Probability in the Engineering and Informational Sciences, **18**, 2004, 339–368. Printed in the U.S.A.

QUICK SIMULATION METHODS FOR ESTIMATING THE UNRELIABILITY OF REGENERATIVE MODELS OF LARGE, HIGHLY RELIABLE SYSTEMS

MARVIN K. NAKAYAMA
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102
E-mail: marvin@njit.edu

PERWEZ SHAHABUDDIN
Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027
E-mail: perwez.shahabuddin@columbia.edu

We investigate fast simulation techniques for estimating the unreliability in large Markovian models of highly reliable systems for which analytical/numerical techniques are difficult to apply. We first show mathematically that for "small" time horizons, the relative simulation error, when using the importance sampling techniques of failure biasing and forcing, remains bounded as component failure rates tend to zero. This is in contrast to naive simulation where the relative error tends to infinity. For "large" time horizons where these techniques are not efficient, we use the approach of first bounding the unreliability in terms of regenerative-cycle-based measures and then estimating the regenerative-cycle-based measures using importance sampling; the latter can be done very efficiently. We first use bounds developed in the literature for the asymptotic distribution of the time to hitting a rare set in regenerative systems. However, these bounds are "close" to the unreliability only for a certain range of time horizons. We develop new bounds that make use of the special structure of the systems that we consider and are "close" to the unreliability for a much wider range of time horizons. These techniques extend to non-Markovian, highly reliable systems as long as the regenerative structure is preserved.

 0269-9648/04 $16.00

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

1. INTRODUCTION

This article deals with the estimation of the unreliability (i.e., probability of system failure within a given time horizon) in large regenerative models of highly reliable systems. For example, we might be interested in the probability that the computer system aboard a space shuttle fails before the mission end time. Similarly, we might be interested in the probability that the web server used for real-time monitoring and displaying scores in the Olympic Games crashes during the period of the games.

The class of systems we study in this article includes a large class of systems that can be modeled by the System Availability Estimator (SAVE) modeling tool (Blum, Goyal, Heidelberger, Lavenberg, Nakayama, and Shahabuddin [2,3]; also see Goyal, Carter, de Souza e Silva, Lavenberg, and Trivedi [13] and Goyal and Lavenberg [14] for earlier versions). This tool is mainly used for the availability/reliability modeling of computer and communication systems. It models systems consisting of components, each of which fails and gets repaired. The system is considered to be up or down depending on the states of the individual components. All components are highly reliable (i.e., their expected failure times are much larger than their expected repair times). There are a limited number of repairmen in the system. In addition, there are component dependencies and interactions like failure propagation (the failure of a component could cause some other components to fail instantaneously), operational dependencies (the operation of an up component requires some other components to be up), and repair dependencies (the repair of a failed component requires some other components to be up). SAVE considers models for which component failure- and repair-time distributions are exponentially distributed; we will call these highly reliable Markovian systems. The techniques in this article can easily be extended to systems for which the repair times of components have general distributions.

Highly reliable Markovian systems can be modeled as continuous-time Markov chains (CTMCs). However, due to the size of the state space (the state space grows exponentially with the number of components in the system), analytical and numerical (nonsimulation) methods are very difficult. These methods include Gaussian elimination (see, e.g., Young [42]) for computing steady-state measures and uniformization (see, e.g., Gross and Miller [16] and Jensen [19]) for computing transient measures. Hence, approximations have to be used. Although some useful approximations based on lumping and state aggregation have been developed for the steady-state case (see, e.g., Muntz, de Souza e Silva, and Goyal [26]), very few exist for the transient setting. Further complications arise because the embedded Markov chain is *stiff* due to the wide difference between the failure rates and the repair rates of components. In any case, when no exact solutions are easily computable, simulation presents a viable alternative.

In the case of simulation, because system failures are rare, it takes long CPU times for naive simulation to produce good estimates. Importance sampling (see, e.g., Glynn and Iglehart [12] for the basic technique) has been used successfully to increase the speed of simulation in highly reliable Markovian systems (see

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

Alexopoulos and Shultes [1], Carrasco [5, 6], Conway and Goyal [7], Geist and Smotherman [9], Goyal, Shahabuddin, Heidelberger, Nicola, and Glynn [15], Juneja and Shahabuddin [20–22], Lewis and Böhm [25], Obal and Sanders [33], Shahabuddin [36], Shultes [39], and references therein). Surveys can be found in Heidelberger [17], Nakayama [28], Nicola, Shahabuddin, and Nakayama [32], and Shahabuddin [37]. The most widely researched importance-sampling technique for Markovian models is failure biasing. In failure biasing (Lewis and Böhm [25]), we artificially accelerate the occurrence of component-failure events with respect to component-repair events so that the system fails more often. The resulting estimates are then adjusted to make them unbiased. Mathematical analyses of failure-biasing techniques for estimating steady-state measures in highly reliable Markovian systems has been done in Nakayama [27,29], Shahabuddin [36], and Strickland [41]. In [36], it was shown that a failure-biasing heuristic called balanced failure biasing (Goyal et al. [15] and Shahabuddin [36]), produces a bounded relative error (BRE) in the estimation of steady-state measures like the unavailability (the relative error of the estimate remains bounded as the failure rates tend to zero). This is in contrast to naive simulation, for which the relative error tends to infinity. Balanced failure biasing was then implemented in SAVE (Blum et al. [2,3]).

In the case of transient measures, an additional technique that is used is forcing (Lewis and Böhm [25]). In this case, the first component-failure transition is accelerated to make it happen within the time horizon. As has been shown experimentally (Goyal et al. [15] and Shahabuddin [35]), for most transient measures, forcing and failure biasing work well for the case where the time horizon is "small."

When the time horizon is "large," simulation using importance sampling is not efficient. Some work in the large-time-horizon case has been done in Carrasco [5] by numerically inverting a discretized Laplace transform. Another approach suggested in Shahabuddin and Nakayama [38] and Shahabuddin [35] is to bound the transient measure in terms of regenerative-cycle-based measures and then directly estimate (using simulation) the regenerative-cycle-based measures. In highly reliable systems, regenerative cycle-based measures can be estimated with bounded relative error using importance sampling (Shahabuddin [36]). In Shahabuddin [35], such bounds were developed for the expected interval unavailability (the expected fraction of time in $(0, t]$ that the system is down). It was also shown that these bounds converge to the expected interval unavailability as component failure rates tend to zero.

This article has two main contributions. First, we prove that for the case of small time horizons, failure biasing and forcing give bounded relative error in the estimation of the unreliability. A crucial proposition (Proposition 1) proved in this article also allows the bounded-relative-error result to be extended to the small-time-horizon estimation of the expected interval unavailability using the infrastructure already developed by Shahabuddin [35]. Second, we investigate the "bounding approach" mentioned in the previous paragraph, for large-time-horizon estimation of the unreliability. Since a highly reliable Markovian system is regenerative, the time to failure is roughly the "geometric sum" of independent and identically distributed

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

(i.i.d.) random variables, where a "success" in the geometric distribution is defined as failure occurring in a regenerative cycle. It is well known that if the probability of success is small, then this geometric sum is approximately exponentially distributed (see, e.g., Keilson [24] and Solovyev [40]). There is some literature regarding bounds for the cumulative distribution function (which, in our case, is the unreliability) of such random variables (e.g., Brown [4], Kalashnikov [23], and Solovyev [40]). We consider bounds proposed in [4] and [23]. Mathematical analysis reveals that these bounds converge to the unreliability (as the component failure rates tend to zero) only for a certain range of time horizons. We develop new improved bounds that take into consideration the special structure of the type of highly reliable systems we consider; the bounds in [4], [23] and [40] do not make use of this special structure. We show that these new bounds converge to the unreliability for a much wider range of time horizons. We implement these bounds and do some experimental comparisons between our bounds and the bounds in [4]. Preliminary versions of some of the results mentioned earlier appeared (without proofs) in Shahabuddin and Nakayama [38].

In Section 2, we briefly review the Markovian framework considered in Shahabuddin [36] and results for regenerative-cycle-based measures. Section 3 contains asymptotic expressions for the unreliability and the variance of unreliability estimates without and with importance sampling. The bounded-relative-error property of the estimate of the unreliability for the small-time-horizon case is proven here. In Section 3.2, we explain why the efficiency deteriorates for large-time-horizon cases. Then, we consider bounds developed in Brown [4] and determine the range of the time horizons for which these bounds converge. Finally, we develop bounds on the unreliability and investigate their convergence. Section 4 contains experimental results. Difficult proofs of intermediate results are relegated to the Appendix.

2. REVIEW OF HIGHLY RELIABLE SYSTEMS AND FAST SIMULATION

2.1. CTMC Models of Highly Reliable Systems

We further elaborate on the description of the highly reliable Markovian systems presented in Section 1. Let there be R types of components in the system with n_i components of type i. Some of the n_i components can act as spares for that type. Let N be the total number of components of all types in the system (i.e., $N = \sum_{i=1}^{R} n_i$). There are many repairman classes in the system, with a finite number of repairmen in each repairman class. Each component type is assigned a repairman class. The component types assigned to a repairman class are divided into priority classes and repaired in a preemptive or nonpreemptive manner. Any service discipline can be used for members of the same priority class.

Let $\{\mathbf{X}(t) : t \geq 0\}$ be a CTMC model (assume left-continuous with right limits) of a highly reliable Markovian system. For the simplest such system, $\mathbf{X}(t) = (X_1(t), X_2(t), \ldots, X_R(t))$, where $X_i(t)$ is the number of components of type i that are failed. For the general, highly reliable Markovian systems we consider, we have to

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

add some more process descriptors in the state of the system. For example, we might have to add an ordered list of components waiting to be repaired in each repairman class. The following analysis is independent of which definition of $\mathbf{X}(t)$ we use.

Transitions in the CTMC correspond either to a component failing (which may cause the instantaneous failure of other components through failure propagation) or a component completing repair. We will refer to the first as a "failure transition" and to the second as a "repair transition." We label the state with all components up as $\mathbf{1}$. We will also use $\mathbf{1}$ to denote the set of states that contains only the single state $\mathbf{1}$. *Assume that* $\mathbf{X}(0) = \mathbf{1}$ *unless stated otherwise*. Let S be the set of states that are accessible from $\mathbf{1}$. For the purpose of simulation analysis, we will restrict $\{\mathbf{X}(t) : t \geq 0\}$ to the set S. We partition S into two subsets: $S = U \cup F$, where U is the set of up states and $F \neq \emptyset$ is the set of down states. Of course, $\mathbf{1} \in U$ and the state where all components are failed is in F. We will need the following assumptions:

(A.1) The CTMC is irreducible over the set S.
(A.2) From all states in S except $\mathbf{1}$, there is at least one repair transition (with positive probability).
(A.3) From all states in U, there is at least one failure transition (with positive probability).
(A.4) From $\mathbf{1}$, there is at least one failure transition to a state in $U - \mathbf{1}$ (with positive probability).

Let $\mathbf{Q} = \{q(\mathbf{x}, \mathbf{y}), \mathbf{x}, \mathbf{y} \in S\}$ be the rate matrix (also called the infinitesimal generator matrix) of the CTMC. In this matrix, we arrange the states in the order of increasing number of components failed. Thus, the first state is one in which no components are failed (i.e., $\mathbf{1}$). This is followed by states in which exactly one component is failed and so on. Let $q(\mathbf{x}) = -q(\mathbf{x}, \mathbf{x})$ denote the total rate out of state $\mathbf{x}$ and $h(\mathbf{x}) = 1/q(\mathbf{x})$ be the mean holding time in that state. From (A.2) and (A.3), $q(\mathbf{x}) > 0$ for all $\mathbf{x} \in S$. Let Φ denote the probability measure on the sample paths of this CTMC. For any $E \subset S$, let $T_E = \inf\{t > 0 : \mathbf{X}(t-) \notin E, \mathbf{X}(t) \in E\}$. Of particular interest are $T_{\mathbf{1}}$ and T_F.

Let $\{\mathbf{Y}_n : n \geq 0\}$ denote the embedded discrete-time Markov chain (DTMC) of $\{\mathbf{X}(t) : t \geq 0\}$; that is, $\{\mathbf{Y}_n : n \geq 0\}$ has transition matrix $\mathbf{P} = \{P(\mathbf{x}, \mathbf{y}) : \mathbf{x}, \mathbf{y} \in S\}$, where $P(\mathbf{x}, \mathbf{x}) = 0$ and $P(\mathbf{x}, \mathbf{y}) = q(\mathbf{x}, \mathbf{y})/q(\mathbf{x})$ for $\mathbf{x} \neq \mathbf{y}$. One can simulate a CTMC by progressively generating the next state of the embedded DTMC and generating the random holding time in that state. For any $E \in S$, let $\tau_E = \inf\{n \geq 1 : \mathbf{Y}_n \in E\}$. Of particular interest are $\tau_{\mathbf{1}}$ and τ_F.

