MTTF ESTIMATION USING IMPORTANCE SAMPLING ON MARKOV MODELS

Héctor CANCELA[*], Gerardo RUBINO[†] and Bruno TUFFIN[‡]

Abstract

Very complex systems occur nowadays quite frequently in many technological areas and they are often required to comply with high dependability standards. To study their availability and reliability characteristics, Markovian models are commonly used. Due to the size and complexity of the systems, and due to the rarity of system failures, both analytical solutions and "crude" simulation can be inefficient or even non-relevant. A number of variance reduction Monte Carlo techniques have been proposed to overcome this difficulty; importance sampling methods are among the most efficient. The objective of this paper is to survey existing importance sampling schemes, to propose some new schemes and improvements on existing ones, and to discuss on their different properties.

[*] UDELAR, Montevideo, Uruguay, *cancela@fing.edu.uy*
[†] ENSTB, Cesson-Sevigné, France & IRISA, Rennes, France, *rubino@irisa.fr*
[‡] IRISA, Rennes, France, *btuffin@irisa.fr*

1 Introduction

Let us consider a multi-component repairable system. The user has defined what an operational state is, and the fact that the system is repairable means that it can come back to such a state after the occurrence of a failure, due to some repairing facility included in. We are interested in evaluating some specific dependability metrics from a model of the system. If X_t denotes the state of the model at time t, where $X_t \in S$, the specifications induce a partition of the state space S into two (disjoint) sets: U, the set of states where the system is up (delivering service as specified), and D, composed by those states where the system is down (the delivered service does not fit anymore the specifications).

The most important dependability metrics are: (i) the *asymptotic availability*, defined by $\Pr(X_\infty \in U)$ (assuming for instance that the model is irreducible and ergodic); (ii) the *MTTF* (Mean Time To Failure), defined as $\mathrm{E}(\tau_D)$ where τ_D is the *hitting time* of the set D (assuming that $X_0 \in U$), that is, $\tau_D = \inf\{t \mid X_t \in D\}$; (iii) the *reliability* at time t, equal to $\Pr(\tau_D > t)$, also assuming that $X_0 \in U$; (iv) the *point availability* at time t, defined by $\Pr(X_t \in U)$; (v) the distribution (or the moments) of the random variable *interval availability* in $[0, t]$, defined by $\frac{1}{t}\int_0^t [X_s \in U] ds$ where $[\mathcal{P}]$ is the indicator function of the predicate $\mathcal{P}$.

A frequent situation is that the model (the stochastic process X) is quite complex and large (that is, $|S| \gg 1$) and that the failed states are *rare*, that is, $\tau_D \gg \tau_0$ with high probability, where τ_0 is the return time to the initial state 0, presumed to be the (only) state where all the components are up (that is, $\tau_0 = \inf\{t > 0 \mid X_t = 0, X_{t^-} \neq 0\}$). The size of the model may make difficult or impossible its exact numerical evaluation and the rarity of the interesting events can do the same with a naive Monte Carlo estimation (see for example Heidelberger [6]).

In the first case, an alternative approach deals with the computation of bounds of the measures of interest. In this line, see Muntz et al. [11] for an efficient scheme devoted to the analysis of the asymptotic availability, extended by Mahévas and Rubino [10] to deal with more general models (and also to the analysis of asymptotic performance measures). In the Monte Carlo area, different importance sampling schemes have been proved to be appropriate, in order to design efficient estimation algorithms. This paper focuses on a basic and widely used dependability measure, the *MTTF*. We analyze some known importance sampling schemes designed to estimate it, we exhibit some improving techniques and we discuss on general properties of this family of methods.

This paper is organized as follows. We give the model specifications in Section 2 and we describe general simulation techniques in Section 3. As we study highly dependable systems, we introduce a rarity parameter in Section 4 and we present the importance sampling schemes in Section 5. Some of these schemes are taken from the literature, but many are new, adapted to specific situations in practice. Section 6 deals with important properties of the estimators: bounded relative error and bounded normal approximation. Comparisons of the algorithms are then given: in Section 7 asymptotically as the rarity parameter goes to 0, and numerically in Section 8. Moreover we show in Section 9 that numerical results can lead to wrong estimations in some cases. This is due to the fact that the events can still be rare, even if occuring more often. This important remark has not been stated in the literature yet. Finally, we conclude in Section 10.

2 The model

The system is represented (modeled) by a finite continuous time homogeneous and irreducible Markov chain $X = \{X_t, t \geq 0\}$. We denote by S the state space of X, and we suppose that $1 \ll |S|\ (< \infty)$.

Let us denote by $Q(x, y)$ the transition rate from state x to state y and by Y the discrete time homogeneous and irreducible Markov chain canonically embedded in X at its jump times. The transition probability $P(x, y)$ that Y visits state y after state x verifies

$$P(x, y) = \frac{Q(x, y)}{\sum_{z:z\neq x} Q(x, z)}. \tag{1}$$

Let us precise here the main characteristics of the model. We assume that the components are (i) either operational (or up), or (ii) unoperational (or down, that is, failed). The same happens with the whole system. As said before, $S = U \cup D$ where U is the set of up states and D is the set of down states, $U \cap D = \emptyset$. The components have also a *class* or *type* belonging to the set $\mathcal{K} = \{1, 2, \cdots, K\}$ of classes. An operational class k component has failure rate $\lambda_k(x)$ when the model is in state x.

In the sequel, we will basically follow the notation used by Shahabuddin [14] and by Nakayama [12], and the assumptions made there. The whole set of transitions is partitioned into two (disjoint) sets $\mathcal{F}$, the set of *failures* and $\mathcal{R}$, the set of *repairs*. To facilitate the reading, we denote $Q(x, y) = \lambda(x, y)$ when $(x, y) \in \mathcal{F}$ and $Q(x, y) = \mu(x, y)$ when $(x, y) \in \mathcal{R}$. We also denote by F_x the set of states that can be reached from x after a failure, and by R_x the set of states that can be reached from x after a repair, that is,

$$F_x = \{y \mid (x, y) \in \mathcal{F}\}, \quad R_x = \{y \mid (x, y) \in \mathcal{R}\}. \tag{2}$$

Recall that we assume that the initial state is fixed and denoted by 0. Since all the components are up in that state, we assume $0 \in U$. We also have $R_0 = \emptyset$ (that is, no repairs from 0 since everything is assumed to work when the system's state is 0).

Let us denote by $\nu_k(x)$ the number of operational components of class k when the model state is x. The intuitive idea of failure and repair translate into the following formal relationships: for all $x \in S$,

$$(x, y) \in \mathcal{F} \Longleftrightarrow \forall k,\ \nu_k(x) \geq \nu_k(y) \text{ and } \exists k \text{ s.t. } \nu_k(x) > \nu_k(y),$$

$$(x, y) \in \mathcal{R} \Longleftrightarrow \forall k,\ \nu_k(x) \leq \nu_k(y) \text{ and } \exists k \text{ s.t. } \nu_k(x) < \nu_k(y).$$

To finish the description of the model, let us specify how the transitions occur. After the failure of some operational class k component when the system state is x, the system jumps to state y with probability $p(y; x, k)$. This allows to take into account the case of *failure propagation*, that is, the situation where the failure of some component induces, with some probability, that a subset of components is shut down (for instance, the failure of the power supply can make some other components unoperational). The probabilities $p(y; x, k)$ are assumed to be defined for all y, x, k; in general, in most of the cases $p(y; x, k) = 0$.

