0.01	2	1.0199e+04	(1.0198e+04, 1.0199e+04)	2.549e-02	4.41e+01	2.86e-08
Classical	0.001	2	1.0345e+06	(9.6769e+05, 1.1012e+06)	3.293e+01	1.50e+03	1.63e+00
New	0.001	2	1.0020e+06	(1.0020e+06, 1.0020e+06)	2.610e-03	4.36e+03	2.97e-08
Classical	0.1	1	2.9078e+01	(2.8967e+01, 2.9190e+01)	1.952e+00	9.06e-02	3.45e-07
New	0.1	1	2.9023e+01	(2.8975e+01, 2.9071e+01)	8.373e-01	8.74e-02	6.13e-08
Classical	0.01	1	2.9772e+02	(2.9498e+02, 3.0045e+02)	4.691e+00	2.05e-01	4.51e-06
New	0.01	1	2.9893e+02	(2.9847e+02, 2.9940e+02)	7.926e-01	5.66e-01	3.56e-07
Classical	0.001	1	2.9755e+03	(2.9084e+03, 3.0426e+03)	1.151e+01	1.51e+00	2.00e-04
New	0.001	1	2.9989e+03	(2.9943e+03, 3.0034e+03)	7.703e-01	5.58e+00	3.31e-06
Classical	0.01	0.5	2.0946e+02	(2.0886e+02, 2.1006e+02)	1.459e+00	9.18e-02	1.95e-07
New	0.01	0.5	2.0930e+02	(2.0886e+02, 2.0973e+02)	1.054e+00	9.61e-02	1.07e-07
Classical	$1e-4$	0.5	2.0109e+04	(2.0033e+04, 2.0186e+04)	1.941e+00	3.30e-01	1.24e-06
New	$1e-4$	0.5	2.0099e+04	(2.0058e+04, 2.0141e+04)	1.049e+00	6.94e-01	7.65e-07
Classical	$1e-6$	0.5	1.9994e+06	(1.9952e+06, 2.0035e+06)	1.054e+00	7.37e+00	8.19e-06
New	$1e-6$	0.5	2.0011e+06	(1.9970e+06, 2.0053e+06)	1.054e+00	6.66e+00	7.40e-06

BRE or VRE for a specific r, the WNRV may not be bounded or vanish. Table 1 shows that the WNRV of the new estimator vanishes for $r = 3$, remains bounded for $r = 2$, and increases for $r = 1$ and $r = 1/2$. In contrast, the classical estimator has WNRV that increases as ε shrinks for all $r > 0$. This is again in line with (36) and (33), even if γ^* and γ_* are not fixed here but depend on ε.

6.2 Highly Reliable Markovian System (HRMS)

We next want to compare the two estimators on an HRMS, a type of model extensively studied in the literature. We consider a system with $c = 3$ component types, with redundancy $n_i = 5$ for each type $i = 1,2,3$. Each component has an exponentially distributed time to failure with rate λ_i for components of type i, where $\lambda_i = \varepsilon$, for some parameter ε. Any failed component has an exponentially distributed repair time with rate 1. Component failure and repair times are all independent. The system is down whenever fewer than two components of anyone type are operational.

Table 2 gives results for the MTTF with $n = 10^5$. For IS, we used either the zero-variance approximation (0-var) of L'Ecuyer and Tuffin (2012) or balanced failure biasing (BFB) of Shahabuddin (1994) with $\alpha = 0.8$. It can be checked that the classical and new MTTF estimators give very similar results, and that the results get closer as ε decreases (for $\varepsilon = 0.001$ they are exactly the same up to the 8 first digits). Hence, the new and classical estimators are similar in performance for this type of model. We do not include CPU times and WRNV because computational times are equivalent and do not change as $\varepsilon \to 0$.

6.3 M/M/1 Queue

We also simulated the queue-length process of an M/M/1 queue. Our goal is to estimate the mean time to reach a given buffer size N. We fix the service rate at 1 and consider different values of the arrival rate λ to investigate both the impact of N and traffic intensity $\rho = \lambda/1$. For IS, we swap the arrival and service rates, as in Parekh and Walrand (1989).

Table 2: Results of MTTF estimation for the HRMS with $c = 3$ and $n_i = 5$.