As is well known, positive-recurrent CTMCs are regenerative (see, e.g., Crane and Iglehart [8] for the definition) in nature. In the following study, we will be considering system regenerations that occur when the system enters state $\mathbf{1}$. In any regenerative cycle, let Z be the random variable denoting the holding time in state $\mathbf{1}$ and let W be the remaining time until either state $\mathbf{1}$ is reached (again) or the system fails. In mathematical terms, $W = \min(T_{\mathbf{1}}, T_F) - Z$, where $T_{\mathbf{1}}$ and T_F are measured from the start of the cycle. We will let $\overline{W}$ (resp., $\underline{W}$) be the random variable having the distribution of W given that a system failure (resp., no system failure) occurs in

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

a regenerative cycle. Another quantity that will be used in the following analysis will be $\gamma \equiv P\{T_F < T_{\mathbf{1}}\} = P(\tau_F < \tau_{\mathbf{1}})$ (i.e., the probability of system failure occurring in a cycle). Assumption (A.3) ensures that γ is not trivially equal to zero. Assumptions (A.2) and (A.4) ensure that $1 - \gamma$ is not equal to zero. For any regenerative-cycle-based random variable, say W, we will use W_i to denote its value in the ith regenerative cycle.

2.2. Modeling Highly Reliable Components

In mathematical models of highly reliable Markovian systems, the failure rate of any component, say component i, is assumed to be of the form $\lambda_i \epsilon^{r_i}$, where ϵ is a small positive parameter called the rarity parameter; r_i and λ_i are positive constants (Shahabuddin [36]; see also Gertsbakh [10]). Let $r_0 = \min\{r_1, \ldots, r_N\} > 0$. Because the repair rates are large compared to failure rates, the repair rate of component i is represented by a constant $\mu_i > 0$. The failure-propagation probabilities are either assumed to be constants or of the same form as the failure rates (i.e., constants multiplied by ϵ raised to positive powers). Once we introduce this ϵ-parameterization, then the performance measures become functions of ϵ. However, for simplicity, we do not specify this dependence in the notation. For example, we continue to use γ for what should ideally be denoted by $\gamma(\epsilon)$. In highly reliable Markovian systems, the aim is to study the performance measures and the variance of their estimators for small ϵ.

This particular ϵ-parameterization guarantees that if $Q(\mathbf{x}, \mathbf{y}) > 0$ (resp., $P(\mathbf{x}, \mathbf{y}) > 0$) for some $\epsilon = \epsilon_0 > 0$, then $Q(\mathbf{x}, \mathbf{y}) > 0$ (resp., $P(\mathbf{x}, \mathbf{y}) > 0$) for all $0 < \epsilon < \epsilon_0$. Given the arrangement of states in $\mathbf{Q}$ that we mentioned earlier, it can easily be seen that all the elements above the diagonal are $O(\epsilon)$ and all elements below the diagonal are $O(1)$. (A function $f(\epsilon)$ is defined to be $O(\epsilon^d)$ (resp., $\underline{O}(\epsilon^d)$), $d \geq 0$, if there exists a constant K such that $|f(\epsilon)| \leq K\epsilon^d$ (resp., $\geq K\epsilon^d$) for all sufficiently small ϵ. A function is said to be $\Omega(\epsilon^d)$, $d \geq 0$, if it is both $O(\epsilon^d)$ and $\underline{O}(\epsilon^d)$; that is, a function is $\Omega(\epsilon^d)$ if it is exactly of order ϵ^d. A function $f(\epsilon)$ is defined to be $o(\epsilon^d)$, $d \geq 0$, if $|f(\epsilon)|/\epsilon^d \to 0$ as $\epsilon \to 0$.) This structure played an important role in the steady-state-simulation analysis of highly reliable Markovian systems in Shahabuddin [36].

One key property of highly reliable Markovian systems that will be used later is the special structure of the regenerative cycle. Because state $\mathbf{1}$ has no repair transitions, $q(\mathbf{1})$ is $\Omega(\epsilon^{r_0})$ since it is the sum of only component failure rates. An important consequence of this is that $E(Z)$ is $\Omega(\epsilon^{-r_0})$. By Assumption (A.2) and the fact that μ_i's are positive constants, we have that $q(\mathbf{x}) = \Omega(1)$ for all $\mathbf{x} \in S - \mathbf{1}$. Finally, as mentioned earlier, repair rates are very large compared to failure rates. Using these three facts, we see that most of the regenerative cycles consist of a long time interval in which the system is in state $\mathbf{1}$, after which there is a single failure transition (which might correspond to more than one component failing because of failure propagation), followed by a short time interval in which the failed

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

components are repaired (and thus the cycle completes). This is the intuition behind the fact shown in Shahabuddin [36] that $E(\min(T_F, T_{\mathbf{1}})) = E(Z + W)$ is of the same order as $E(Z)$ (i.e., $\Omega(\epsilon^{-r_0})$). Using similar methods, we can show that the expected regenerative-cycle time, $E(T_{\mathbf{1}})$, is also $\Omega(\epsilon^{-r_0})$ and that $E(\overline{W})$, $E(\underline{W})$, and $E(\underline{W}^2)$ are $O(1)$. For completeness, the formal proofs are given in the Appendix. It was also shown in [36] that $\gamma = \Omega(\epsilon^r)$, where r is some nonnegative number depending on the structure of the system. Using the fact that the mean time to system failure (MTTF) for regenerative systems can be expressed as $E(\min(T_F, T_{\mathbf{1}}))/\gamma$ (see, e.g., Goyal et al. [15]), we see that the MTTF is $\Omega(\epsilon^{-(r_0+r)})$.

2.3. Importance Sampling for Highly Reliable Systems

Consider the problem of estimating the probability, α, of a rare event $\{X \in A\}$, where X is a random sample path with probability measure $\mathbf{\Phi}$ and A is a rare set. The "naive" way of estimation is to generate samples of X, say $X_1, X_2, \ldots, X_n$, from $\mathbf{\Phi}$ and then form the sample mean $\sum_{i=1}^{n} I(X_i \in A)/n$ where $I(\cdot)$ is the indicator function of the event inside the bracket. The variance of this estimator is $(\alpha - \alpha^2)/n \approx \alpha/n$, as α is small. The expected half-width of the $100(1-\delta)\%$ confidence interval, HW, is proportional to $\sqrt{\alpha/n}$. Thus, the relative error (RE), which is defined by $\mathrm{RE} = HW/\alpha$, is proportional to $\sqrt{1/\alpha n}$. For fixed n, as $\alpha \to 0$, $\mathrm{RE} \to \infty$. This is the problem with naive estimation of the probability of rare events.

In importance sampling, we use a change of measure $\mathbf{\Phi}_{\mathrm{new}}$, such that for each sample path $x \in A$, $\mathbf{\Phi}_{\mathrm{new}}(x) > 0$ if $\mathbf{\Phi}(x) > 0$. Then, we can express α as

$$\begin{aligned}
\alpha &= E(I(X \in A)) \\
&= \int_{x \in A} d\mathbf{\Phi}(x) \\
&= \int_{x \in A} L(x)\, d\mathbf{\Phi}_{\mathrm{new}}(x) \\
&= E_{\Phi_{\mathrm{new}}}(I(X \in A)L(X)), \qquad \textbf{(1)}
\end{aligned}$$

where $L(x) = d\mathbf{\Phi}(x)/d\mathbf{\Phi}_{\mathrm{new}}(x)$ when $d\mathbf{\Phi}_{\mathrm{new}}(x) > 0$ and 0 otherwise. The function L is the Radon–Nikodym derivative; it is also called the likelihood ratio. The subscript in the second expectation operator denotes the new probability measure under which the expectation is taken. Equation (1) forms the basis of the technique of importance sampling. The last expectation term suggests that we use the probability measure $\mathbf{\Phi}_{\mathrm{new}}(\cdot)$ and generate the samples $(X_i, L(X_i))$. Then, a new unbiased estimator is given by

$$\frac{1}{n} \sum_{i=1}^{n} (I(X_i \in A)L(X_i))$$

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

and its variance is

$$\frac{1}{n}\left(E_{\Phi_{\text{new}}}(I(X \in A)L^2(X)) - \alpha^2\right) = \frac{1}{n}\left(E(I(X \in A)L(X)) - \alpha^2\right).$$

The main problem in importance sampling is to find an easily implementable $\mathbf{\Phi}_{\text{new}}$ so that $E(I(X \in A)L(X)) \ll \alpha$, and so the variance of the new estimator is significantly less than that of the naive one.

The most common importance-sampling technique used for highly reliable systems is failure biasing, proposed by Lewis and Böhm [25]. The basic idea behind failure biasing is to make component-failure transitions in the embedded DTMC happen with a probability that is much higher than in the original system. In state $\mathbf{1}$, there are no repair transitions. Therefore, we do not need to failure bias in this state. However, in states that have both failure and repair transitions, the total probability of repair transitions ≈ 1 and the total probability of failure transitions ≈ 0. In such states, the total probability of failure transitions is increased to θ, where θ is some constant (i.e., it is independent of ϵ) between 0 and 1 that is significantly larger than the failure transition probabilities (in practice, θ is typically taken to be 0.5). Therefore, the total probability of repair transitions is decreased to $1 - \theta$. In a version of failure biasing called balanced failure biasing (Goyal et al. [15] and Shahabuddin [36]), the probability of each failure transition (that had positive probabilities in the original system), conditioned on the event that the transition that occurs is a failure transition, is made the same (this is also done in state $\mathbf{1}$). The probability of each repair transition, conditioned on the event that the transition that occurs is a repair transition, is left unchanged. Let $\mathbf{P}'$ be the new transition-probability matrix corresponding to balanced failure biasing. Note that $P'(\mathbf{x}, \mathbf{y}) > 0$ if and only if $P(\mathbf{x}, \mathbf{y}) > 0$.

We will now review the order of magnitude results for the variance associated with importance sampling in the estimation of γ. To obtain a sample of $I(T_F < T_{\mathbf{1}})$, we need only simulate one regenerative cycle of the CTMC. Thus, γ is called a "regenerative-cycle-based measure." Other examples of regenerative-cycle-based measures are $E(W)$, $E(WI(T_F < T_{\mathbf{1}}))$, $E(WI(T_F > T_{\mathbf{1}}))$, $E(T_{\mathbf{1}})$, and $E(\min\{T_F, T_{\mathbf{1}}\})$. Let $\mathbf{\Phi}''$ be a change of measure on the sample paths of the CTMC where we use $\mathbf{P}'$ until time $\min(T_F, T_{\mathbf{1}})$ and then $\mathbf{P}$ after that. For any stopping time τ of the embedded DTMC, define

$$L_\tau = \prod_{i=0}^{\tau-1} \frac{P(\mathbf{Y}_i, \mathbf{Y}_{i+1})}{P'(\mathbf{Y}_i, \mathbf{Y}_{i+1})}.$$

Let $\tau_{\min} = \min(\tau_F, \tau_{\mathbf{1}})$. Then, in the same spirit as Eq. (1), we can write

$$E(I(T_F < T_{\mathbf{1}})) = E_{\Phi''}(I(T_F < T_{\mathbf{1}})L_{\tau_{\min}}).$$

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

The variance of the importance sampling estimator, $\sigma^2_\gamma(\mathbf{\Phi}'') \equiv \mathrm{Var}_{\Phi''}(I(T_F < T_{\mathbf{1}})L_{\tau_{\min}})$, is given by $\sigma^2_\gamma(\mathbf{\Phi}'') = E_{\Phi''}(I(T_F < T_{\mathbf{1}})L^2_{\tau_{\min}}) - \gamma^2$. Let $\sigma^2_\gamma(\mathbf{\Phi})$ be the variance of the naive estimator.

The following theorem was proved in Shahabuddin [36] under the following strengthened version of Assumption (A.4):

(A.4′) For all $\mathbf{x} \in F$, $P(\mathbf{1},\mathbf{x})$ is $o(1)$ (this implies that there exists $\mathbf{x} \in U - \mathbf{1}$ such that $P(\mathbf{1},\mathbf{x})$ is $\Omega(1)$).

THEOREM 1: *Both γ and $\sigma^2_\gamma(\mathbf{\Phi})$ are $\Omega(\epsilon^r)$, where r is a positive constant depending on the structure of the system, the r_i's, and the failure propagation probabilities. Also, $E_{\Phi''}(I(T_F < T_{\mathbf{1}})L^2_{\tau_{\min}})$ is $\Omega(\epsilon^{2r})$ and, so, $\sigma^2_\gamma(\mathbf{\Phi}'')$ is $O(\epsilon^{2r})$.*

This theorem implies that we get a bounded RE in the estimation of γ. One can prove similar bounded RE results for other "rare" regenerative-cycle-based measures like $E(WI(T_F < T_{\mathbf{1}}))$. Moreover, observe that Assumption (A.4′) is not at all restrictive. If it does not hold, then γ is $\Omega(1)$ and we do not have to use importance sampling to estimate it. Unless otherwise stated, in the following sections we will assume that Assumptions (A.1), (A.2), (A.3), and (A.4′) hold.

For transient measures like the unreliability and expected interval unavailability, in addition to failure biasing, we have to use another technique called forcing (Lewis and Böhm [25]). If the time horizon is orders of magnitude less than $E(Z)$, then the system will fail very rarely in $[0, t]$, even though we failure bias. To avoid this, the time of the first event is sampled from the distribution of Z conditioned on the fact that it is less than t. Because Z is exponentially distributed with rate $q(\mathbf{1})$, the time of the first event that we use in the simulation is sampled from the distribution function given by

$$F(s) = \frac{1 - e^{-q(\mathbf{1})s}}{1 - e^{-q(\mathbf{1})t}},$$

where $0 \le s \le t$.