Observe that

$$\forall (x, y) \in \mathcal{F},\ \ \lambda(x, y) = \sum_{k=1}^{K} \nu_k(x) \lambda_k(x) p(y; x, k). \tag{3}$$

Concerning the repairs, the only needed assumption is that from every state different from the initial one, there is at least one repair transition, that is,

$$\forall x \neq 0,\ R_x \neq \emptyset.$$

This excludes the case of *delayed repairs*, corresponding to systems where the repair facilities are activated only when there are "enough" failed units.

3 Regenerative Monte Carlo scheme

The regenerative approach to evaluate the *MTTF* consists of using the following expression:

$$MTTF = \frac{\mathrm{E}(\min(\tau_D, \tau_0))}{\gamma} \tag{4}$$

where $\gamma = Pr(\tau_D < \tau_0)$ (see Goyal et al. [5]). To estimate $\mathrm{E}(\min(\tau_D, \tau_0))$ and γ, we generate independent cycles C_1, C_2, $\ldots$, that is, sequences of adjacent states starting and ending with state 0, and not containing it in any other position, and we estimate the corresponding expectations.

Observe first that the numerator and the denominator in the r.h.s. of Eq. (4) can be computed directly from the embedded discrete time chain Y, that is, working in discrete time. To formalize this, let us denote by $\mathcal{C}$ the set of all the cycles and by $\mathcal{D}$ the set of the cycles passing through D. The probability of a cycle $c \in \mathcal{C}$ is

$$q(c) = \prod_{(x,y)\in c} P(x,y). \tag{5}$$

An estimator of $MTTF$ is then

$$\widehat{MTTF} = \frac{\sum_{i=1}^{I} G(C_i)}{\sum_{i=1}^{I} H(C_i)} \tag{6}$$

where for any cycle c, we define $G(c)$ as the sum of the expectations of the sojourn times in all its states until reaching D or being back to 0, and $H(c)$ is equal to 1 if $c \in \mathcal{D}$, and to 0 otherwise. Observe that, to estimate the denominator in the expression of the $MTTF$, when a cycle reaches D, the path construction is stopped since we already know that $\tau_D < \tau_0$.

Using the Central Limit Theorem, we have (see Goyal et al. [5])

$$\frac{\sqrt{I}(\widehat{MTTF} - MTTF)}{\sigma/\overline{H}_I} \to N(0,1), \tag{7}$$

with $\overline{H}_I = \dfrac{1}{I}\sum_{i=1}^{I} H(C_i)$, and

$$\sigma^2 = \sigma_q^2(G) - 2\, MTTF\, \mathrm{Cov}_q(G,H) + MTTF^2 \sigma_q^2(H), \tag{8}$$

where $\sigma_q^2(F)$ denotes the variance of F under the probability measure q. A confidence interval can thus be obtained.

The estimation of the numerator in Eq. (4) presents no problem even in the rare event context since in that case $\mathrm{E}(\min(\tau_D, \tau_0)) \approx \mathrm{E}(\tau_0)$. The estimation of γ, however, is difficult or even impossible using the standard Monte Carlo scheme in the rare event case. Indeed, the expectation of the first time that event "$\tau_D < \tau_0$" occurs is about $1/\gamma$, then large for highly reliable systems. For its estimation, we can follow an importance sampling approach. The idea is to change the underlying measure such that *all* the cycles in the interesting set $\mathcal{D}$ receive a higher weight. This is not possible in general, and what we in fact do is to change the transition probabilities $P()$'s into $P'()$'s with an appropriate choice such that we expect that the weight $q'()$ of *most of the interesting cycles* will increase.

The following method, called MS-DIS for *Measure Specific Dynamic Importance Sampling* and introduced in Goyal et al. [5], uses independent simulations for the numerator and denominator of Eq. (4). On the total I sample paths, ξI are reserved for the estimation of $E(\min(\tau_D, \tau_0))$ and $(1-\xi)I$ for the estimation of γ. As the estimation of $E(\min(\tau_D, \tau_0))$

is simple, we use the crude estimator $\overline{G}_{\xi I} = (\xi I)^{-1} \sum_{i=1}^{\xi I} G(C_{i,q})$ where $C_{i,q}$, is the ith path sampled under probability measure q. The importance sampling technique is applied to the estimation of γ. A new estimator of the $MTTF$ is then

$$\widehat{MTTF} = \frac{\overline{G}_{\xi I}}{\overline{H}_{(1-\xi)I}} \tag{9}$$

with

$$\overline{H}_{(1-\xi)I} = ((1-\xi)I)^{-1} \sum_{i=1}^{(1-\xi)I} H(C_i) \frac{q(C_{i,q'})}{q'(C_{i,q'})} \tag{10}$$

using independent paths $C_{i,q'}$, $1 \le i \le (1-\xi)I$, sampled under the new probability measure q', and independent of the $C_{i,q}$, $1 \le i \le \xi I$. We have then

$$\frac{\sqrt{I}(\widehat{MTTF} - MTTF)}{\sigma / \overline{H}_{(1-\xi)I}} \to N(0,1) \tag{11}$$

with

$$\sigma^2 = \frac{\sigma_q^2(G)}{\xi} + (MTTF)^2 \frac{\sigma_{q'}^2(Hq/q')}{1-\xi}. \tag{12}$$

A dynamic choice of ξ can also be made to reduce σ^2.

In Section 5, we review the main schemes proposed for the estimation of γ, and we propose some new ways of performing the estimations, which will be shown to behave better in appropriate situations. Next section first discusses the formalization of the rare event situation, in order to be able to develop the analysis of those techniques.

4 The rarity parameter

We must formalize the fact that failures are rare or slow, and that repairs are fast. Following Shahabuddin [14], we introduce a *rarity parameter* ε. We assume that the failure rates of class

k components have the following form:

$$\lambda_k(x) = a_k(x)\varepsilon^{i_k(x)} \tag{13}$$

where either the real $a_k(x)$ is strictly positive and the integer $i_k(x)$ is greater than or equal to 1, or $a_k(x) = i_k(x) = 0$. To simplify things, we naturally set $a_k(x) = 0$ if $\nu_k(x) = 0$. No particular assumption is necessary about the $p(y; x, k)$'s, so, we write

$$p(y; x, k) = b_k(x, y)\varepsilon^{j_k(x,y)} \tag{14}$$

with real $b_k(x, y) \geq 0$, integer $j_k(x, y) \geq 0$, and $j_k(x, y) = 0$ when $b_k(x, y) = 0$. Concerning the repair rates, we simply state

$$\mu(x, y) = \Theta\left(1\right), \tag{15}$$

where $f(\varepsilon) = \Theta\left(\varepsilon^d\right)$ means that there exists two constants $k_1, k_2 > 0$ such that $k_1\varepsilon^d \leq |f(\varepsilon)| \leq k_2\varepsilon^d$ (recall that for every state $x \neq 0$, there exists at least one state y s.t. $\mu(x, y) > 0$). We can thus observe that the rarity of the interesting event "$\tau_D < \tau_0$" increases when ε decreases.

The form of the failure rates of the components has the following consequence on the failure transitions in X: for all $(x, y) \in \mathcal{F}$,

$$\lambda(x, y) = \Theta\left(\varepsilon^{m(x,y)}\right) \tag{16}$$

where

$$m(x, y) = \min_{k: a_k(x) b_k(x,y) > 0} \{i_k(x) + j_k(x, y)\} \tag{17}$$

(observe that if $F_x \neq \emptyset$, then for all $y \in F_x$ we necessarily have $m(x, y) \geq 1$).