Estimator	ε	Est.	Confidence Interval	Rel. Err.
Classical, 0-var	0.1	3.4090e+02	(3.1811e+02, 3.6370e+02)	1.08e+01
New, 0-var	0.1	3.4007e+02	(3.1766e+02, 3.6248e+02)	1.06e+01
Classical, BFB	0.1	2.9984e+02	(2.6856e+02, 3.3113e+02)	1.68e+01
New, BFB	0.1	2.9949e+02	(2.6899e+02, 3.3000e+02)	1.64e+01
Classical, 0-var	0.01	1.7553e+06	(1.7417e+06, 1.7688e+06)	1.24e+00
New, 0-var	0.01	1.7554e+06	(1.7419e+06, 1.7689e+06)	1.24e+00
Classical, BFB	0.01	1.7412e+06	(1.6790e+06, 1.8033e+06)	5.76e+00
New, BFB	0.01	1.7412e+06	(1.6790e+06, 1.8033e+06)	5.76e+00
Classical, 0-var	0.001	1.6785e+10	(1.6760e+10, 1.6811e+10)	2.46e-01
New, 0-var	0.001	1.6785e+10	(1.6760e+10, 1.6811e+10)	2.46e-01
Classical, BFB	0.001	1.6600e+10	(1.5849e+10, 1.7350e+10)	7.29e+00
New, BFB	0.001	1.6600e+10	(1.5849e+10, 1.7350e+10)	7.29e+00

Table 3 gives results for $n = 10^6$ total number of observations of $(T \wedge \tau, \mathscr{I}(T < \tau))$. Although the new estimator sometimes increases the relative error, it always decreases the WNRV, as can be seen through the *WNRV improvement factor* (WNRVIF), which for the new estimator is defined as the ratio of the WNRV of the new estimator to that of the classical estimator. For example, for $\lambda = 0.9$ and $N = 500$, the new estimator provides an improvement of 66%.

Table 3: Results for the mean time to reach to buffer size N for an M/M/1 queue.

Estimator	λ	N	Est.	0.95 Confidence Interval	Rel. Err.	CPU time (sec)	WNRV	WNRVIF
Classical	0.1	5	1.234e+05	(1.233e+05, 1.235e+05)	4.013e-01	1.019e-01	1.64e-08	
New	0.1	5	1.234e+05	(1.233e+05, 1.235e+05)	3.994e-01	9.135e-02	1.46e-08	1.10
Classical	0.1	10	1.234e+10	(1.233e+10, 1.235e+10)	4.090e-01	1.793e-01	3.00e-08	
New	0.1	10	1.235e+10	(1.234e+10, 1.236e+10)	4.148e-01	1.585e-01	2.73e-08	1.10
Classical	0.1	50	1.234e+50	(1.233e+50, 1.235e+50)	4.673e-01	6.218e-01	1.36e-07	
New	0.1	50	1.235e+50	(1.234e+50, 1.236e+50)	4.745e-01	5.459e-01	1.23e-07	1.10
Classical	0.5	5	1.139e+02	(1.136e+02, 1.143e+02)	1.510e+00	1.234e-01	2.81e-07	
New	0.5	5	1.139e+02	(1.136e+02, 1.142e+02)	1.447e+00	9.878e-02	2.07e-07	1.36
Classical	0.5	10	4.077e+03	(4.063e+03, 4.092e+03)	1.845e+00	1.802e-01	6.13e-07	
New	0.5	10	4.077e+03	(4.062e+03, 4.092e+03)	1.827e+00	1.574e-01	5.25e-07	1.17
Classical	0.5	50	4.491e+15	(4.472e+15, 4.511e+15)	2.228e+00	5.397e-01	2.68e-06	
New	0.5	50	4.491e+15	(4.471e+15, 4.510e+15)	2.230e+00	4.075e-01	2.03e-06	1.32
Classical	0.5	100	5.060e+30	(5.035e+30, 5.084e+30)	2.488e+00	8.645e-01	5.35e-06	
New	0.5	100	5.050e+30	(5.025e+30, 5.076e+30)	2.560e+00	5.734e-01	3.76e-06	1.42
Classical	0.9	50	1.868e+04	(1.843e+04, 1.893e+04)	6.873e+00	8.051e-01	3.80e-05	
New	0.9	50	1.890e+04	(1.865e+04, 1.915e+04)	6.774e+00	6.821e-01	3.13e-05	1.21
Classical	0.9	100	3.766e+06	(3.709e+06, 3.823e+06)	7.734e+00	1.248e+00	7.47e-05	
New	0.9	100	3.737e+06	(3.680e+06, 3.794e+06)	7.785e+00	9.614e-01	5.83e-05	1.28
Classical	0.9	500	7.444e+24	(7.299e+24, 7.589e+24)	9.919e+00	3.649e+00	3.59e-04	
New	0.9	500	7.595e+24	(7.445e+24, 7.745e+24)	1.007e+01	2.136e+00	2.17e-04	1.66