3. ESTIMATION OF THE UNRELIABILITY

Given a finite time horizon t, the unreliability, $U(t)$, is defined to be the probability that the system fails before time t given that it starts in state $\mathbf{1}$; that is,

$$U(t) = P(T_F < t). \tag{2}$$

We wish to estimate the unreliability of the system for different orders of magnitude of the time horizon. A modeling technique that is used in Shahabuddin [35] and Shahabuddin and Nakayama [38] is to represent t as being $\Omega(\epsilon^{-r_t})$, where $r_t \ge 0$, and

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

then model different orders of magnitude of the time horizon by varying r_t. Hence, for $r_t = 0$, t is of the same order as the expected repair times, and for $r_t = r_0$, t is of the same order as the expected first component-failure time in the system (which is of the same order as the expected regenerative cycle time). For $r_t = r_0 + r$, t is of the same order as the MTTF.

In the following subsection, we will prove that a combination of forcing and importance sampling gives bounded RE in the estimation of the unreliability for the case where $r_t = 0$; that is, the time horizon is "small." Before that, we will need to significantly extend the importance-sampling theory developed for "rare" regenerative-cycle-based measures (which was partially reviewed in Section 2.3) to a "nonrare" measure. In particular, the following proposition is crucial to the main result of the next subsection. The proof is technical, so it is deferred to the Appendix.

PROPOSITION 1: $E_{\Phi''}(I(T_F > T_{\mathbf{1}})L^2_{\tau_{\min}}) - 1 = \Omega(1)$.

3.1. The Small-Time-Horizon Case

We can express the unreliability as $U(t) = E(I(A))$, where A is the event $\{T_F < t\}$. Let $\mathbf{\Phi}'$ be the new measure on the sample paths of $\{\mathbf{X}(t) : t \geq 0\}$, in which in each replication we use $\mathbf{P}'$ until the system fails and $\mathbf{P}$ from then on. For the unreliability, for each sample, we only need to simulate the CTMC until either the system fails or the time horizon is exceeded. Let $\mathbf{\Phi}'_{\text{Forcing}}$ be the measure corresponding to both balanced failure biasing and forcing. The variances, without and with balanced failure biasing, are denoted by $\sigma^2_{U(t)}(\mathbf{\Phi})$ and $\sigma^2_{U(t)}(\mathbf{\Phi}')$, respectively. The variance with balanced failure biasing and forcing is denoted by $\sigma^2_{U(t)}(\mathbf{\Phi}'_{\text{Forcing}})$. The following theorem provides the orders of magnitude of the unreliability and the variances of its estimators when the time horizon is small.

THEOREM 2: *Consider the case where* $t = \Omega(\epsilon^{r_t})$ *with* $r_t = 0$. *Then, both* $U(t)$ *and* $\sigma^2_{U(t)}(\mathbf{\Phi})$ *are* $\Omega(\epsilon^{r+r_0})$. *Also,* $\sigma^2_{U(t)}(\mathbf{\Phi}') = O(\epsilon^{2r+r_0})$ *and* $\sigma^2_{U(t)}(\mathbf{\Phi}'_{Forcing}) = O(\epsilon^{2(r+r_0)})$ (r *is the same as in the order of magnitude expression for* γ *in Theorem 1).*

From Theorem 2, we get the following corollary:

COROLLARY 1: *The RE using* $\mathbf{\Phi}'_{Forcing}$ *(corresponding to a fixed* $100(1-\delta)\%$ *level of confidence), for a fixed number n of replications, remains bounded as* $\epsilon \to 0$.

Lemmas 1 and 2 give bounds on $U(t)$ that are used to prove the order of magnitude result for $U(t)$ in Theorem 2. The result for $\sigma^2_{U(t)}(\mathbf{\Phi})$ is a consequence of the fact that $\sigma^2_{U(t)}(\mathbf{\Phi}) = U(t) - U^2(t)$. To simplify notation, let $q \equiv q(\mathbf{1})$. Let K be the random variable denoting the number of times the Markov chain is in state $\mathbf{1}$ (including the first time) before hitting a state in F. Clearly, K has a geometric distribution with parameter (i.e., success probability) γ.

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

LEMMA 1: *Let* $\bar{U}(t) = 1 - e^{-qt\gamma}$. *Then, for all* ϵ *and* t,

$$U(t) \leq \bar{U}(t).$$

PROOF: Let W_{tot} be the total amount of time the CTMC spends in states other than **1** before hitting F. Let $f_{W_{\text{tot}}|k}(\cdot)$ be the density of W_{tot} given that $K = k$ and let $\text{Erlang}(q,k)$ denote an Erlang random variable with rate q and shape parameter k. For a real-valued a, let $(a)^+ = \max(a,0)$. Then,

$$\begin{aligned}
U(t) &= \sum_{k=1}^{\infty} \int_{s=0}^{\infty} P(T_F < t | K = k, W_{\text{tot}} = s) f_{W_{\text{tot}}|k}(s)\, ds (1-\gamma)^{k-1}\gamma \\
&= \sum_{k=1}^{\infty} \int_{s=0}^{\infty} P(\text{Erlang}(q,k) \leq (t-s)^+) f_{W_{\text{tot}}|k}(s)\, ds (1-\gamma)^{k-1}\gamma \\
&\leq \sum_{k=1}^{\infty} \int_{s=0}^{\infty} P(\text{Erlang}(q,k) \leq t) f_{W_{\text{tot}}|k}(s)\, ds (1-\gamma)^{k-1}\gamma \\
&= \sum_{k=1}^{\infty} P(\text{Erlang}(q,k) \leq t)(1-\gamma)^{k-1}\gamma \\
&= \int_{s=0}^{t} qe^{-qs}\gamma \sum_{k=1}^{\infty} \frac{[qs(1-\gamma)]^{k-1}}{(k-1)!}\, ds \\
&= \int_{s=0}^{t} q\gamma e^{-q\gamma s}\gamma\, ds \\
&= 1 - e^{-q\gamma t}.
\end{aligned}$$

■

LEMMA 2: *Let* $q_{min} = \min(q(\mathbf{x}) : \mathbf{x} \in U - \mathbf{1})$. *Then, for all* ϵ *and* t,

$$U(t) \geq \Omega(\epsilon^r)(1 - e^{-qt/k})(1 - e^{-q_{min}t/k})^{k-1},$$

where k *is some constant.*

PROOF: From Shahabuddin [36], it can be shown that there exists a sequence of transitions that start from state **1** and reach a state in F without reentering state **1** such that the product of their probabilities is $\Omega(\epsilon^r)$. All other paths have probability of order $O(\epsilon^r)$. In highly reliable systems terminology (see, e.g., Gertsbakh [10]), this is one of the "most likely paths" to system failure. Let $(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k)$ be the sequence of states visited in one such path, where $\mathbf{x}_0 = \mathbf{1}$ and $\mathbf{x}_k \in F$ and $\mathbf{x}_i \in U - \mathbf{1}$ for $1 \leq i < k$. Then,

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

$$
\begin{aligned}
U(t) &= \sum_{i=1}^{\infty} P\{T_F \le t, \tau_F = i\} \ge P\{T_F \le t, \tau_F = k\} \\
&= \sum_{\mathbf{y}_1 \in U, \ldots, \mathbf{y}_{k-1} \in U, \mathbf{y}_k \in F} P\{\mathbf{Y}_0 = \mathbf{1}, \mathbf{Y}_1 = \mathbf{y}_1, \ldots, \mathbf{Y}_k = \mathbf{y}_k, T_F \le t\} \\
&\ge P\{\mathbf{Y}_0 = \mathbf{1}, \mathbf{Y}_1 = \mathbf{x}_1, \ldots, \mathbf{Y}_k = \mathbf{x}_k, T_F \le t\} \\
&= \int_{t_0=0}^{t} \int_{t_1=0}^{t-t_0} \cdots \int_{t_{k-1}=0}^{t-(t_0+t_1+\cdots+t_{k-2})} P(\mathbf{1}, \mathbf{x}_1) P(\mathbf{x}_1, \mathbf{x}_2) \cdots P(\mathbf{x}_{k-1}, \mathbf{x}_k) \\
&\quad \times q(\mathbf{x}_0) e^{-q(\mathbf{x}_0)t_0} q(\mathbf{x}_1) e^{-q(\mathbf{x}_1)t_1} \cdots q(\mathbf{x}_{k-1}) e^{-q(\mathbf{x}_{k-1})t_{k-1}} \, dt_0 \, dt_1 \cdots dt_{k-1} \\
&\ge \int_{t_0=0}^{t/k} \int_{t_1=0}^{t/k} \cdots \int_{t_{k-1}=0}^{t/k} P(\mathbf{1}, \mathbf{x}_1) P(\mathbf{x}_1, \mathbf{x}_2) \cdots P(\mathbf{x}_{k-1}, \mathbf{x}_k) \\
&\quad \times q(\mathbf{x}_0) e^{-q(\mathbf{x}_0)t_0} q(\mathbf{x}_1) e^{-q(\mathbf{x}_1)t_1} \cdots q(\mathbf{x}_{k-1}) e^{-q(\mathbf{x}_{k-1})t_{k-1}} \, dt_0 \, dt_1 \cdots dt_{k-1} \\
&\ge P(\mathbf{1}, \mathbf{x}_1) P(\mathbf{x}_1, \mathbf{x}_2) \cdots P(\mathbf{x}_{k-1}, \mathbf{x}_k) \\
&\quad \times (1 - e^{-qt/k})(1 - e^{-q(\mathbf{x}_1)t/k}) \cdots (1 - e^{-q(\mathbf{x}_{k-1})t/k}) \\
&\ge \Omega(\epsilon^r)(1 - e^{-qt/k})(1 - e^{-q_{\min} t/k})^k,
\end{aligned}
$$

where the inequality in the fifth line above follows from the fact that the region of integration in the fifth line is contained in that of the previous line. ■

In the same spirit as Eq. (1), we can write $U(t) = E(I(T_F < t)) = E_{\Phi'}(I(T_F < t)L_{\tau_F})$ because we terminate a replication once the failed state is reached. Hence, now we can use $\mathbf{\Phi}'$ and obtain samples of $(I(T_F < t), L_{\tau_F})$. Consequently, if we use balanced failure biasing, the new variance is given by $\sigma^2_{\hat{U}(t)}(\mathbf{\Phi}') \equiv \mathrm{Var}_{\Phi'}(I(T_F < t)L_{\tau_F}) = E_{\Phi'}(I(T_F < t)L^2_{\tau_F}) - U^2(t)$.

Lemma 3, which follows, gives an upper bound on $E_{\Phi'}(I(T_F < t)L^2_{\tau_F})$ which will enable us to evaluate the order of magnitude of $\sigma^2_{\hat{U}(t)}(\mathbf{\Phi}')$. Let $D \equiv E_{\Phi''}(I(T_F < T_{\mathbf{1}})L^2_{\tau_{\min}}) = E(I(T_F < T_{\mathbf{1}})L_{\tau_{\min}})$ and $B \equiv E_{\Phi''}(I(T_F > T_{\mathbf{1}})L^2_{\tau_{\min}}) = E(I(T_F > T_{\mathbf{1}})L_{\tau_{\min}})$, where Φ'' has been defined in Section 2.3. The order of magnitude of D (resp., $B - 1$) is given in Theorem 1 (resp., Proposition 1).

The key to this result lies in the fact that we can decompose the likelihood ratio L_{τ_F} into a product of likelihood ratios over individual cycles. For any sample path with $K = k$, we can write

$$L_{\tau_F} = L^{(1)} L^{(2)} \cdots L^{(k)},$$

where $L^{(i)}$ represents the likelihood ratio over the ith regenerative cycle in the sample path. Given $K = k$, the $L^{(i)}$ are mutually (conditionally) independent, with $L^{(1)}, \ldots, L^{(k-1)}$ having the distribution of $L_{\tau_{\min}}$ given that $\{\tau_{\mathbf{1}} < \tau_F\}$ and $L^{(k)}$ has the distribution of $L_{\tau_{\min}}$ given that $\{\tau_{\mathbf{1}} > \tau_F\}$.