Let us look now at the transition probabilities of Y. For any $x \neq 0$, since we assume that $R_x \neq \emptyset$, we have

$$(x, y) \in \mathcal{F} \Longrightarrow P(x, y) = \Theta\left(\varepsilon^{m(x,y)}\right),\ m(x, y) \geq 1, \tag{18}$$

and

$$(x, y) \in \mathcal{R} \Longrightarrow P(x, y) = \Theta(1). \tag{19}$$

For the initial state, we have that for all $y \in F_0$,

$$P(0, y) = \Theta\left(\varepsilon^{m(0,y) - \min_{z \in F_0} m(0,z)}\right). \tag{20}$$

Observe here that if $\mathrm{argmin}_{z \in F_0} m(0, z) = w \in D$, then we have $P(0, w) = \Theta(1)$ and there is no rare event problem. This happens in particular if $F_0 \cap U = \emptyset$. So, the interesting case for us (the rare event situation) is the case of $P(0, w) = o(1)$ for all $w \in F_0 \cap D$. In other words, the case of interest is when (i) $F_0 \cap U \neq \emptyset$ and (ii) $\exists y \in F_0 \cap U$ s.t. $\forall \omega \in F_0 \cap D$, $m(0, y) < m(0, \omega)$.

A simple consequence of the previous assumptions is that for any cycle c, its probability $q(c)$ is $q(c) = \Theta\left(\varepsilon^h\right)$ where the integer h is $h \geq 1$. If we define

$$\mathcal{C}_h = \{c \in \mathcal{C} \mid q(c) = \Theta\left(\varepsilon^h\right)\}, \tag{21}$$

then we have (see Shahabuddin [14])

$$\gamma = \sum_{c \in \mathcal{D}} q(c) = \Theta(\varepsilon^r) \tag{22}$$

where $r = \mathrm{argmin}\{h \mid \mathcal{C}_h \neq \emptyset\} \geq 1$. We see formally now that γ decreases as $\varepsilon \to 0$.

5 Importance sampling schemes

In this section, we describe different importance sampling schemes for analyzing highly reliable Markovian systems. Some of the methods (those presented in subsections 5.1, 5.2, 5.5, 5.6 and the first method of subsection 5.7) have already been presented in the literature; the remaining ones are new contributions, presented here for the first time.

To simplify the description of the different schemes, let us introduce the following notation. For any state x, we denote by $f_x(y)$ the transition probability $P(x, y)$, for each $y \in F_x$. In the same way, for any state x, let us denote $r_x(y) = P(x, y)$ for each $y \in R_x$. Using an importance sampling scheme means that instead of P we use a different matrix P', leading to new $f'_x()$'s and $r'_x()$'s. The transition probabilities associated with the states of D are not concerned in the estimation of γ since when a cycle reaches D, it is "stopped" as we explained in Section 3.

5.1 Failure biasing (FB) (Lewis and Böhm [8], Conway and Goyal [4])

This is the most straightforward method: to increase the probability of regenerative cycles including system failures, we increase the probability of the failure transitions. We must choose a parameter $\alpha \in (0, 1)$, which is equal to $f'_x(F_x)$ for all $x \neq 0$ (typically, $0.5 \leq \alpha \leq 0.9$). The transition probabilities are then changed as follows.

- $\forall x \in U,\ x \neq 0,\ \forall y \in F_x : \quad f'_x(y) = \alpha \dfrac{f_x(y)}{f_x(F_x)};$
- $\forall x \in U,\ x \neq 0,\ \forall y \in R_x : \quad r'_x(y) = (1 - \alpha) \dfrac{r_x(y)}{r_x(R_x)}.$

The $f_0()$'s are not modified (since we already have $f_0(F_0) = \Theta(1)$). Observe that the total probability of failure from x is now equal to α (that is, for any $x \in U - \{0\}$, $f'_x(F_x) = \alpha$).

5.2 Selective failure biasing (SFB) (Goyal et al. [5])

The idea here is to separate the failure transitions from x $(x \in U)$ into two disjoint sets: those consisting of the *first* failure of a component of some class k (and called *initial failures*), and the remaining ones (called *non-initial failures*). Following this, the set of states F_x is partitioned into two (disjoint) sets IF_x and NIF_x, where

$$IF_x = \{y \mid (x, y) \text{ is an } \textit{initial} \text{ failure}\},$$

$$NIF_x = \{y \mid (x, y) \text{ is a } \textit{non-initial} \text{ failure}\}.$$

The idea is then to increase the probability of a non-initial failure, that is, to make the failure of some class k components more probable than in the original model, if there is at least one component of that class that has already failed.

To implement this, we must choose two parameters $\alpha, \beta \in (0, 1)$ (typically, $0.5 \le \alpha, \beta \le 0.9)$ and change the transition probabilities in the following way:

- $\forall x \in U,\ x \neq 0,\ \forall y \in IF_x,\ \ f'_x(y) = \alpha(1-\beta)\dfrac{f_x(y)}{f_x(IF_x)},$
  and $\forall y \in NIF_x,\ \ f'_x(y) = \alpha\beta\dfrac{f_x(y)}{f_x(NIF_x)};$

  for $x = 0$, we use the same formulae with $\alpha = 1$; in the same way, if $IF_x = \emptyset$, we use $\beta = 1$ and if $NIF_x = \emptyset$, we set $\beta = 0$.

- $\forall x \in U,\ x \neq 0,\ \forall y \in R_x,\ \ r'_x(y) = (1-\alpha)\dfrac{r_x(y)}{r_x(R_x)}.$

In this scheme, as in the FB method, the total failure probability from x is $f'_x(F_x) = \alpha$, but now we have a further refinement, leading to $f'_x(NIF_x) = \alpha\beta$ and $f'_x(IF_x) = \alpha(1-\beta)$.

5.3 Selective failure biasing for "series-like" systems (SFBS)

The implicit assumption in SFB is that the criteria used to define an operational state (that is, the type of considered system) is close to the situation where the system is up if and only if, for each component class k, the number of operational components is greater or equal some threshold l_k, and if neither the initial number of components N_k nor the level l_k are "very dependent" on k. Now, assume that this last part of the assumptions does not hold, that is, assume that from the dependability point of view, the system is a series of l_k-out-of-N_k modules, but that the N_k's and the l_k's are strongly dependent on k. A reasonable way to improve SFB is to make more probable the failures of the class k components for which $\nu_k(x)$ is closer to the threshold l_k.

Consider a state $x \in U$ and define a class k *critical in* x if $\nu_k(x)-l_k = \min_{k'=1,\cdots,K} (\nu_{k'}(x) - l_{k'})$; otherwise, the class is *non-critical*. Now, for a state $y \in F_x$, the transition (x, y) is *critical* if there is some critical class k in x such that $\nu_k(y) < \nu_k(x)$. We denote by $F_{x,c}$ the subset of F_x composed of the critical failures, that is,

$$F_{x,c} = \{y \in F_x \mid (x, y) \text{ is critical } \}.$$

We also define $F_{x,nc}$, the set of *non-critical* failures, by $F_{x,nc} = F_x - F_{x,c}$. Then, a specialized SFB method, which we call SFBS, can be defined by the following modification of the $f_x()$'s (we omit the frontiers' case which is handled as for SFB):

- $\forall x \in U,\ \forall y \in F_{x,nc},\ \ f'_x(y) = \alpha(1-\beta)\dfrac{f_x(y)}{f_x(F_{x,nc})}$,
  and $\forall y \in F_{x,c},\ \ f'_x(y) = \alpha\beta\dfrac{f_x(y)}{f_x(F_{x,c})}$.
- $\forall y \in R_x,\ \ r'_x(y) = (1-\alpha)\dfrac{r_x(y)}{r_x(R_x)}$.

See Section 7 for the numerical behavior of this method and the the gain that can be obtained when using it instead of SFB.