7 CONCLUDING REMARKS

The classical MSIS ratio estimator of the expected hitting time to a set can perform poorly when $E[T \wedge \tau; T < \tau]$ makes a non-negligible contribution to the numerator $E[T \wedge \tau]$ in the ratio formula (4) for μ. We modified the classical estimator to obtain a new estimator, which can perform much better in certain settings, as we showed via a theoretical analysis of a simple model and through numerical experiments. It would be interesting to precisely characterize the types of models for which the new estimator will be more efficient than the classical estimator. Also, we are further investigating CLTs in which simultaneously the CPU time

b grows large and the rarity parameter ε shrinks, as in Proposition 5 of Glynn et al. (2017), which may be more appropriate to study models such as the $(Q+2)$-state model in Section 5.

ACKNOWLEDGMENTS

This work has been supported in part by the National Science Foundation under Grant No. CMMI-1537322. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Billingsley, P. 1995. *Probability and Measure*. 3rd ed. New York: John Wiley and Sons.

Billingsley, P. 1999. *Convergence of Probability Measures*. 2nd ed. New York: John Wiley and Sons.

Glynn, P. W., and D. L. Iglehart. 1989. "Importance Sampling for Stochastic Systems". *Management Science* 35(11):1367–1393.

Glynn, P. W., M. K. Nakayama, and B. Tuffin. 2017. "On the Estimation of the Mean Time to Failure by Simulation". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A.D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 1844–1855. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Goyal, A., P. Shahabuddin, P. Heidelberger, V. Nicola, and P. W. Glynn. 1992. "A Unified Framework for Simulating Markovian Models of Highly Dependable Systems". *IEEE Transactions on Computers* C-41(1):36–51.

Kalashnikov, V. 1994. *Topics on Regenerative Processes*. Boca Raton: CRC Press.

L'Ecuyer, P., J. H. Blanchet, B. Tuffin, and P. W. Glynn. 2010. "Asymptotic Robustness of Estimators in Rare-Event Simulation". *ACM Transactions on Modeling and Computer Simulation* 20(1):6:1–6:41.

L'Ecuyer, P., and B. Tuffin. 2012. "Approximating Zero-Variance Importance Sampling in a Reliability Setting". *Annals of Operations Research* 189(1):277–297.

Parekh, S., and J. Walrand. 1989. "A Quick Simulation Method for Excessive Backlogs in Networks of Queues". *IEEE Transactions on Automatic Control* 34(1):54–66.

Rubino, G., and B. Tuffin. 2009. *Rare Event Simulation using Monte Carlo Methods*. Chichester, UK: John Wiley.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*. New York: John Wiley and Sons.

Shahabuddin, P. 1994. "Importance Sampling for Highly Reliable Markovian Systems". *Management Science* 40(3):333–352.

Shahabuddin, P., V. F. Nicola, P. Heidelberger, A. Goyal, and P. W. Glynn. 1988. "Variance Reduction in Mean Time to Failure Simulations". In *Proceedings of the 1988 Winter Simulation Conference*, edited by M. Abrams, P. Haigh, and J. Comfort, 491–499. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

AUTHOR BIOGRAPHIES

MARVIN K. NAKAYAMA is a professor in the Department of Computer Science at the New Jersey Institute of Technology. He received an M.S. and Ph.D. in operations research from Stanford University and a B.A. in mathematics-computer science from U.C. San Diego. He is a recipient of a CAREER Award from the National Science Foundation, and a paper he co-authored received the Best Theoretical Paper Award for the 2014 Winter Simulation Conference. He is an associate editor for *ACM Transactions on Modeling and Computer Simulation*, and served as the simulation area editor for the *INFORMS Journal on Computing* from 2007–2016. His research interests include simulation, modeling, statistics, risk analysis, and energy. His email address is marvin@njit.edu.

BRUNO TUFFIN received his PhD degree in applied mathematics from the University of Rennes 1 (France) in 1997. Since then, he has been with Inria in Rennes. His research interests include developing Monte Carlo and quasi-Monte Carlo simulation techniques for the performance evaluation of telecommunication systems and telecommunication-related economical models. He is currently Area Editor for *INFORMS Journal on Computing* and Associate Editor for *ACM Transactions on Modeling and Computer Simulation*. He has written or co-written three books (two devoted to simulation): *Rare event simulation using Monte Carlo methods* published by John Wiley & Sons in 2009, *La simulation de Monte Carlo* (in French), published by Hermes Editions in 2010, and *Telecommunication Network Economics: From Theory to Applications*, published by Cambridge University Press in 2014. His email address is bruno.tuffin@inria.fr.