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

LEMMA 3: *For all ϵ and t,*

$$E_{\Phi'}(I(T_F < t)L_{\tau_F}^2) \leq \frac{D}{B-1}(e^{(B-1)qt} - 1). \tag{3}$$

PROOF: In the same spirit as Eq. (1),

$$E_{\Phi'}(I(T_F < t)L_{\tau_F}^2) = E(I(T_F < t)L_{\tau_F})$$

$$= \sum_{k=1}^{\infty} E(I(T_F < t)L_{\tau_F}|K = k)\gamma(1-\gamma)^{k-1}$$

$$= \sum_{k=1}^{\infty} E(I(Z_1 + W_1 + \cdots + Z_k + W_k < t)L_{\tau_F}|K = k)\gamma(1-\gamma)^{k-1}$$

$$\leq \sum_{k=1}^{\infty} E(I(Z_1 + Z_2 + \cdots + Z_k < t)L_{\tau_F}|K = k)\gamma(1-\gamma)^{k-1}. \tag{4}$$

Given $K = k$, $(Z_1 + Z_2 + \cdots + Z_k)$ is (conditionally) independent of L_{τ_F}. Therefore, we can write

$$E(I(Z_1 + Z_2 + \cdots + Z_k < t)L_{\tau_F}|K = k)$$
$$= E(I(Z_1 + Z_2 + \cdots + Z_k < t)|K = k)E(L_{\tau_F}|K = k)$$

and

$$E(L_{\tau_F}|K = k) = E(L^{(1)}L^{(2)}\cdots L^{(k)}|K = k)$$
$$= (E(L_{\tau_{\min}}|\tau_1 < \tau_F))^{k-1}E(L_{\tau_{\min}}|\tau_1 > \tau_F)$$
$$= \left(\frac{B}{(1-\gamma)}\right)^{k-1}\frac{D}{\gamma}.$$

Substituting this in Eq. (4), we obtrain

$$E_{\Phi'}(I(T_F < t)L_{\tau_F}^2) \leq \sum_{k=1}^{\infty} P(\text{Erlang}(q,k) < t)B^{k-1}D.$$

Carrying out the necessary algebra, we get the result of the lemma. ■

We will now describe the contribution of forcing to the reduction in variance. Recall that $\mathbf{\Phi}'_{\text{Forcing}}$ denotes the new measure on the sample paths of $\{\mathbf{X}(t) : t \geq 0\}$ when with balanced failure biasing (i.e., $\mathbf{\Phi}'$) we also use forcing. Then, we have the following lemma.

LEMMA 4: $\sigma^2_{U(t)}(\mathbf{\Phi}'_{Forcing}) = (1 - e^{-qt})E_{\Phi'}(I(T_F \leq t)L_{\tau_F}^2) - U^2(t)$.

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

PROOF: Let L_{Forcing} be the likelihood ratio incurred due to forcing. Note that $L_{\text{Forcing}} = 1 - e^{-qt}$. Then,

$$\begin{aligned}\sigma^2_{U(t)}(\mathbf{\Phi}'_{\text{Forcing}}) &= E_{\Phi'_{\text{Forcing}}}(I(T_F \le t)L^2_{\tau_F}L^2_{\text{Forcing}}) - U^2(t)\\ &= E_{\Phi'}(I(T_F \le t)L^2_{\tau_F}L_{\text{Forcing}}) - U^2(t)\\ &= (1 - e^{-qt})E_{\Phi'}(I(T_F \le t)L^2_{\tau_F}) - U^2(t).\end{aligned}$$ ■

PROOF OF THEOREM 2: From Lemma 1, we get that $U(t)$ is $O(\epsilon^{r+r_0})$. Because q is $\Omega(\epsilon^{r_0})$, we have that $(1 - e^{-qt/k})$ is $\Omega(\epsilon^{r_0})$. Moreover, since $q_{\min}$ is $\Omega(1)$, we have that $(1 - e^{-q_{\min}t/k})$ is also $\Omega(1)$. It then follows that $U(t)$ is $\underline{O}(\epsilon^{r+r_0})$ and we get the first part of the theorem.

From Theorem 1, we see that $D = \Omega(\epsilon^{2r})$. Using Proposition 1 and the fact that $e^x = 1 + x + o(x)$, we get that $e^{(B-1)qt} - 1$ is $\Omega(\epsilon^{r_0})$. Therefore, by Lemma 3, we get that $E_{\Phi'}(I(T_F < t)L^2_{\tau_F})$ is $O(\epsilon^{2r+r_0})$. Hence, $\sigma^2_{U(t)}(\mathbf{\Phi}') = E_{\Phi'}(I(T_F < t)L^2) - U^2(t)$ is $O(\epsilon^{2r+r_0})$. Then, using Lemma 4, we get that

$$\begin{aligned}\sigma^2_{U(t)}(\mathbf{\Phi}'_{\text{Forcing}}) &= (1 - e^{-qt})E_{\Phi'}(I(T_F < t)L^2_{\tau_F}) - U^2(t)\\ &\le qtE_{\Phi'}(I(T_F < t)L^2_{\tau_F}) - U^2(t)\\ &= \Omega(\epsilon^{r_0})E_{\Phi'}(I(T_F < t)L^2_{\tau_F}) - U^2(t)\\ &= \Omega(\epsilon^{r_0})O(\epsilon^{2r+r_0}) - \Omega(\epsilon^{2(r+r_0)})\\ &= O(\epsilon^{2r+2r_0}).\end{aligned}$$ ■

3.2. The Large-Time-Horizon Case

Consider the following generalization of Theorem 2, which gives the orders of magnitude of the variances of our estimators for large time horizons.

THEOREM 3: *Consider the case of large time horizons where* $t = \Omega(\epsilon^{-r_t})$ *with* $0 \le r_t \le r_0$. *Then, both* $U(t)$ *and* $\sigma^2_{U(t)}(\mathbf{\Phi})$ *are* $\Omega(\epsilon^{r+r_0-r_t})$ *and* $\sigma^2_{U(t)}(\mathbf{\Phi}'_{Forcing}) = O(\epsilon^{2(r+r_0-r_t)})$.

The proof of this theorem follows from Lemmas 1–4 and Proposition 1 by substituting $t = \Omega(\epsilon^{r_t})$ in all of the expressions involving t and using the same method as in the proof of Theorem 2.

We saw in the previous section that for small t, for the bounded RE property to hold, the variance reduction using importance sampling had to be of the same order as the unreliability. From Theorem 3, we see that for the case of large time horizons, we also get a variance reduction that is of the same order as the unreliability as long as $r_t \le r_0$.

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

Let

$$\overline{V}(t) = \frac{D}{B-1}\,(e^{(B-1)qt} - 1).$$

We can easily show that for $0 < r_t \leq r_0$, $E_{\Phi'}(I(T_F < t)L^2_{\tau_F})/\overline{V}(t) \to 1$ as $\epsilon \to 0$. We conjecture that this holds even for $r_0 < r_t < r_0 + r_t$. If this conjecture is true, then we have an approximate expression for the variance for the case when ϵ is small. Then, it is easy to see why simulation using importance sampling becomes very inefficient for the case where $r_t > r_0$. This is also explained by the results in Glynn [11]; the variance of the likelihood ratio increases (roughly) exponentially fast in the number of transitions of the Markov chain. In experiments in Nicola, Nakayama, Heidelberger, and Goyal [30], it is shown that even for the case where $r_t = r_0$, one has to tune the value of the failure-biasing parameter θ by trial and error, which can be computationally expensive. Hence, for the case where $r_t \geq r_0$, it is best to use the "bounding approach," as mentioned in Section 1.

Motivated by this, we now investigate bounds on the unreliability. As mentioned in Section 1, because the system is regenerative, the time to failure is roughly a geometric sum of i.i.d. random variables. There is some literature on bounds on the distribution function of such random variables, which, in our case, corresponds to the unreliability. We investigate the bounds given in [4]. For completeness, we first present the bound in its original form. We then adapt it to the reliability model described in this article. Let V_i (generically denoted by V) be i.i.d. nonnegative random variables and let V' be another nonnegative random variable independent of the V_i's. Let $T = \sum_{i=1}^{N_0-1} V_i + V'$, where N_0 is a geometric random variable, independent of the V_i's and V', with probability of "success" p (i.e., $P(N_0 = i) = (1-p)^{i-1}p$ for $i \geq 1$). It is well known in the literature (e.g., Keilson [24] and Solovyev [40]) that as $p \to 0$, T converges in distribution to an exponentially distributed random variable with mean $E(T)$. Here, $E(T) = (1-p)E(V)/p + E(V')$. Theorem 2.2 of Brown [4] gives the following bound:

$$|P(T \leq t) - (1 - e^{-t/E(T)})| \leq \left(\frac{E(V^2)}{E^2(V)} + \frac{E(V')}{E(V)}\,\frac{1}{(1-p)}\right)p. \qquad \textbf{(5)}$$

In our setting, a "success" in the geometric random variable definition corresponds to system failure happening in a regenerative cycle. More specifically, if we define $N_0 \equiv K$, $V_i \equiv Z_i + \underline{W}_i$ for $0 \leq i \leq K-1$, $V' \equiv Z_K + \overline{W}_K$ (recall that $\underline{W}_i$ is the random variable W_i conditioned on no failure event occurring in that cycle and that $\overline{W}_i$ is the random variable W_i conditioned on a failure event occurring in that cycle), and $p \equiv \gamma$, then $T \equiv T_F$. An important point to note is that in our setting, even though W_i is not independent of K, the $\underline{W}_i$, $0 \leq i \leq K-1$, and $\overline{W}_K$ are independent of K. Also,

$$E(T_F) = \frac{(1-\gamma)}{\gamma}\,E(Z + \underline{W}) + E(Z + \overline{W}) = \frac{(1/q) + E(W)}{\gamma}.$$

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

Define $U_b(t) = 1 - e^{-t/E(T_F)}$. Let $\bar{I} \equiv I(T_F < T_1)$ (resp., $\underline{I} \equiv I(T_F > T_1)$) and define

$$\begin{aligned} U_{err,b}(t) &= \left(\frac{E((Z+\underline{W})^2)}{E^2(Z+\underline{W})} + \frac{E(Z+\overline{W})}{E(Z+\underline{W})} \frac{1}{(1-\gamma)} \right) \gamma \\ &= [\gamma(3 - 5\gamma + 2\gamma^2) + E(W\underline{I})q(4\gamma - 2\gamma^2 - 1 + qE(W)) \\ &\quad + E(W^2\underline{I})q^2(\gamma - \gamma^2) - q^2E^2(W\underline{I}) + E(W)q(1-\gamma)] \\ &\quad \times [(1-\gamma)^2 + 2q(1-\gamma)E(W\underline{I}) + q^2E^2(W\underline{I})]^{-1}. \end{aligned} \tag{6}$$

Then, using Eq. (5), we see that upper and lower bounds on $U(t) = P(T_F \le t)$ are given by $\bar{U}_b(t) = U_b(t) + U_{\text{err},b}(t)$ and $\underline{U}_b(t) = U_b(t) - U_{\text{err},b}(t)$, respectively. The bounds are in terms of regenerative-cycle-based measures.

The quality of any upper bound and lower bound combination depends on how close they are to each other. Define the *relative error of the bounds* (this should not be confused with the relative error, RE, in the simulation context, that was defined earlier), REB, to be the difference between the upper bound and the lower bound divided by twice the measure of interest. (The "twice" is motivated by the fact that if we use the arithmetic mean of the upper and lower bound as an approximation for our measure of interest, then the relative error between the approximation and the actual value is always less than REB.) In this case, the REB is given by $\text{REB}_b(t) \equiv (\bar{U}_b(t) - \underline{U}_b(t))/2U(t) = U_{\text{err},b}(t)/U(t)$.

THEOREM 4: *If $r_t > r_0$, then $\underline{U}_{err,b}(t)/U_b(t) \to 0$ as $\epsilon \to 0$; if $r_t = r_0$, then $\underline{U}_{err,b}(t)/U_b(t)$ is $\Omega(1)$; if $r_t < r_0$, then $\underline{U}_{err,b}(t)/U_b(t) \to \infty$ as $\epsilon \to 0$. Hence, $REB_b(t) \to 0$, as $\epsilon \to 0$ if and only if $r_t > r_0$.*

PROOF: We use the representation given in Eq. (6). Assumptions (A.2) and (A.4′) imply that $1 - \gamma$ is $\Omega(1)$. Using the orders of γ, $1 - \gamma$, $E(\underline{W})$, $E(\overline{W})$, $E(\underline{W}^2)$, and $E(Z)$ (see Sect. 2.2), we get that $E(Z + \overline{W}) = \Omega(\epsilon^{-r_0})$, $E(Z + \underline{W}) = \Omega(\epsilon^{-r_0})$, and $E((Z + \underline{W})^2) = \Omega(\epsilon^{-2r_0})$. Consequently, $U_{\text{err},b}(t) = \Omega(\epsilon^r)$. Moreover, since $E(T_F) = \Omega(\epsilon^{-r-r_0})$, $U_b(t) = \Omega(\epsilon^{r+r_0-r_t})$ for $r_t < r + r_0$ and $U_b(t) = \Omega(1)$ for $r_t \ge r + r_0$. The result for all three cases follows from these facts. The last part of the theorem follows from the fact that

$$\frac{U_{\text{err},b}(t)}{U_b(t)} \left(\frac{1}{1 + U_{\text{err},b}(t)/U_b(t)} \right) \le \text{REB}_b(t) \le \frac{U_{\text{err},b}(t)}{U_b(t)} \left(\frac{1}{1 - U_{\text{err},b}(t)/U_b(t)} \right).$$

Therefore, $\text{REB}_b(t) \to 0$ if and only if $U_{\text{err},b}(t)/U_b(t) \to 0$. ■

We also tried using the bounds in Kalashnikov [23], but there seems to be some error in the bounds. Moreover, as mentioned earlier, the bounds do not make use of the special structure of the regenerative cycles; that is, they do not make use of the fact that $\min\{T_1, T_F\} = Z + W$, where Z is exponentially distributed and $E(Z) \gg E(W)$. The following theorem provides bounds that make use of this special structure.