5.4 Selective failure biasing for "parallel-like" systems (SFBP)

This is the dual of SFBS. Think of a system working as sets of l_k-out-of-N_k modules in parallel, $1 \leq k \leq K$. Consider a state $x \in U$ and define a class k *critical in* x if $\nu_k(x) \geq l_k$; otherwise, the class is *non-critical*. Now, for a state $y \in F_x$, the transition (x, y) is *critical* if there is some critical class k in x such that $\nu_k(y) < \nu_k(x)$. As before, the set of states $y \in F_x$ such that (x, y) is critical, is denoted by $F_{x,c}$, and $F_{x,nc} = F_x - F_{x,c}$.

A first idea is to follow the analogous scheme as for the SFBS case: using in the same way two parameters α and β, the principle would be to accelerate the critical transitions first, then the non-critical ones, by means of the respective weights $\alpha\beta$ and $\alpha(1-\beta)$. This leads to the following rules:

- $\forall x \in U,\ \forall y \in F_{x,nc},\ \ f'_x(y) = \alpha(1-\beta)\dfrac{f_x(y)}{f_x(F_{x,nc})}$,
  and $\forall y \in F_{x,c},\ \ f'_x(y) = \alpha\beta\dfrac{f_x(y)}{f_x(F_{x,c})}$.
- $\forall y \in R_x,\ \ r'_x(y) = (1-\alpha)\dfrac{r_x(y)}{r_x(R_x)}$.

As we will see in Section 7, there is no need for the β parameter, and the method we call SFBP is then defined by the following rules:

- $\forall x \in U,\ \forall y \in F_{x,c},\ \ f'_x(y) = \alpha\dfrac{f_x(y)}{f_x(F_{x,c})}$,
  and $\forall y \in F_{x,nc},\ \ f'_x(y) = (1-\alpha)\dfrac{f_x(y)}{r_x(F_x) + f_x(F_{x,nc})}$.

- $\forall y \in R_x, \;\; r'_x(y) = (1-\alpha)\dfrac{r_x(y)}{r_x(F_x) + f_x(F_{x,nc})}$.

As we see, we only accelerate the critical transitions, the non-critical ones are handled in the same way as the repairs.

5.5 Inverse Failure Biasing (IFB) (Papadopoulos [13])

IFB has been inspired from the importance sampling theory applied to the $M/M/1$ queue [6] where service and arrival rates are exchanged. The rule is the following:

- If $x = 0, \forall y \in F_0, \;\; f'_0(y) = \dfrac{1}{|F_0|}$.
- $\forall x \in U, \; x \neq 0 \; \forall y \in F_x, \;\; f'_x(y) = \dfrac{r_x(R_x)}{|F_x|}$,
  and $\forall y \in R_x, \;\; r'_x(y) = \dfrac{f_x(F_x)}{|R_x|}$.

With these new transition probabilities, repairs are $O(\varepsilon)$; then they are less likely to occur. This scheme should be efficient when applied to a system such that all important paths to failure (those in $\mathcal{C}_r$ where r is defined in Equation (22)) are paths without any repair. If this is not the case, from theoretical considerations (developed in Section 6) we believe that this scheme will perform poorly; we will investigate this numerically in Section 8.

5.6 Distance-based selected failure biasing (DSFB) (Carrasco [3])

We assume that there may be some propagation of failures in the system. For all $x \in U$, its distance $d(x)$ to D is defined as the minimal number of components whose failure put the model in a down state, that is,

$$d(x) = \min_{y \in D} \sum_k \left(\nu_k(x) - \nu_k(y)\right).$$

Obviously, for any $y \in F_x$ we have $d(y) \le d(x)$. A failure (x, y) is said *dominant* if and only if $d(x) > d(y)$ and it is *non-dominant* iff $d(x) = d(y)$. The *criticality* of $(x, y) \in \mathcal{F}$ is

$$c(x, y) = d(x) - d(y) \ge 0.$$

The idea of this algorithm is to take into account the different criticalities to control more deeply the failure transitions in the importance sampling scheme. It is assumed, of course, that the user can compute the distances $d(x)$ for any operational state x with low cost.

Define recursively the following partition of F_x:

$$F_{x,0} = \{y \in F_x \mid c(x, y) = 0\},$$

and $F_{x,l}$ is the set of states $y \in F_x$ such that $c(x, y)$ is the smallest criticality value greater than $c(x, w)$ for any $w \in F_{x,l-1}$. In symbols, if we denote, for all $l \ge 1$,

$$G_{x,l} = F_x - F_{x,0} - F_{x,1} - \cdots - F_{x,l-1},$$

then we have

$$F_{x,l} = \{y \in G_{x,l} \mid y = \text{argmin}\{c(x, z),\ z \in G_{x,l}\}\}.$$

Let us denote by V_x the number of criticality values greater than 0 of failures from x, that is,

$$V_x = \text{argmax}\{l \ge 1 \mid F_{x,l} \ne \emptyset\}.$$

The method proposed by Carrasco [3] has three parameters $\alpha, \beta, \beta_c \in (0, 1)$. The new probability transitions are

- $\forall x \in U,\ \forall y \in F_{x,0},\quad f'_x(y) = \alpha(1-\beta)\dfrac{f_x(y)}{f_x(F_{x,0})};$
  $\forall l$ s.t. $1 \le l < V_x,\ \forall y \in F_{x,l},\quad f'_x(y) = \alpha\beta(1-\beta_c)\beta_c^{l-1}\dfrac{f_x(y)}{f_x(F_{x,l})};$
  $\forall y \in F_{x,V_x},\quad f'_x(y) = \alpha\beta\beta_c^{V_x-1}\dfrac{f_x(y)}{f_x(F_{x,V_x})};$

- $\forall x \neq 0,\ \forall y \in R_x,\ \ r'_x(y) = (1-\alpha)\dfrac{r_x(y)}{r_x(R_x)}.$

As before, we must define what happens at the "frontiers" of the transformation. If $F_{x,0} = \emptyset$, then we use $\beta = 1$. If $x = 0$, then we set $\alpha = 1$.

It seems intuitively clear that we must, in general, give a higher weight to the failures with higher criticalities. This is not the case of the approach originally proposed by Carrasco [3].

Just by "inverting" the order of the weights of the failures arriving at the $F_{x,l}$, $l \geq 1$, we obtain a new version which gives higher probabilities to failure transitions with higher criticalities. The Distance-based Selective Failure Biaising (DSFB) which we define here corresponds to the following algorithm:

- $\forall x \in U,\ \forall y \in F_{x,0},\ \ f'_x(y) = \alpha(1-\beta)\dfrac{f_x(y)}{f_x(F_{x,0})};$
  $\forall y \in F_{x,1},\ \ f'_x(y) = \alpha\beta\beta_c^{V_x - 1}\dfrac{f_x(y)}{f_x(F_{x,1})};$
  $\forall l$ s.t. $2 \leq l \leq V_x,\ \forall y \in F_{x,l},\ \ f'_x(y) = \alpha\beta(1-\beta_c)\beta_c^{V_x - l}\dfrac{f_x(y)}{f_x(F_{x,l})};$
- $\forall x \neq 0,\ \forall y \in R_x,\ \ r'_x(y) = (1-\alpha)\dfrac{r_x(y)}{r_x(R_x)}.$

5.7 Balanced methods

If we except IFB, the previous methods classify the transitions from a fixed state into a number of disjoint sets, and assign modified global probabilities to each of these sets; but they do not modify the relative weights of the transitions belonging to the same set. An alternative is to assign uniform probabilities to all transitions from x leading to the same subset of F_x. This can be done independently of the number and the definition of those sets, so that we can find *balanced versions* of all the previously mentioned methods with the only exception of IFB, as already stated.