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

THEOREM 5:

(a) Let $\bar{U}(t) = 1 - e^{-q\gamma t}$. *Define* $l \equiv l(t,q) = \max(\sqrt{t}, t\sqrt{q})$ *and let* $\underline{U}(t) \equiv \bar{U}(t) - U_{err}(t)$, *where*

$$U_{err}(t) \equiv \left(e^{-\gamma q(t-l)} - e^{-\gamma qt} + \frac{E(W)}{\gamma l}(1 - e^{-\gamma q(t-l)}) - \frac{q(t-l)E(WI(T_1 < T_F))}{l} e^{-\gamma q(t-l)} \right). \quad (7)$$

Then, $\underline{U}(t) \leq U(t) \leq \bar{U}(t)$ *for all* ϵ *and* t.

(b) Let $REB(t)$ *denote the REB in this case. For* $r_t > 0$, $REB(t) \equiv (\bar{U}(t) - \underline{U}(t))/2U(t) = U_{err}(t)/2U(t) \to 0$ *as* $\epsilon \to 0$.

Remark 1: These bounds are in terms of regenerative-cycle-based measures.

Remark 2: These bounds converge for a much wider range of the time horizon as compared to the range in Theorem 4.

4. EXPERIMENTAL RESULTS WITH A LARGE MARKOVIAN MODEL

We took an example of a large computing system originally considered in Goyal et al. [15] and subsequently in many other articles. The system is depicted in Figure 1. It consists of two sets of processors with two processors per set, two sets of disk controllers with two controllers per set, and six clusters of disks with four disks per cluster. In a disk cluster, data are replicated so that we can have one disk fail without affecting the system. The failure rates of processors, controllers, and disks are assumed to be $\frac{1}{2000}$, $\frac{1}{2000}$, and $\frac{1}{6000}$ per hour, respectively. If a processor of a given

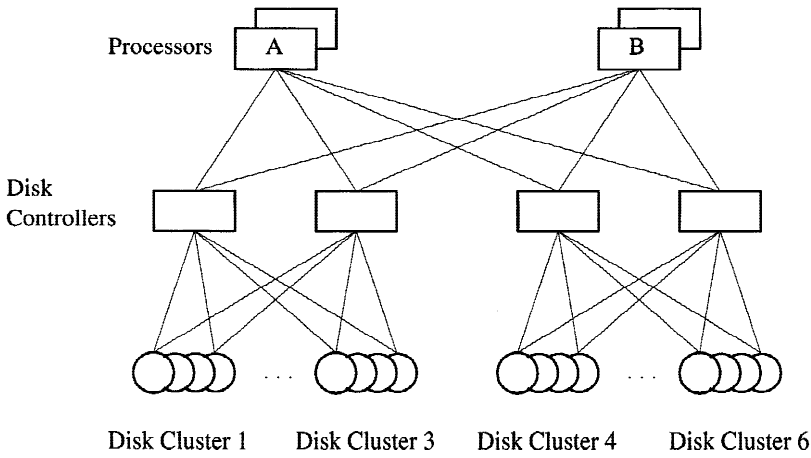

FIGURE 1. Block diagram of the computing system modeled.

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

set fails, it has a 0.01 probability of causing the operating processor of the other set to fail. Each unit in the system can fail in one of two failure modes which occur with equal probability. The repair rates for all mode 1 and all mode 2 failures are 1 per hour and $\frac{1}{2}$ per hour, respectively. The system is defined to be operational if all data are accessible to both processor types, which means that at least one processor in each set, one controller in each set, and three out of four disks in each of the six disk clusters are operational. We also assume that operational components continue to fail at the given rates when the system failed.

To keep the state space within manageable limits (in order to facilitate comparison with approximate numerical results from SAVE), Goyal et al. [15] assumed that after each transition (whether failure or repair), the repairman picks a component at random from the set of failed components. In this way, the state variable does not have to include the order of components waiting at the repair queue. For the purpose of comparison, we first use the same repair discipline. Later, we also consider the same example with first-come first-served (FCFS), where the numerical methods implemented in SAVE cannot be used.

In order to see the effect of our simulation schemes, we estimate the unreliability for different values of the time horizon using different techniques. The results are presented in Table 1. The time horizon is given in the first column. This model has 1,265,625 states and is thus very difficult to solve by exact techniques. Goyal et al. [15] used SAVE to numerically compute approximations for values of the time horizon up to 1024, but give no bounds on the approximation error. We used SAVE to complete these computations for the other time horizons. All these are reproduced in column 2. The third column gives the estimate and the RE corresponding to 99% confidence intervals (CIs) if we use naive simulation. The fourth column gives the estimate and the RE using failure biasing and forcing. Each of the naive and importance-sampling-simulation cases were simulated for 400,000 events. In the fifth and sixth columns, we estimate the bounds of Brown [4] mentioned earlier (henceforth referred to as M.B. bounds). We do this by running 1 simulation of 400,000 events and estimating γ, $E(W)$, $E(WI(T_1 < T_F))$, and $E(W^2 I(T_1 < T_F))$. We then used them to compute the bounds for *all t*. These regenerative-cycle-based measures can be estimated using the dynamic importance sampling (DIS) approach with balanced failure biasing as described in Goyal et al. [15]. For building confidence intervals, we use the delta method (e.g., see Serfling [34, p. 124]) to establish a central limit theorem.

Next, we estimate the bounds which we developed. In this case, one has to first estimate the regenerative-cycle-based measures γ, $E(W)$, and $E(WI(T_1 < T_F))$. The last two columns of Table 1 give the experimental results using these new bounds. We again use 1 simulation run of 400,000 events to estimate the bound for all t. Compared to the M.B. bounds, it is simpler to build confidence intervals in this case, as there are fewer measures to be estimated and the expressions are less complicated.

In this example, $E(T_1) \approx 125$ and $E(T_F) \approx 152{,}240$. For time horizons that are significantly larger than 125, one should not expect balanced failure biasing and forcing to work well. One can observe in Table 1 that for time horizon 1024 and

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

TABLE 1. Estimates of the Unreliability, Brown's [4] Bounds, and the New Bounds

t	Approx. (SAVE) ($\times 10^{-3}$)	Naive Sim. Est. and RE ($\times 10^{-3}$)	Imp. Samp. Est. and RE ($\times 10^{-3}$)	M.B. LB Est. and RE ($\times 10^{-3}$)	M.B. UB Est. and RE ($\times 10^{-3}$)	New LB Est. and RE ($\times 10^{-3}$)	New UB Est. and RE ($\times 10^{-3}$)
4	0.0153	0.0184 ± 97%*	0.0154 ± 4%	−2.3 ± 3.7%	2.34 ± 3.7%	0.002 ± 29%	0.025 ± 3.7%
16	0.0873	0.0871 ± 49%*	0.0902 ± 3.8%	−2.2 ± 3.7%	2.42 ± 3.7%	0.04 ± 4.8%	0.10 ± 3.7%
64	0.380	0.417 ± 28%*	0.381 ± 4.3%	−1.9 ± 3.7%	2.71 ± 3.7%	0.258 ± 3.9%	0.399 ± 3.7%
256	1.55	1.71 ± 22%	1.58 ± 5.6%	−0.7 ± 3.6%	3.89 ± 3.7%	1.25 ± 3.7%	1.59 ± 3.7%
1,024	6.23	6.28 ± 21%	6.22 ± 21%*	3.96 ± 3.7%	8.60 ± 3.7%	5.32 ± 3.7%	6.36 ± 3.7%
2,048	12.4	15.1 ± 19%	10.6 ± 37%*	10.2 ± 3.7%	14.9 ± 3.7%	10.7 ± 3.7%	12.7 ± 3.7%
4,096	24.9	24.2 ± 21%	9.29 ± 38%*	22.6 ± 3.7%	27.2 ± 3.7%	21.4 ± 3.7%	25.2 ± 3.7%
8,192	47.8	52.5 ± 20%	14.7 ± 118%*	46.9 ± 3.6%	51.5 ± 3.6%	42.5 ± 3.6%	49.8 ± 3.6%
16,384	95.3	96.1 ± 20%	10.6 ± 46%*	93.6 ± 3.5%	98.3 ± 3.5%	83.2 ± 3.6%	97.1 ± 3.5%

Note: The asterisk indicates that the estimate of the quantity and/or its variance was highly unstable.

beyond, the estimates of the unreliability and/or their REs, using balanced failure biasing and forcing, become highly unstable. In fact, we suspect infinite variance, either in the estimation of the quantity or its RE, in this range of the time horizon; that is, no matter how long we run the simulation, we will never get stability. The estimates of the M.B. bounds are "satisfactory" (we define this to be less than 10% REB) only for time horizon 4096 and beyond. However, for all time horizons where failure biasing and forcing do not work well (i.e., time horizon 1024 and beyond), the estimates of the new bounds are satisfactory. For very large values of t, estimates using M.B. bounds do better than those using the new ones. Hence, these could be used for better accuracy for the higher time ranges.

To verify the robustness of our methods for simulating rare events, we consider a more "rare" case where all the failure rates and the failure propagation probabilities are reduced by a factor of 100. We also use the FCFS service discipline. In this case, $E(T_1) \approx 12{,}500$ and $E(T_F) \approx 16.44 \times 10^8$. As earlier, a total of 400,000 events were simulated for each case. The results are presented in Table 2. In this case, importance sampling starts to become unstable from time horizons 10^5 and beyond, and we note the same relative trends in the bounds as we did earlier. In addition, the RE using naive simulation goes up about 10 times from the previous less rare case; however, the relative errors in the simulation of the bounds are almost unchanged. All of the (estimated) REs in the estimates of the bounds, except for $t = 10$, ranged from 3.7% to 3.8%.

In these experiments, we used balanced failure biasing to estimate the regenerative-cycle-based measures. As mentioned earlier, this is guaranteed to produce a bounded RE in the estimation of these measures. However, other failure-biasing schemes that are efficient in practice (e.g., the failure distance scheme of Carrasco [5]) can also be used in the estimation of these regenerative-cycle-based measures.

5. DISCUSSION AND OPEN PROBLEMS

In this article, we discussed the estimation of the unreliability in large Markovian models of highly reliable systems. We show that for small time horizons, simulation using the importance-sampling techniques of failure biasing and forcing are provably effective. For large time horizons, simulation using the above techniques becomes difficult. In this case, the approach used is to first bound the unreliability in terms of regenerative-cycle-based measures and then estimate the regenerative-cycle-based measures using importance sampling. We explore bounds existing in the literature and develop some bounds of our own.

For the small-time-horizon case ($r_t = 0$), this work complements the work in Heidelberger, Shahabuddin, and Nicola [18] and Nicola et al. [30], which deals with the estimation of transient measures and proving corresponding BRE results for non-Markovian reliability models. However, the frameworks used for simulation and thus importance sampling in the non-Markovian setting are very different from those used for Markovian models. In particular, a discrete-event-simulation

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

TABLE 2. Estimates of the Unreliability and the Bounds for Another Model

t	Naive Sim. Est. and RE	Imp. Samp. Est. and RE	M.B. LB Est.	M.B. UB Est.	New LB Est.	New UB Est.
10	N/A	5.02E − 9 ± 3.9%	−2.27E−5	2.27E−5	1.90E−9	6.05E−9
10^2	N/A	6.03E − 8 ± 4.1%	−2.26E−5	2.28E−5	4.51E−8	6.05E−8
10^3	N/A	5.98E − 7 ± 4.2%	−2.21E−5	2.33E−5	5.53E−7	6.05E−7
10^4	N/A	5.94E − 6 ± 4.7%	−1.66E−5	2.87E−5	5.85E−6	6.05E−6
10^5	4.27E − 5 ± 257%*	4.88E − 5 ± 12%*	3.78E−5	8.32E−5	5.95E−5	6.05E−5
10^6	4.04E − 4 ± 257%*	6.86E − 5 ± 16%*	5.82E−4	6.28E−4	5.96E−4	6.05E−4
10^7	7.97E − 3 ± 181%*	8.65E − 5 ± 36%*	6.01E−3	6.01E−3	5.94E−3	6.03E−3
10^8	7.41E − 2 ± 175%*	8.52E − 5 ± 35%*	5.87E−2	5.87E−2	5.78E−2	5.87E−2

Note: All of the (estimated) REs in the estimates of the bounds, except for $t = 10$, ranged from 3.7% to 3.8%. The asterisk indicates that the estimate of the quantity and/or its variance was highly unstable. The N/A indicates that no samples of system failure before time t were obtained.

approach and/or uniformization is used in [18] and [30], whereas the CTMC approach is recommended and used for large Markovian models (e.g., SAVE). As was the case in this article, the *direct* importance-sampling approaches described in [18] and [30] do not work when the time horizon is large.

For the large-time-horizon non-Markovian case, we can easily extend the bounding approach in this article, for the case when failure times are exponentially distributed but repair times are generally distributed. This is because the regenerative property is still preserved and one can use the techniques in Nicola et al. [30] and Nicola, Shahabuddin, Heidelberger, and Glynn [31] to estimate the regenerative-cycle-based measures efficiently. However, results on the convergence of the bounds to the actual measure, although intuitively apparent, are difficult to prove rigorously. The development of efficient large-time-horizon simulation techniques in models for which both the failure times and repair times are nonexponentially distributed is still an open problem, because, then, the regenerative structure is lost.

It should be mentioned that even though we use balanced failure biasing for estimating the transient measures (for small time horizons) and the regenerative-cycle-based measures, one could also have used balanced failure transition distance biasing (Carrasco [5,6]) or the balanced likelihood ratio method (Alexopoulos and Shultes [1] and Shultes [39]). Balanced failure transition distance biasing uses structural information about the system to failure bias and, thus, tends to produce more accurate estimates. However, there is an implementation overhead that comes with using more information that is not present in the case of balanced failure biasing. Bounded relative error results for regenerative-cycle-based measures in the case of balanced failure transition distance biasing follow as straightforward extensions of the work in Shahabuddin [36] (see, e.g., Nakayama [27,29] and Nicola et al. [32]); we expect the same to be true for transient measures in the case of small time horizons. The balanced likelihood ratio method has empirically been shown to work better than balanced failure biasing on systems with larger redundancy (minimum number of component failures required for the system to fail) than those originally considered in SAVE (Blum et al. [2,3]). Bounded relative error results for estimating regenerative-cycle-based measures using the balanced likelihood ratio method, and an extension of the technique that uses structural information about the system are also described in Alexopoulos and Shultes [1] and Shultes [39].

Acknowledgments

The work of the first author was supported by National Science Foundation (NSF) CAREER Award DMI-96-24469 and NSF grant DMI-99-00117. The work of the second author was supported in part by NSF grant DMS-95-08709, NSF CAREER Award DMI-96-25291, and a 1998 IBM University Partnership Program Award. The authors would like to thank Chia-wei Hu for computational assistance.

References

1. Alexopoulos, C. & Shultes, B.C. (2001). Estimating reliability measures of highly-dependable Markovian systems, using balanced likelihood ratios. *IEEE Transactions on Reliability* 50: 265–280.
2. Blum, A., Goyal, A., Heidelberger, P., Lavenberg, S.S., Nakayama, M.K., & Shahabuddin, P. (1994). Modeling and analysis of system dependability using the System Availability Estimator. In M. Malek,

D.S. Fussel, A. Dahbura, & T. Nanya (eds.), *Digest of papers: The twenty-fourth annual international symposium on fault-tolerant computing*. New York: IEEE Press, pp. 137–141.
3. Blum, A., Heidelberger, P., Lavenberg, S.S., Nakayama, M.K., & Shahabuddin, P. (1993). System Availability Estimator (SAVE) language reference and user's manual version 4.0. Research Report RA 219 S, IBM T.J. Watson Research Center, Yorktown Heights, NY.
4. Brown, M. (1990). Error bounds for exponential approximations of geometric convolutions. *The Annals of Probability* 18(3): 1388–1402.
5. Carrasco, J.A. (1991). Efficient transient simulation of failure/repair Markovian models. In V.K. Agarwal & E. Cerny (eds.), *Proceedings of the tenth symposium on reliable distributed systems*. New York: IEEE Press, pp. 152–161.
6. Carrasco, J.A. (1992). Failure distance-based simulation of repairable fault-tolerant systems. In G. Balbo & G. Serazzi (eds.), *Computer performance evaluation: Modelling techniques and tools*, pp. 351–366. Amsterdam: Elsevier.
7. Conway, A.E. & Goyal, A. (1987). Monte Carlo simulation of computer system availability/reliability models. In J.P. Shen, D.P. Siewiorek, F. Cristian, & J. Goldberg (eds.), *Proceedings of the seventeenth symposium of fault-tolerant computing*. New York: IEEE Press, pp. 230–235.
8. Crane, M.A. & Iglehart, D.L. (1975). Simulating stable stochastic systems, III: Regenerative processes and discrete event simulations. *Operations Research* 23(1): 33–45.
9. Geist, R.M. & Smotherman, M. (1989). Ultrahigh reliability estimates through simulation. In *Proceedings of the annual reliability and maintainability symposium*. New York: IEEE Press, pp. 350–355.
10. Gertsbakh, I.B. (1984). Asymptotic methods in reliability theory: A review. *Advances in Applied Probability* 16: 147–175.
11. Glynn, P.W. (1994) Importance sampling for Markov chains: Asymptotics for the variance. *Stochastic Models* 10: 701–717.
12. Glynn, P.W. & Iglehart, D.L. (1989). Importance sampling for stochastic simulations. *Management Science* 35(11): 1367–1393.
13. Goyal, A., Carter, W.C., de Souza e Silva, E., Lavenberg, S.S., & Trivedi, K.S. (1986). The System Availability Estimator. In H. Kopetz & M. Dal Cin (eds.), *Proceedings of the sixteenth symposium on fault tolerant computing*, pp. 84–89. New York: IEEE.
14. Goyal, A. & Lavenberg, S.S. (1987). Modelling and analysis of computer system availability. *IBM Journal of Research and Development* 31(6): 651–664.
15. Goyal, A., Shahabuddin, P., Heidelberger, P., Nicola, V.F., & Glynn, P.W. (1992). A unified framework for simulating Markovian models of highly dependable systems. *IEEE Transactions on Computers* C-41(1): 36–51.
16. Gross, D. & Miller, D.R. (1984). The randomization technique as a modeling tool and solution procedure for transient Markov processes. *Operations Research* 32: 343–361.
17. Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation* 5(1): 43–85.
18. Heidelberger, P., Shahabuddin, P., & Nicola, V.F. (1994). Bounded relative error in estimating transient measures of highly dependable Markovian systems. *ACM Transactions on Modeling and Computer Simulation* 4(2): 137–164.
19. Jensen, A. (1953). Markoff chains as an aid in the study of Markoff processes. *Skandinavisk Aktuarietidskrift* 36: 87–91.
20. Juneja, S. & Shahabuddin, P. (1992). Fast simulation of Markovian reliability/availability models with general repair policies. In D. Pradhan, J. Stiffler, J. Lala, & I. Koren (eds.), *Proceedings of the twenty-second annual international symposium on fault tolerant computing*. New York: IEEE Computer Society Press, pp. 150–159.
21. Juneja, S. & Shahabuddin, P. (2001). Efficient simulation of Markov chains with small transition probabilities. *Management Science* 47: 547–562.
22. Juneja, S. & Shahabuddin, P. (2001). A splitting-based importance-sampling algorithm for fast simulation of Markov reliability models with general repair policies. *IEEE Transactions on Reliability* 50: 235–245.

23. Kalashnikov, V.V. (1989). Analytical and simulation estimates of reliability for regenerative models. *Systems Analysis Modelling Simulation* 6: 833–851.
24. Keilson, J. (1979). *Markov chain models—rarity and exponentiality*. New York: Springer-Verlag.
25. Lewis, E.E. & Böhm, F. (1984). Monte Carlo simulation of Markov unreliability models. *Nuclear Engineering and Design* 77: 49–62.
26. Muntz, R.R., de Souza e Silva, E., & Goyal, A. (1989). Bounding availability of repairable computer systems. *IEEE Transactions on Computers* 38: 1714–1723.
27. Nakayama, M.K. (1994). A characterization of the simple failure biasing method for simulations of highly reliable Markovian systems. *ACM Transactions on Modeling and Computer Simulation* 4(1): 52–88.
28. Nakayama, M.K. (1994). Fast simulation methods for highly dependable systems. In D.A. Sadowski, A.F. Seila, J.D. Tew, & S. Manivannan (eds.), *Proceedings of the 1994 winter simulation conference*. New York: IEEE Press, pp. 221–228.
29. Nakayama, M.K. (1996). General conditions for bounded relative error in simulation of highly reliable Markovian systems. *Advances in Applied Probability* 28: 687–727.
30. Nicola, V.F., Nakayama, M.K., Heidelberger, P., & Goyal, A. (1993). Fast simulation of highly dependable systems with general failure and repair processes. *IEEE Transactions on Computers* 42(8): 1440–1452.
31. Nicola, V.F., Shahabuddin, P., Heidelberger, P., & Glynn, P.W. (1993). Fast simulation of steady-state availability in non-Markovian highly dependable systems. In J.C. Laprie & A. Goyal (eds.), *Proceedings of the 23rd annual international symposium on fault tolerant computing*. New York: IEEE Computer Society Press, pp. 491–498.
32. Nicola, V.F., Shahabuddin, P., & Nakayama, M.K. (2001). Techniques for fast simulation of models of highly dependable systems. *IEEE Transactions on Reliability* 50: 246–264.
33. Obal, W.D. II & Sanders, W.H. (1994). Importance sampling simulation in ULTRA-SAN. *Simulation* 62: 98–111.
34. Serfling, R.J. (1980). *Approximation theorems in mathematical statistics.* New York: Wiley.
35. Shahabuddin, P. (1994). Fast transient simulation of Markovian models of highly dependable systems. *Performance Evaluation, Special Issue: Performance '93—16th IFIP Working Group 7.3 International Symposium on Computer Performance Modeling, Measurement and Evaluation* 20, 1–3, 267–286.
36. Shahabuddin, P. (1994). Importance sampling for the simulation of highly reliable Markovian systems. *Management Science* 40(3): 333–352.
37. Shahabuddin, P. (2001). Rare event simulation in stochastic models. In B.R. Haverkort, R. Marie, K. Trivedi, & G. Rubino (eds.), *Performability modeling techniques and tools*. New York: Wiley, pp. 163–178.
38. Shahabuddin, P. & Nakayama, M.K. (1993). Estimation of unreliability and its derivatives for large time horizons in Markovian systems. In E.C. Russell, W.E. Biles, G.W. Evans, & M. Mollaghasemi (eds.), *Proceedings of the 1993 winter simulation conference*. New York: IEEE Press, pp. 422–429.
39. Shultes, B.C. (1997). Regenerative techniques for estimating performance measures of highly dependable systems with repair. Ph.D. thesis, Georgia Institute of Technology, Atlanta.
40. Solovyev, A.D. (1971). Asymptotic behavior of the time of first occurrence of a rare event. *Engineering Cybernetics* 9: 1038–1048.
41. Strickland, S.G. (1995). Optimal importance sampling for quick simulation of highly reliable systems. In E.C. Russell, W.E. Biles, G.W. Evans, & M. Mollaghasemi (eds.), *Proceedings of the 1993 winter simulation conference*. New York: IEEE Press, pp. 437–444.
42. Young, D. (1971). *Iterative solution of large linear systems.* New York: Academic Press.

APPENDIX

We first briefly review the notation and framework considered in Shahabuddin [36]. Let $\mathbf{P}_R$ be the matrix constructed from $\mathbf{P}$ where the rows and columns corresponding to state **1** and states

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

in F are removed. Since all states represented in $\mathbf{P}_R$ have at least one ongoing repair transition, $\mathbf{P}_R$ is a matrix in which the positive elements above the diagonal (i.e., probabilities of failure transitions) are of the form $c\epsilon^d + o(\epsilon^d)$, $c > 0$, $d > 0$, where c and d are generic representations of constants. The positive elements below the diagonal (i.e., probabilities of repair transitions) are of the form $c + o(1)$, $c > 0$. For $\mathbf{x} \in U$, let $p_F(\mathbf{x}) = \sum_{\mathbf{y} \in F} P(\mathbf{x}, \mathbf{y})$, $p_{\mathbf{1}}(\mathbf{x}) = P(\mathbf{x}, \mathbf{1})$, and $p_0(\mathbf{x}) = P(\mathbf{1}, \mathbf{x})$. Let the vectors $\mathbf{p}_F$, $\mathbf{p}_1$, and $\mathbf{p}_0$ be given by $\mathbf{p}_F = \{p_F(\mathbf{x}) : \mathbf{x} \in U - \mathbf{1}\}$, $\mathbf{p}_1 = \{p_{\mathbf{1}}(\mathbf{x}) : \mathbf{x} \in U - \mathbf{1}\}$, and $\mathbf{p}_0 = \{p_0(\mathbf{x}) : \mathbf{x} \in U - \mathbf{1}\}$, respectively. All vectors are assumed to be column vectors. The positive elements of $\mathbf{p}_1$ are of the form $c + o(1)$, $c > 0$, the positive elements of $\mathbf{p}_F$ are $o(1)$, and the positive elements of $\mathbf{p}_0$ are $O(1)$. From Assumption (A.4), we have that at least one element of $\mathbf{p}_0$ is positive. Let $v_\gamma(\mathbf{x}) = P(T_F < T_{\mathbf{1}} | \mathbf{Y}_0 = \mathbf{x})$ and $\mathbf{v}_\gamma = \{v_\gamma(\mathbf{x}) : \mathbf{x} \in U - \mathbf{1}\}$. By Assumption (A.3), all elements of $\mathbf{v}_\gamma$ are positive. By Assumption (A.1), starting from any state $\mathbf{x} \in U - \mathbf{1}$, the Markov chain hits $\mathbf{1} \cup F$ with probability 1. Consequently, for $\mathbf{x} \in U$, $P(T_{\mathbf{1}} < T_F | \mathbf{Y}_0 = \mathbf{x}) = 1 - v_\gamma(\mathbf{x})$. By Assumption (A.2), we have that all elements of $\mathbf{e} - \mathbf{v}_\gamma$ are positive, where $\mathbf{e}$ is a vector of 1's.