Before looking the balanced versions in detail, let us observe that sometimes the systems are already "balanced" themselves, that is, there are no significant differences between the magnitude of the transition probabilities. In these cases, the unbalanced and balanced versions of the same method will basically behave in the same manner.

Balanced FB

Analyzing the FB method, it was proved (first by Shahabuddin in [14]) that balancing it improves its behaviour when there are transition probabilities from the same state x which differ by orders of magnitude. The Balanced FB method is then defined by

- $\forall x \neq 0,\ \forall y \in F_x,\ \ f'_x(y) = \alpha \dfrac{1}{|F_x|}$;
- $\forall x \neq 0,\ \forall y \in R_x,\ \ r'_x(y) = (1-\alpha)\dfrac{r_x(y)}{r_x(R_x)}$.

If $x = 0$, then we set $\alpha = 1$ in the algorithm.

Balanced SFB

The Balanced SFB scheme consists of the following rules:

- $\forall x \neq 0,\ \forall y \in IF_x,\ \ f'_x(y) = \alpha(1-\beta)\dfrac{1}{|IF_x|}$,
  and $\forall y \in NIF_x,\ \ f'_x(y) = \alpha\beta\dfrac{1}{|NIF_x|}$;

  for $x = 0$, we use the same formulae with $\alpha = 1$; in the same way, if $IF_x = \emptyset$, we use $\beta = 1$ and if $NIF_x = \emptyset$, we set $\beta = 0$.

- $\forall x \neq 0,\ \forall y \in R_x,\ \ r'_x(y) = (1-\alpha)\dfrac{r_x(y)}{r_x(R_x)}$.

Balanced SFBS

We describe now the transformations associated with the Balanced SFBS scheme, except for the repairs and the frontier cases, which are as in the Balanced SFB's method:

- $\forall x,\ \forall y \in F_{x,nc},\ \ f'_x(y) = \alpha(1-\beta)\dfrac{1}{|F_{x,nc}|}$,
  and $\forall y \in F_{x,c},\ \ f'_x(y) = \alpha\beta\dfrac{1}{|F_{x,c}|}$.

Balanced SFBP

The Balanced SFBP method is defined by the following rules:

- $\forall x \in U,\ \forall y \in F_{x,c},\ \ f'_x(y) = \alpha\dfrac{1}{|F_{x,c}|}$,
  and $\forall y \in F_{x,nc},\ \ f'_x(y) = (1-\alpha)\dfrac{1}{|R_x| + |F_{x,nc}|}$.
- $\forall y \in R_x,\ \ r'_x(y) = (1-\alpha)\dfrac{1}{|R_x| + |F_{x,nc}|}$.

It can be observed that, for the Balanced SFBP scheme, we do not take the repair probabilities proportionally to the original ones. Indeed, we have grouped repairs and non-initial failures, so taking the new transition probabilities proportional to the original ones would give rare events for the non-initial failures. Thus this small change, i.e. a uniform distribution over $F_{x,nc} \cup R_x$, balances all the transitions.

Balanced DSFB

The Balanced DSFB scheme is

- $\forall x \in U,\ \forall y \in F_{x,0},\ \ f'_x(y) = \alpha(1-\beta)\dfrac{1}{|F_{x,0}|}$ (as for B_3);
  $\forall y \in F_{x,1},\ \ f'_x(y) = \alpha\beta\beta_c^{V_x-1}\dfrac{1}{|F_{x,1}|}$;
  $\forall l$ s.t. $2 \le l \le V_x,\ \forall y \in F_{x,l},\ \ f'_x(y) = \alpha\beta(1-\beta_c)\beta_c^{V_x-l}\dfrac{1}{|F_{x,l}|}$;
- $\forall x \neq 0,\ \forall y \in R_x,\ \ r'_x(y) = (1-\alpha)\dfrac{r_x(y)}{r_x(R_x)}$.

6 Bounded relative error and bounded normal approximation

Shahabuddin [14] defines the concept of bounded relative error as follows:

Definition 6.1 *Let σ^2 denote the variance of the estimator of γ and z_δ the $1-\delta/2$ quantile of the standard normal distribution. Then the relative error for a sample size M is defined by*

$$RE = z_\delta \frac{\sqrt{\sigma^2/M}}{\gamma}. \tag{23}$$

We say that we have a bounded relative error (BRE) if RE remains bounded as $\varepsilon \to 0$.

If the estimator enjoys this property, only a fixed number of iterations is required to obtain a confidence interval having a fixed error no matter how rarely failures occur.

Tuffin [16, 15] introduced the concept of bounded normal approximation to justify the use of the central limit theorem. Recall first the following version of the Berry-Esseen Theorem (Bentkus and Götze [1]).

For a random variable Z, let $\varrho = E(|Z-E(Z)|^3)$, $\sigma^2 = E((Z-E(Z))^2)$ and let $\mathcal{N}$ be the standard normal distribution. For $Z_1, \cdots, Z_I$ i.i.d. copies of Z, define $\overline{Z}_I = I^{-1}\sum_{i=1}^{I} Z_i$,

$\hat{\sigma}_I^2 = I^{-1} \sum_{i=1}^{I} (Z_i - \overline{Z}_I)^2$ and let F_I be the distribution of the centered and normalized sum $(Z_1 + \cdots + Z_I)/(\hat{\sigma}_I \sqrt{I}) - E(Z)\sqrt{I}/\hat{\sigma}_I$. Then there exists an absolute constant $a > 0$ such that, for each x and I

$$|F_I(x) - \mathcal{N}(x)| \leq \frac{a\varrho}{\sigma^3 \sqrt{I}}. \tag{24}$$

Thus it is interesting to control the quantity ϱ/σ^3 because, in this way, the validity of the normal approximation, and then, of the coverage of the confidence interval, is guaranteed. A discussion on this point can be found in the work by Tuffin [16, 15]. Following [16], we define the bounded normal approximation as follows.

Definition 6.2 *If ϱ denotes the third order moment and σ the standard deviation of the estimator of γ, we say that we have a bounded normal approximation (BNA) if ϱ/σ^3 is bounded when $\varepsilon \to 0$.*

Necessary and sufficient conditions for both properties are known (see Nakayama [12] for BRE and Tuffin [15, 16] for BNA). It has been proven by Nakayama [12] (see also Shahabuddin [14]) that Balanced FB leads to the BRE property and it has been also shown that this is not true for unbalanced methods. Similarly, from Shahabuddin's work [14], or using Theorem 2 in Nakayama [12], it can be shown that any of the balanced algorithms gives BRE.

In [13], it is argued that IFB verifies BRE for balanced systems as well as for some other classes. Unfortunately, this stands only if the paths in $\mathcal{C}_r$ are direct paths to failure, i.e., do not include any repair transition. This property does not hold in general for balanced systems; and can be difficult to check in a particular case. A simple counter-example would be the case of a system made up of two classes of components, say for example CPUs and disks, with two units in each class and with failure propagations from the CPUs to the disks (see Figure 1 where those failures propagations are given by the downward vertical links). All failure rates are assumed to be $\Theta(\varepsilon)$. Thus $\gamma = \Theta(\varepsilon^r)$ with $r = 1$. Here, a path including a failure of a CPU which contaminates a disk, then a repair of the CPU, and finally a failure of the still operational disk (the dashed path in Figure 1), belongs to $\mathcal{C}_r$ but includes a repair transition, which contradicts [13]'s assumption. In this example, the IFB method would not have the BRE property (contrary to what is stated in [13]).