By a matrix or a vector being $o(\epsilon^d)$ (resp., $O(\epsilon^d)$, $\underline{O}(\epsilon^d)$, $\Omega(\epsilon^d)$), $d \geq 0$, we mean that all elements of that matrix or vector are $o(\epsilon^d)$ (resp., $O(\epsilon^d)$, $\underline{O}(\epsilon^d)$, $\Omega(\epsilon^d)$). It has been shown in Shahabuddin [36] (see the proof of Lemma 3 in that article) that there exists a nonnegative integer constant N_0 such that

$$\sum_{i=N_0+1}^{\infty} \mathbf{P}_R^i \quad \text{is} \quad o(1). \tag{A.1}$$

Therefore,

$$\sum_{i=0}^{\infty} \mathbf{P}_R^i \quad \text{is} \quad O(1). \tag{A.2}$$

It was shown in Shahabuddin [36] that $\mathbf{v}_\gamma = \sum_{i=0}^{\infty} \mathbf{P}_R^i \mathbf{p}_F$. Similarly, one can show that $\mathbf{e} - \mathbf{v}_\gamma = \sum_{i=0}^{\infty} \mathbf{P}_R^i \mathbf{p}_{\mathbf{1}}$. From Eq. (A.2) and the fact mentioned earlier that $\mathbf{p}_F$ is $o(1)$, we get that $\mathbf{v}_\gamma$ is $o(1)$.

Let $\tilde{\mathbf{P}}$ be a matrix constructed out of $\mathbf{P}$ by removing the row and column corresponding to state $\mathbf{1}$. $\tilde{\mathbf{P}}$ has the same structure as $\mathbf{P}_R$ in the sense that the positive elements above the diagonal are of the form $c\epsilon^d + o(\epsilon^d)$, $c > 0$, $d > 0$, and the positive elements below the diagonal are of the form $c + o(1)$, $c > 0$. Thus,

$$\sum_{i=0}^{\infty} \tilde{\mathbf{P}}^i \quad \text{is} \quad O(1) \tag{A.3}$$

for exactly the same reason as Eq. (A.2).

LEMMA A.1: $E(T_{\mathbf{1}}) = \Omega(\epsilon^{-r_0})$.

PROOF: Let $\tilde{\mathbf{h}} = \{h(\mathbf{x}) : \mathbf{x} \in S - \mathbf{1}\}$ (recall that $h(\mathbf{x}) \equiv 1/q(\mathbf{x})$), $\tilde{w}(\mathbf{x}) = E(T_{\mathbf{1}} | \mathbf{X}_0 = \mathbf{x})$, and $\tilde{\mathbf{w}} = \{\tilde{w}(\mathbf{x}) : \mathbf{x} \in S - \mathbf{1}\}$. Then, we have that $\tilde{\mathbf{w}} = \tilde{\mathbf{h}} + \tilde{\mathbf{P}}\tilde{\mathbf{h}}$, from which we get $\tilde{\mathbf{w}} = \sum_{i=0}^{\infty} \tilde{\mathbf{P}}^i \tilde{\mathbf{h}}$. From Assumption (A.2) and the fact that the repair rates are $\Omega(1)$, we get that $\tilde{\mathbf{h}}$ is $\Omega(1)$, and so from Eq. (A.3), $\tilde{\mathbf{w}}$ is $O(1)$. In addition, $\mathbf{p}_0$ is $O(1)$, so $E(T_{\mathbf{1}}) \equiv \tilde{w}(\mathbf{1}) = h(\mathbf{1}) + \mathbf{p}_0^T \tilde{\mathbf{w}} = \Omega(\epsilon^{-r_0}) + O(1) = \Omega(\epsilon^{-r_0})$. ■

LEMMA A.2: *$E(\overline{W})$ and $E(\underline{W})$ are $O(1)$.*

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

PROOF: For $\mathbf{x} \in U$, let $\overline{w}(\mathbf{x}) = E(\min(T_{\mathbf{1}}, T_F) I(T_F < T_{\mathbf{1}}) | \mathbf{Y}_0 = \mathbf{x})$, and $\underline{w}(\mathbf{x}) = E(\min(T_{\mathbf{1}}, T_F) I(T_{\mathbf{1}} < T_F) | \mathbf{Y}_0 = \mathbf{x})$. Let $\overline{\mathbf{w}} = \{\overline{w}(\mathbf{x}) : \mathbf{x} \in U - \mathbf{1}\}$ and $\underline{\mathbf{w}} = \{\underline{w}(\mathbf{x}) : \mathbf{x} \in U - \mathbf{1}\}$. Let $\mathbf{h} = \{h(\mathbf{x}) : \mathbf{x} \in U - \mathbf{1}\}$. Note that $\underline{\mathbf{w}} = \mathbf{h} \circ (\mathbf{e} - \mathbf{v}_\gamma) + \mathbf{P}_R \underline{\mathbf{w}}$ (the "$\circ$" denotes the scalar product) from which we get $\underline{\mathbf{w}} = \sum_{i=0}^{\infty} \mathbf{P}_R^i (\mathbf{h} \circ (\mathbf{e} - \mathbf{v}_\gamma))$. Then, using the fact from Section 2.1 that $1 - \gamma > 0$, we get that

$$\begin{aligned} E(\underline{W}) &= E(\min(T_{\mathbf{1}}, T_F) - Z | T_{\mathbf{1}} < T_F) \\ &= \frac{E((\min(T_{\mathbf{1}}, T_F) - Z) I(T_{\mathbf{1}} < T_F))}{(1 - \gamma)} \\ &= \frac{\mathbf{p}_0^T \underline{\mathbf{w}}}{\mathbf{p}_0^T (\mathbf{e} - \mathbf{v}_\gamma)}. \end{aligned} \tag{A.4}$$

In a similar fashion, we get that $\overline{\mathbf{w}} = \sum_{i=0}^{\infty} \mathbf{P}_R^i (\mathbf{h} \circ \mathbf{v}_\gamma)$ and using the fact from Section 2.1 that $\gamma > 0$, we get that

$$\begin{aligned} E(\overline{W}) &= E(\min(T_{\mathbf{1}}, T_F) - Z | T_F < T_{\mathbf{1}}) \\ &= \frac{\mathbf{p}_0^T \overline{\mathbf{w}}}{\gamma} \\ &= \frac{\mathbf{p}_0^T \overline{\mathbf{w}}}{p_F(\mathbf{1}) + \mathbf{p}_0^T \mathbf{v}_\gamma} \\ &\leq \frac{\mathbf{p}_0^T \overline{\mathbf{w}}}{\mathbf{p}_0^T \mathbf{v}_\gamma}. \end{aligned} \tag{A.5}$$

By Assumption (A.4) and the fact that all elements of $\mathbf{v}_\gamma$ are positive, $\mathbf{p}_0^T \mathbf{v}_\gamma > 0$. We will now show that $\underline{\mathbf{w}}$ (resp., $\overline{\mathbf{w}}$) and $\mathbf{e} - \mathbf{v}_\gamma$ (resp., $\mathbf{v}_\gamma$) are of the same order. From Assumption (A.2) and the fact that the repair rates are $\Omega(1)$, we get that $\mathbf{h}$ is $\Omega(1)$. Consequently, if an element of $\mathbf{e} - \mathbf{v}_\gamma$ (resp., $\mathbf{v}_\gamma$) is $\Omega(\epsilon^d)$ for some $d \geq 0$, then the corresponding element of $\mathbf{h} \circ (\mathbf{e} - \mathbf{v}_\gamma)$ (resp., $\mathbf{h} \circ \mathbf{v}_\gamma$) is also $\Omega(\epsilon^d)$ for the same d. Because $\sum_{i=0}^{\infty} \mathbf{P}_R^i = O(1)$, the corresponding element of $\underline{\mathbf{w}}$ (resp., $\overline{\mathbf{w}}$) is similarly $O(\epsilon^d)$ for the same d. Using this fact in Eq. (A.4) (resp., Eq. (A.5)), we see that $E(\underline{W})$ (resp., $E(\overline{W})$) is $O(1)$. ■

LEMMA A.3: *$E(\underline{W}^2)$ is $O(1)$.*

PROOF: We use notation from the proof of Lemma A.2. Furthermore, let $H(\mathbf{x})$ denote the holding-time random variable in state $\mathbf{x}$. Clearly, $H(\mathbf{x})$ is exponentially distributed with rate $q(\mathbf{x})$. Define $l(\mathbf{x}) = E(H^2(\mathbf{x})) = 2/q^2(\mathbf{x})$, $\mathbf{l} = \{l(\mathbf{x}) : \mathbf{x} \in U - \mathbf{1}\}$, $\underline{q}(\mathbf{x}) = E((\min(T_{\mathbf{1}}, T_F))^2 \times I(T_{\mathbf{1}} < T_F) | \mathbf{Y}_0 = \mathbf{x})$, and $\underline{\mathbf{q}} = \{\underline{q}(\mathbf{x}) : \mathbf{x} \in U - \mathbf{1}\}$. Then, one can easily derive the recursive equation that $\underline{\mathbf{q}} = \mathbf{l} \circ (\mathbf{e} - \mathbf{v}_\gamma) + 2\mathbf{h} \circ (\mathbf{P}_R \underline{\mathbf{w}}) + \mathbf{P}_R \underline{\mathbf{q}}$ from which one gets that $\underline{\mathbf{q}} = \sum_{i=0}^{\infty} \mathbf{P}_R (\mathbf{l} \circ (\mathbf{e} - \mathbf{v}_\gamma) + 2\mathbf{h} \circ (\mathbf{P}_R \underline{\mathbf{w}}))$. In that case,

$$\begin{aligned} E(\underline{W}^2) &= \frac{E((\min(T_F, T_{\mathbf{1}}) - Z)^2 I(T_{\mathbf{1}} < T_F))}{(1 - \gamma)} \\ &= \frac{\mathbf{p}_0^T \underline{\mathbf{q}}}{\mathbf{p}_0^T (\mathbf{e} - \mathbf{v}_\gamma)}. \end{aligned} \tag{A.6}$$

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

Using the fact that $\mathbf{v}_\gamma$ is $o(1)$, we get that the elements of $\mathbf{e} - \mathbf{v}_\gamma$ are $\Omega(1)$. In addition, from what was stated in the proof of Lemma A.2, $\underline{\mathbf{w}}$ is $O(1)$. The $\mathbf{l}$ and $\mathbf{h}$ are $\Omega(1)$, $\sum_{i=0}^{\infty} \mathbf{P}_R^i$ is $O(1)$, and, therefore, $\underline{\mathbf{q}}$ is $O(1)$. Then, the proof follows from Eq. (A.6). ■

PROOF OF PROPOSITION 1: For any matrix (vector), say $\mathbf{P}_R$, with elements of the form $c + o(1)$, $c \geq 0$, let $\hat{\mathbf{P}}_R$ be the matrix containing the "c part" of the elements. Using the fact mentioned earlier that $\mathbf{v}_\gamma$ is $o(1)$, we get that

$$\sum_{i=0}^{\infty} \mathbf{P}_R^i \mathbf{p_1} = \mathbf{e} - \mathbf{v}_\gamma = \mathbf{e} - o(1). \tag{A.7}$$

Using Eq. (A.1) and the fact that the positive elements of $\mathbf{p}_1$ are of the form $c + o(1)$, $c > 0$, we have that

$$\sum_{i=0}^{\infty} \mathbf{P}_R^i \mathbf{p_1} = \left(\sum_{i=0}^{N_0} \mathbf{P}_R^i + o(1) \right) \mathbf{p_1} = \sum_{i=0}^{N_0} (\hat{\mathbf{P}}_R + o(1))^i (\hat{\mathbf{p}}_\mathbf{1} + o(1)) + o(1)$$

$$= \sum_{i=0}^{N_0} \hat{\mathbf{P}}_R^i \hat{\mathbf{p}}_\mathbf{1} + o(1). \tag{A.8}$$

Comparing Eq. (A.8) with Eq. (A.7), we see that

$$\sum_{i=0}^{N_0} \hat{\mathbf{P}}_R^i \hat{\mathbf{p}}_\mathbf{1} = \mathbf{e}. \tag{A.9}$$