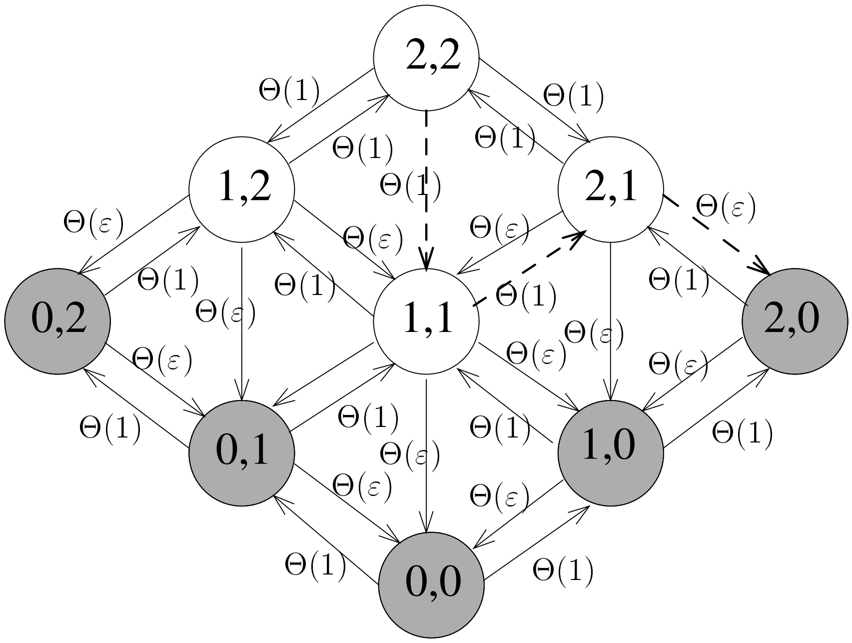

Figure 1: Counter example for IFB. A state i, j means that i CPUs and j disks are up. D is composed by the states in grey. The transition probabilities of the embedded DTMC are given.

Returning to the general discussion, the following result shows that BRE and BNA are not independent.

Theorem 6.3 *(from Tuffin [15, 16]). If we have BNA, we have BRE. Nevertheless, there exist systems with BRE but without BNA.*

This means that we must not only check for the BRE property: the critical one is BNA. It is also proven by Tuffin [16, 15] that any balanced method verifies the BNA property, so balancing all the methods leads to good properties. Using the necessary and sufficient conditions for BRE and BNA, i.e. using Theorem 2 in Nakayama's paper [12] and Theorem 4 in Tuffin's paper [16], it is immediate to see that, in fact, any change of measure independent of the rarity parameter ε verifies the BRE and BNA properties (for the BRE property, this has been observed first by Shahabuddin in [14]).

7 Asymptotic comparison of methods

Given a specified system, we can wonder which scheme, among the several ones described in Section 5, is the most appropriate. This section has two folds. First, we explain why we do not use a β parameter in the SFBP scheme, as we do in the SFB and SFBS cases. Second, we make some asymptotic comparisons of the discussed techniques. We consider only balanced schemes because they are the only ones, among the methods described in Section 5, to verify in general the desirable BRE and BNA properties.

The asymptotic efficiency (as $\varepsilon \to 0$) is controlled by two quantities: the asymptotic variance of the estimator and the mean number of transitions needed by embedded chain Y to hit D when it does it before coming back to 0.

7.1 On the SFBP choice

- We want to compare the variance of the two considered choices in SFBP (with or without a β parameter), in the case of a system structured as a set of l_k-out-of-N_k modules in parallel, $k = 1, \cdots, K$, i.e. the case of interest. To do this, let us denote by $f'_{x,\beta}(y)$ the transition probability associated with a SFBP scheme using a β parameter, as shown before, in the first part of Eq. 5.4. Let s be the integer such that $\sigma^2 = \Theta(\varepsilon^s)$. We can observe that the most important paths for the variance estimation, i.e. the paths $c \in \mathcal{D}$ verifying $q^2(c)/q'(c) = \Theta(\varepsilon^s)$ are typically composed of critical transitions (x, y) for which the failure SFBP probability $f'_x(y)$ (without using β) verify

$$f'_x(y) = f'_{x,\beta}(y)/\beta, \tag{25}$$

i.e., transitions driving closer to the failure states. So, if we note σ^2 (resp. σ^2_β) the variance of the estimator without (resp. with) the β parameter, $\sigma^2/\sigma^2_\beta < 1$ as $\varepsilon \to 0$.

- Let us denote by $|c|$ the number of transitions in cycle $c \in \mathcal{D}$ until hitting D. The expected number of transitions necessary to hit D under the modified measure q' is

$$E(T) = \sum_{c \in \mathcal{D}} |c| q'(c). \tag{26}$$

¿From Equation (25), we see that $E(T)$ is smaller if we do not use the β parameter.

¿From both of these points of view, we conclude that not using a β parameter in SFBP scheme is a good idea.

7.2 Comparison of Balanced schemes

Using the balanced schemes, all the variances are of the same order (i.e. $O(\varepsilon^{2r})$) because each path is in $\Theta(1)$ (see Shahabuddin [14] or Nakayama [12] for a proof). Then, we can point out the following facts:

- The variances are of the same order with all the balanced schemes. Nevertheless the constant of the $\Theta(\cdot)$ may be quite different. The analysis of this constant is much more difficult in this general case than for the SFBP schemes previously presented and appears to depend too much on the specific model parameters to allow any kind of general claim about it.

- The preceding point suggests basing the choice between the different methods mainly on the mean hitting time to D given in Eq. (26). To get the shortest computational time, our heuristic is the following:

  - if there are many propagation faults in the system, we suggest the use of a Balanced DSFBP scheme;
  - if there is no (or very few) propagation faults and if the system is working as a series of l_k-out-of-N_k modules, the balanced SFBS scheme seems the appropriate one;
  - if there no (or very few) propagation faults and if the system is working as a set of l_k-out-of-N_k modules in parallel, $1 \leq k \leq K$, we suggest the use of the Balanced SFBP method;
  - there remains the case of a poorly structured system, or one where it is not clear if the structure function is rather of the series type, or of the parallel one; in those cases, the general Balanced FB scheme can also lead to a useful variance reduction.

8 Numerical illustrations

All the systems used in the numerical illustrations given in this section were modeled and evaluated using a specific library (called BB as *balls & buckets* framework) (from Cancela [2]), on a SPARCstation 10 Model 602 workstation. In all cases, the estimated measure is $\gamma = \Pr(\tau_D < \tau_0)$.

We are not going to compare all the methods discussed before in both versions, unbalanced and balanced. Our aim is to get a feeling of what can be obtained in practice, and to give some general guidelines to choose among the different methods.

First, let us consider methods FB, IFB, SFB and SFBS. When the modeled systems have a structure close to a series of l_k-out-of-N_k modules, it seems clear that both SFB and SFBS are better than FB. If the values $N_k - l_k$ (that is, the number of redundant components for class k) do not (or do slightly) depend on k, SFB and SFBS should have more or less the same behaviour; but when some components have significant differences in these values, SFBS should outperform SFB. To look at how these rules of thumb work out in a particular case, we study two versions of a Tandem computer, described by Katzmann in [7] (we follow here a later description made by Liceaga and Siewiorek [9]). This computer is composed of a multiprocessor p, a dual disk controller k, two RAID disk drives d, two fans f, two power supplies ps, and one dual interprocessor bus b. In addition to a CPU, each processor contains its own memory. When a component in a dual fails, the subsystem is reconfigured into a simplex. This Tandem computer requires all subsystems, one fan, and one power supply for it to be operational. The failure rates $\lambda_k(x)$ are 5ε, 2ε, 4ε, 0.1ε, 3ε and 0.3ε for the processors, the disk controller, the disks, the fans, the power supplies and the bus respectively, with $\varepsilon = 10^{-5}$ f/hr. There is only one repairman and the repair rates are $\mu_k(x) = 30$ r/hr, for all the components, except for the bus, which has repair rate $\mu_k(x) = 15$ r/hr.

We first consider a version of this computer where both the multiprocessor and the disks have two units, and only one is needed for the system to be working. In this case, $N_k = 2$ and $l_k = 1$ for all k. Table 1 presents the variances and computing times for the FB, the IFB, the SFB and the SFBS methods, observed when estimating γ with a sample size $M = 10^5$, and parameters $\alpha = 0.7$, $\beta = 0.8$. As expected, we can observe that for this situation, algorithms SFB and SFBS are equivalent (both in precision and in execution time); their variance is an order of magnitude better than the variance of the FB algorithm, which is also slower. The slight difference in the execution time between SFB and SFBS comes from the fact that in the latter there is a little bit of supplementary computations to do, with basically the same cycle structure.

The performance of the IFB method was close to that of the FB algorithm in these tests. This good behavior is not surprising since this is a favorable case for IFB. The IFB method has a slightly better precision than the FB algorithm, with esentially the same execution time.

Method	Variance	Time (sec.)
FB	2.98×10^{-16}	92
IFB	1.92×10^{-16}	94
SFB	3.43×10^{-17}	48
SFBS	3.43×10^{-17}	53

Table 1: Methods FB, IFB, SFB, and SFBS for a series h_k-out-of-N_k system with no dependence on k

Let us now consider this same architecture, but with with a four-unit multiprocessor (only one of the four processors is required to have an operational system); and with each RAID being composed by 5 drives, only 3 of which are required. In this case, N_k and l_k vary for different k. Table 2 presents the variances and computing times for the FB, the IFB, the SFB and the SFBS methods, observed when estimating γ with a sample size $M = 10^5$, and parameters $\alpha = 0.7$, $\beta = 0.8$. As in the previous case, the FB and the IFB algorithms are the least performant; but now we observe how SFBS obtains a better precision (at a lower computational cost) than SFB.

Method	Variance	Time (sec.)
FB	5.90×10^{-18}	131
IFB	3.90×10^{-18}	124
SFB	9.23×10^{-19}	69
SFBS	6.20×10^{-19}	61

Table 2: Methods FB, IFB, SFB and SFBS for a series h_k-out-of-N_k system with dependence on k

Consider now a model of a replicated database; there are four sites, and each site has a whole copy of the database, on a RAID disk cluster. We take all clusters identical, with the same redundancies (7-out-of-9), and with failure rate (for each disk) of $\varepsilon = 10^{-2}$. There is one repairman per class, and the repair rate is 1. We consider that the system is up if there is at least one copy of the database in working order: then the structure function of this system is a parallel l_k-out-of-N_k. We compare in Table 3 the behaviour of FB, IFB, SFB, and SFBP algorithms for this system, where all component classes k have the same redundancy; the SFBP method performs much better than both FB and IFB, which in their turn are better than SFB (this is expected because SFB works for a series-like structure function).

Method	Variance	Time (sec.)
FB	2.17×10^{-27}	320
IFB	1.73×10^{-27}	230
SFB	8.74×10^{-26}	353
SFBP	8.89×10^{-28}	221

Table 3: Methods FB, IFB, SFB, and SFBP for a parallel h_k-out-of-N_k system with no dependence on k

Consider now a model with failure propagation, the fault-tolerant database system presented by Muntz et al. in [11]. The components of this system are: a front-end, a database, and two processing subsystems formed by a switch, a memory, and two processors. These components may fail with rates 1/2400, 1/2400, 1/2400, 1/2400 and 1/120 respectively. There is a single repairman who gives priority to the front-end and the database, followed by the switches and memory units, followed by the processors; all with repair rate 1. If a processor fails it contaminates the database with probability .001. The systems is operational if the front-end, the database, and a processing subsystem are up; a processing subsystem is up if the switch, the memory, and a processor are up. We illustrate in Table 4 the results obtained with the FB, IFB, SFB, and DSFB techniques using $\alpha = 0.7$, $\beta = 0.8$, $\beta_c = 0.2$, for a sample size $M = 10^5$. The DSFB technique is much superior in this context, both in precision (a two-order reduction in the variance) and in computational effort. Its reduced execution time is due to the fact that, going much faster to the states where the system is down than with the other methods, the cycle lengths are much shorter.

Method	Variance	Time (sec.)
FB	1.014×10^{-6}	108
IFB	1.056×10^{-6}	44
SFB	1.016×10^{-6}	105
DSFB	2.761×10^{-8}	41

Table 4: Methods FB, IFB, SFB, and DSFB for a system with failure propagations

Our last example illustrates the use of simulation techniques to evaluate a model with a very large state space. The system is similar to one presented in [5], but has more components and as a result the underlying Markov chain has a larger state space. The system is composed of two sets of 4 processors each, 4 sets of 2 dual-ported controllers, and 8 sets of disk arrays composed by 4 units. Each controller cluster is in charge of 2 disk arrays; each processor has access to all the controllers clusters. The system is up if there are at least one processor (of either class), one controller of each cluster, and three disks of each array, in operational order.

The failure rates for the processors and the controllers are 1/2000; for the disk arrays we consider four different failure rates (each corresponds to two arrays), namely 1/4000, 1/5000, 1/8000 and 1/10000. We consider a single case of failure propagation: when a processor of a cluster fails, there is a 0.10 probability that a processor of the other cluster is affected. Each failure has two modes; the repair rates depend on the mode, and take the value 1 for the first mode and 0.5 for the second.

The system has more than 7.4×10^{14} states in its state space; this precludes even the generation of the state space, and makes it impossible to think of using exact techniques.

We illustrate in Table 5 the results obtained with the crude, FB, IFB, SFB and DSFB techniques using $\alpha = 0.7$, $\beta = 0.8$, $\beta_c = 0.2$, for a sample size $M = 10^5$. Since this is a complex case the execution times are larger than those observed for the previous cases; but even the slowest method, FB, takes less than 27 minutes to complete the experiment. In all cases, when using importance sampling techniques the variances obtained are between 2 and 3 orders of magnitude smaller than the variance of the crude simulation technique; this allows to estimate γ to a higher precision with the same number of replications. The technique which a priori seems the more appropriate to this kind of system with failure propagations is DSFB; the experimental results confirm this, as DSFB not only has the best variance, but also the second best execution time among the importance sampling techniques compared, and only twice the execution time of the crude technique. The best execution time among the importance sampling techniques corresponds to IFB, which has almost the same speed as the crude method; unfortunately, it is also the least precise method, with a variance very similar to the crude Monte Carlo one. This clearly shows the limitations of IFB when applied to complex systems like the one being evaluated here; in particular, because there are failure propagation, many of the most important failure paths (those with higher probabilities) will include repair transitions, and will be under-represented by the IFB biasing strategy.

The data numerical values in this example have been chosen such that with the relatively small number M of iterations, even the crude method allows to obtain a confidence interval. At the same time, this allows to underline the importance of the concept of *efficiency*: even FB is more efficient than the crude technique, since we must take into account both the execution time and the obtained precision. In the Table, we put in the last column the relative efficiency of each method with respect to the crude one, that is, the product of the variance and the execution time of the crude technique divided by the corresponding product of the specific algorithm.