As in Shahabuddin [36], we construct the matrix $\mathbf{B} = \{B(\mathbf{x}, \mathbf{y}) : \mathbf{x}, \mathbf{y} \in S\}$ by setting $B(\mathbf{x}, \mathbf{y}) = P^2(\mathbf{x}, \mathbf{y})/P'(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}$, $\mathbf{y}$ such that $P'(\mathbf{x}, \mathbf{y}) \neq 0$; $B(\mathbf{x}, \mathbf{y}) = 0$ otherwise. The $\mathbf{B}_R$, $\mathbf{b}_F$, $\mathbf{b_1}$, and $\mathbf{b}_0$ are constructed from $\mathbf{B}$ the same way as $\mathbf{P}_R$, $\mathbf{p}_F$, $\mathbf{p_1}$, and $\mathbf{p}_0$ were constructed from $\mathbf{P}$. Note that $\mathbf{B}_R$ is a matrix in which the positive elements above the diagonal are of the form $c\epsilon^d + o(\epsilon^d)$, $c > 0$, $d \geq 2$, and the positive elements below the diagonal are of the form $c + o(1)$, $c > 0$. It has been shown in Shahabuddin [36] that for the same N_0 as earlier, $\sum_{i=N_0+1}^{\infty} \mathbf{B}_R = o(1)$. Thus, $\sum_{i=0}^{\infty} \mathbf{B}_R^i$ is O(1). It has also been shown in Shahabuddin [36] that $D = \mathbf{b}_0^T (\sum_{i=0}^{\infty} \mathbf{B}_R^i \mathbf{b}_F) + \sum_{\mathbf{y} \in F} B(\mathbf{1}, \mathbf{y})$. Using a similar method, we can show that $B = \mathbf{b}_0^T (\sum_{i=0}^{\infty} \mathbf{B}_R^i \mathbf{b_1})$. The positive elements of $\mathbf{b_1}$ are of the form $c + o(1)$, where $c > 0$. Hence,

$$\sum_{i=0}^{\infty} \mathbf{B}_R^i \mathbf{b_1} = \left(\sum_{i=0}^{N_0} \mathbf{B}_R^i + o(1) \right) \mathbf{b_1} = \sum_{i=0}^{N_0} (\hat{\mathbf{B}}_R + o(1))^i (\hat{\mathbf{b}}_\mathbf{1} + o(1)) + o(1) = \sum_{i=0}^{N_0} \hat{\mathbf{B}}_R^i \hat{\mathbf{b}}_\mathbf{1} + o(1)$$

$$= \left(\sum_{i=0}^{N_0} \left(\frac{1}{1-\theta} \right)^i \hat{\mathbf{P}}_R^i \right) \left(\frac{1}{1-\theta}\, \hat{\mathbf{p}}_\mathbf{1} \right) + o(1) \tag{A.10}$$

$$\geq \frac{1}{1-\theta} \left(\sum_{i=0}^{N_0} \hat{\mathbf{P}}_R^i \hat{\mathbf{p}}_\mathbf{1} \right) + o(1) = \frac{1}{1-\theta}\, \mathbf{e} + o(1). \tag{A.11}$$

Equation (A.10) follows from the form of the elements of $\mathbf{P}'$ and the fact that in the Markov chain corresponding to $\mathbf{P}$, the sum of all the repair transition probabilities from any state other than $\mathbf{1}$ is $1 - o(1)$. The last equality follows from Eq. (A.9). From Eq. (A.11), we get that

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

$$B = \mathbf{b}_0^T \left(\sum_{i=0}^{\infty} \mathbf{B}_R^i \mathbf{b}_\mathbf{1} \right) \geq \frac{1}{1-\theta} \mathbf{b}_0^T \mathbf{e} + o(1) = \frac{1}{1-\theta} \sum_{\mathbf{x} \in U - \mathbf{1}} \frac{P^2(\mathbf{1}, \mathbf{x})}{(1/r(\mathbf{1}))} + o(1),$$

where $r(\mathbf{1})$ is the number of positive probability failure transitions from $\mathbf{1}$. The last equality follows from the fact that $P'(\mathbf{1}, \mathbf{x}) = 1/r(\mathbf{1})$.

The elements of $P(\mathbf{1}, \mathbf{x})$ can be represented in the form $c + o(1)$, $c \geq 0$. Let $\tilde{U} \subseteq U - \mathbf{1}$ be the set of $\mathbf{x}$ for which $P(\mathbf{1}, \mathbf{x})$ is of the form $c + o(1)$, $c > 0$. From Assumption (A.4′), we see that $\sum_{\mathbf{x} \in \tilde{U}} \hat{P}(\mathbf{1}, \mathbf{x}) = 1$. Since $r(\mathbf{1}) \geq |\tilde{U}|$, we have that

$$\begin{aligned} B &\geq \frac{1}{1-\theta} |\tilde{U}| \sum_{\mathbf{x} \in \tilde{U}} (\hat{P}(\mathbf{1}, \mathbf{x}) + o(1))^2 + o(1) \\ &= \frac{1}{1-\theta} |\tilde{U}| \sum_{\mathbf{x} \in \tilde{U}} \hat{P}^2(\mathbf{1}, \mathbf{x}) + o(1) \\ &\geq \frac{1}{1-\theta} + o(1). \end{aligned}$$

The last inequality follows because the minimum of the function $f(z_1, z_2, \ldots z_k) = k \sum_{i=1}^k z_k^2$ subject to the constraints $\sum_{i=1}^k z_i = 1$, $z_i \geq 0$, $z_i \in \mathbf{R}$, $\forall i$, is 1. Therefore,

$$B - 1 \geq \frac{\theta}{1-\theta} + o(1),$$

so $B - 1 = \underline{O}(1)$. Using the fact that $\sum_{i=0}^{\infty} \mathbf{B}_R^i$, $\mathbf{b}_0$, and $\mathbf{b}_\mathbf{1}$ are $O(1)$, we get that $B - 1 = \mathbf{b}_0^T (\sum_{i=0}^{\infty} \mathbf{B}_R^i \mathbf{b}_1) - 1$ is $O(1)$. Hence, $B - 1 = \Omega(1)$. ■

Proof of Theorem 5:

(a) The $U(t) \leq \bar{U}(t)$ bound can be proved by exactly the same method as Lemma 1. We will now prove the lower bound.

Let $f_{(q,k)}(x)$ be the density of an Erlang random variable with rate q and shape parameter k (i.e., the density of $Z_1 + Z_2 + \cdots + Z_K$ given that $K = k$). Then,

$$\begin{aligned} U(t) &= P(Z_1 + W_1 + Z_2 + W_2 + \cdots + Z_K + W_K \leq t) \\ &= \sum_{k=1}^{\infty} \int_{x=0}^{t} P\left(W_1 + \cdots + W_k \leq t - x | K = k, \sum_{i=1}^{k} Z_i = x \right) \\ &\quad \times f_{(q,k)}(x)\, dx\, \gamma(1-\gamma)^{k-1} \\ &= \bar{U}(t) - \sum_{k=1}^{\infty} \int_{x=0}^{t} P\left(W_1 + \cdots + W_k > t - x | K = k, \sum_{i=1}^{k} Z_i = x \right) \\ &\quad \times f_{(q,k)}(x)\, dx\, \gamma(1-\gamma)^{k-1}. \end{aligned}$$

We now have to prove that the second term above is upper bounded by $U_{\text{err}}(t)$. Given $K = k$, the W_i's are (conditionally) independent of the Z_i's. Also given $K = k$, the $W_1, W_2, \ldots, W_{k-1}$ are i.i.d. and have the distribution of $\underline{W}$. Moreover, given $K = k$, W_k is independent of the $W_1, W_2, \ldots W_{k-1}$ and has the distribution of $\overline{W}$. Thus, we can rewrite the second term and bound it as follows:

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

$$\sum_{k=1}^{\infty} \int_{x=0}^{t-l} P(\underline{W}_1 + \cdots + \underline{W}_{k-1} + \overline{W}_k > t - x) f_{q,k}(x)\, dx\, \gamma(1-\gamma)^{k-1}$$

$$+ \sum_{k=1}^{\infty} \int_{x=t-l}^{t} P(\underline{W}_1 + \cdots + \underline{W}_{k-1} + \overline{W}_k > t - x) f_{q,k}(x)\, dx\, \gamma(1-\gamma)^{k-1}$$

$$\leq \sum_{k=1}^{\infty} \int_{x=0}^{t-l} \frac{(k-1)E(\underline{W}) + E(\overline{W})}{(t-x)} f_{q,k}(x)\, dx\, \gamma(1-\gamma)^{k-1}$$

(using Chebyshev's inequality)

$$+ \sum_{k=1}^{\infty} \int_{x=t-l}^{t} f_{q,k}(x)\, dx\, \gamma(1-\gamma)^{k-1}$$

$$\leq \frac{1}{l} \sum_{k=1}^{\infty} \int_{x=0}^{t-l} ((k-1)E(\underline{W}) + E(\overline{W})) f_{q,k}(x)\, dx\, \gamma(1-\gamma)^{k-1}$$

$$+ (e^{-\gamma q(t-l)} - e^{-\gamma qt})$$

$$= \frac{E(\overline{W})}{l}(1 - e^{-\gamma q(t-l)}) + \frac{E(\underline{W})(1-\gamma)}{\gamma l}$$

$$\times (1 - e^{-\gamma q(t-l)} - \gamma q(t-l)e^{-\gamma q(t-l)}) + (e^{-\gamma q(t-l)} - e^{-\gamma qt}) \qquad \textbf{(A.12)}$$

$$= U_{\text{err}}(t).$$

(b) We will first show that $U_{\text{err}}(t)/\bar{U}(t) \to 0$ as $\epsilon \to 0$. The result then follows from the fact that

$$\text{REB}(t) = \frac{U_{\text{err}}(t)}{2U(t)} \leq \frac{U_{\text{err}}(t)}{2\bar{U}(t)} \left(\frac{1}{1 - U_{\text{err}}(t)/\bar{U}(t)} \right).$$

We use the representation of $U_{\text{err}}(t)$ given by Eq. (A.12).

First, let us study the properties of l. If $r_t < r_0$ (resp., $r_t > r_0$), then for all sufficiently small ϵ, $t < 1/q$ (resp., $t > 1/q$), which implies that $\sqrt{t} > t\sqrt{q}$ (resp., $\sqrt{t} < t\sqrt{q}$). Thus, l is $\Omega(\epsilon^{-r_t/2})$ (resp., l is $\Omega(\epsilon^{-r_t+r_0/2})$). If $r_t = r_0$, then $l = \sqrt{t}$ or $l = t\sqrt{q}$, but in either case, $l = \Omega(\epsilon^{-r_t/2})$. Since $r_t > 0$ and for $r_t > r_0$, $r_t - r_0/2 > 0$, we have that $1/l$ is $o(1)$. Since $r_0 > 0$ and $r_t > 0$, we have that $l/t = o(1)$. It then follows that $t - l$ is $\Omega(\epsilon^{-r_t})$.

Consider the case where $r_t < r + r_0$, so that $\epsilon^{r+r_0-r_t} \to 0$ as $\epsilon \to 0$. Using the well-known fact that $1 - e^{-x} = x + o(x)$, we get that $\bar{U}(t)$ is $\Omega(\epsilon^{r+r_0-r_t})$. Hence, all we have to show is that $U_{\text{err}}(t) = o(\epsilon^{r+r_0-r_t})$. Since $t - l$ is $\Omega(\epsilon^{-r_t})$, $(1 - e^{-\gamma q(t-l)}) = \gamma q(t-l) + o(\epsilon^{r+r_0-r_t})$ is $\Omega(\epsilon^{r+r_0-r_t})$. Lemma A.2 shows that both $E(\underline{W})$ and $E(\overline{W})$ are $O(1)$. Since $1/l$ is $o(1)$, the first term in Eq. (A.12) is $o(\epsilon^{r+r_0-r_t})$. Using the well-known fact that $1 - e^{-x} - xe^{-x}$ is $x^2/2 + o(x^2)$ and the fact that $t - l$ is $\Omega(\epsilon^{-r_t})$, one can easily show that $(1 - e^{-\gamma q(t-l)} - \gamma q(t-l)e^{-\gamma q(t-l)})$ is $\gamma q(t-l) \times \Omega(\epsilon^{r+r_0-r_t})$. Considering separately the two cases of $0 < r_t \leq r_0$ and $r_0 < r_t < r + r_0$, one can show that $q(t-l)/l$ is $o(1)$ and, therefore, the second term in Eq. (A.12) is $o(\epsilon^{r+r_0-r_t})$. Finally,

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press

$$e^{-\gamma q(t-l)} - e^{-\gamma qt} = (1 - e^{-\gamma qt}) - (1 - e^{\gamma q(t-l)})$$
$$= \gamma qt \frac{l}{t} + o(\epsilon^{r+r_0-r_t}).$$

Using the fact mentioned earlier that l/t is $o(1)$, we can show that the last part of Eq. (A.12) is $o(\epsilon^{r+r_0-r_t})$.

Now, consider the case where $r_t \geq r + r_0$. In this case, $\bar{U}(t)$ is $\Omega(1)$. Thus, all we have to show is that $U_{\text{err},t}(t)$ is $o(1)$. For $r_t = r + r_0$, $1 - e^{-\gamma q(t-l)}$ and $1 - e^{-\gamma q(t-l)} - \gamma q(t-l)e^{-\gamma q(t-l)}$ are $\Omega(1)$. For $r_t > r + r_0$, one can show the same by using the fact that $1 - e^{-1/x}$ and $1 - e^{-1/x} - (1/x)e^{-1/x}$ approach 1 as $x \to 0$. Using the fact that $1/l$ is $o(1)$, one can show that the first term of Eq. (A.12) is $o(1)$. Note that $r_t \geq r + r_0$ implies that $r_t > r_0$. Consequently, $l = t\sqrt{q}$ and $1/\gamma l$ is $\Omega(\epsilon^{r_t - r - r_0/2}) = o(1)$. Using this fact, one can show that the second term of Eq. (A.12) is $o(1)$. Express the last term of Eq. (A.12) as $e^{-\gamma qt}(e^{\gamma qt\sqrt{q}} - 1)$. For $r_t > r + r_0$, it is easy to see that as $\epsilon \to 0$, this term tends to 0 and, hence, it is $o(1)$. For $r_t = r + r_0$, $e^{-\gamma qt}$ is $\Omega(1)$ and $(e^{\gamma qt\sqrt{q}} - 1)$ is $o(1)$, and so the last term is also $o(1)$. ■

https://doi.org/10.1017/S0269964804183058 Published online by Cambridge University Press