Method	95% confidence interval for γ	Variance	Time	Rel. eff.
crude	$[2.257 \times 10^{-4}, 4.542 \times 10^{-4}]$	3.399×10^{-9}	2' 39"	1
FB	$[2.448 \times 10^{-4}, 2.704 \times 10^{-4}]$	4.247×10^{-11}	19' 55"	10.65
IFB	$[2.340 \times 10^{-4}, 4.659 \times 10^{-4}]$	3.499×10^{-9}	2' 41"	0.96
SFB	$[2.577 \times 10^{-4}, 2.641 \times 10^{-4}]$	2.607×10^{-12}	12' 5"	285.94
DSFB	$[2.601 \times 10^{-4}, 2.644 \times 10^{-4}]$	1.189×10^{-12}	5' 38"	1344.78

Table 5: Crude, FB, IFB, SFB and DSFB methods for a very large system

9 Some numerical aspects of rarity

In this section, we discuss a problem which, to our knowledge, has not been dealt with in the literature about simulation of highly reliable Markovian systems: for numerical reasons, non-BRE behavior in practice (when implementing these methods on a computer) can exist but not be detected when the number M of trials is fixed and $\varepsilon \to 0$. Thus, while looking stable, confidence intervals can give wrong estimates and wrong confidence levels if the failure biasing scheme is not properly chosen.

Let us consider a very simple system with 2 components, one of class 1, one of class 2. The state space is $S = \{0, 1, 2, 3\}$ where 0 is the state with both components up, in state 1 the component of class 1 is down and the other one is up, 2 represents the opposite situation and in state 3, both components are down. We do not make any particular assumption about the repairs. Assume that the probability transitions verify $P(0, 1) \approx \varepsilon$, $P(0, 2) \approx 1$, $P(1, 3) \approx \varepsilon$

and $P(2,3) \approx \varepsilon^2$.

There are 4 cycles: $c_1 = (0,1,3)$, $c_2 = (0,2,3)$, $c_3 = (0,1,0)$ and $c_4 = (0,2,0)$. The set of cycles through D is $\mathcal{D} = \{c_1, c_2\}$. The respective probabilities are $q(c_1) \approx \varepsilon^2$, $q(c_2) \approx \varepsilon^2$, $q(c_3) \approx \varepsilon$ and $q(c_4) \approx 1$. Consider the FB basic technique. After the measure change, the new probability transitions are: $P'(0,1) \approx \varepsilon,\ P'(0,2) \approx 1$, $P'(1,3) = \alpha$ and $P'(2,3) = \alpha$, leading to the new cycle probabilities $q'(c_1) \approx \alpha\varepsilon$, $q'(c_2) \approx \alpha$, $q'(c_3) \approx (1-\alpha)\varepsilon$ and $q'(c_4) \approx 1-\alpha$.

Computing γ gives $\gamma \approx 2\varepsilon^2$ and $\sigma^2 \approx \varepsilon^3/\alpha$. This implies that we do not have bounded relative error since

$$\frac{\sigma}{\gamma} \approx \frac{1}{\sqrt{2\alpha}}\varepsilon^{-\frac{1}{2}}. \tag{27}$$

Since the analysis is done for a fixed number M of trials, the equivalents of the $q'(c_i)$'s show that for ε small enough, it will be very unlikely that cycles c_1 and c_3 are sampled under q'. We will obtain approximatively $M\alpha$ samples of cycle c_2 and $M(1-\alpha)$ samples of cycle c_4. Denoting by L_m the "weighted" likelihood of the m^{th} replication, that is,

$$L_m = \frac{q(C_m)}{q'(C_m)} 1_{(\tau_D - \tau_0)}(C_m), \tag{28}$$

the estimation of the relative error is

$$\begin{aligned} \widehat{RE} &= z_\delta \sqrt{\frac{M}{M-1}\sum_{m=1}^{M} \frac{L_m^2}{\left(\sum_{i=1}^{M} L_i\right)^2} - \frac{1}{M-1}} \\ &\approx z_\delta \sqrt{\frac{M}{M-1}\alpha M \frac{\left(\varepsilon^2/\alpha\right)^2}{\left(\alpha M \varepsilon^2/\alpha\right)^2} - \frac{1}{M-1}} \\ &= z_\delta \sqrt{\frac{1}{(M-1)\alpha} - \frac{1}{M-1}}. \end{aligned}$$

In other words, the observed relative error $\widehat{RE}$ is bounded while the theoretical RE is not. This is, again, a negative consequence of the rare event situation. This is also a supplementary reason to only use balanced importance sampling methods, for which this problem does not exist.

10 Conclusion

We discussed the importance sampling methods designed to estimate the MTTF of a complex system modeled by a Markov chain, in the rare events context. We reviewed the best among already existing methods, and we proposed some modifications of the existing and some new ones, behaving better in some well identified situations.

We also analyzed the main properties of the considered techniques: the bounded relative error concept, the bounded normal approximation concept, their relationships and the relationships with the balanced versions of the estimation algorithms. In particular, we give a case where one of these methods, the IFB one, does not have the BRE property, in contradiction with what is stated in [13].

We compared numerically all the methods, to show some examples of their expected behavior for different classes of systems.

Finally, we showed that apparent numerical robustness when the rarity parameter goes to zero might be wrong if not using appropriate sampling schemes.

The discussion should be helpful in (i) choosing among the available techniques and (ii) in designing new variance reduction algorithms for the same or for other dependability measures.

References

[1] V. Bentkus and F. Götze. The Berry-Esseen bound for Student's statistic. *The Annals of Probability*, 24(1):491–503, 1996.

[2] H. Cancela. *Évaluation de la sûreté de fonctionnement : modèles combinatoires et Markoviens*. PhD thesis, Université de Rennes 1, December 1996.

[3] J. A. Carrasco. Failure distance based on simulation of repairable fault tolerant systems. In *Proceedings of the 5th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, pages 351–365, 1991.

[4] A.E. Conway and A. Goyal. Monte Carlo simulation of computer system availability/reliability models. In *Proceedings of the Seventeenth Symposium on Fault-Tolerant Computing*, pages 230–235, July 1987.

[5] A. Goyal, P. Shahabuddin, P. Heidelberger, V. F. Nicola, and P. W. Glynn. A unified framework for simulating Markovian models of highly dependable systems. *IEEE Transactions on Computers*, 41(1):36–51, January 1992.

[6] P. Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulations*, 54(1):43–85, January 1995.

[7] J. Katzmann. System architecture for non-stop computing. In *14th IEEE Comp. Soc. International Conf*, pages 77–80, 1977.

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

[9] C. Liceaga and D. Siewiorek. Automatic specification of reliability models for fault tolerant computers. *NASA technical paper 3301*, July 1993.

[10] S. Mahévas and G. Rubino. Bound computation of dependability and performance measures. *IEEE Transactions on Computers*, to appear.

[11] R.R. Muntz, E. de Souza e Silva, and A. Goyal. Bounding availability of repairable computer systems. *IEEE Transactions on Computers*, 38(12):1714–1723, 1989.

[12] M. K. Nakayama. General Conditions for Bounded Relative Error in Simulations of Highly Reliable Markovian Systems. *Advances in Applied Probability*, 28:687–727, 1996.

[13] C. Papadopoulos. A New Technique for MTTF Estimation in Highly Reliable Markovian Systems. *Monte Carlo Methods and Applications*, 4(2):95–112, 1998.

[14] P. Shahabuddin. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science*, 40(3):333–352, March 1994.

[15] B. Tuffin. *Simulation accélérée par les méthodes de Monte Carlo et quasi-Monte Carlo : théorie et applications*. PhD thesis, Université de Rennes 1, October 1997.

[16] B. Tuffin. Bounded Normal Approximation in Highly Reliable Markovian Systems. *Journal of Applied Probability*, 36(4):974–986, 1999.