Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

HIGHLY RELIABLE MARKOVIAN SYSTEMS INTERVAL AVAILABILITY ESTIMATION BY IMPORTANCE SAMPLING

Bruno Tuffin

Inria
Campus Universitaire de Beaulieu
35042 Rennes Cedex, FRANCE

ABSTRACT

This paper describes how importance sampling can be applied to efficiently estimate the average interval availability of highly reliable Markovian systems, made of components subject to failures and repairs. We describe a methodology for approximating the zero-variance change of measure. The method is illustrated to be very efficient on a small example, compared with standard importance sampling strategies developed in the literature.

1 INTRODUCTION

Multi-component repairable systems are heterogeneous systems involving various types of elements subject to repairs and failures. Analyzing models representing those systems is of importance in areas such as transport, telecommunications, computer systems or space research. When designing a fault-tolerant system, it is important to select the components (in case of computers, this could be disks, processors...) and the repair policies rendering it highly dependable at a minimal cost.

Those systems are often modeled by Markov chains (remark also that the case of non-Markovian systems has been handled for instance by Nicola et al. (1993), Nicola et al. (1993)), but the size of the model is usually out of reach for analytic or numerical-analysis-based methods, hence the use of simulation. As a second issue, system failures are usually very rare, so that a standard (or crude) simulation of the system requires extremely large sample sizes to get an accurate estimation of the measure of interest. The (Markovian) models involving rare failures are commonly called highly reliable Markovian systems (HRMS). Several variance reduction techniques have been developed to cope with the general problem of rare event simulation (Rubino and Tuffin 2009, Blanchet and Lam 2011). This paper will focus on the so-called *importance sampling* (IS) technique (Glynn and Iglehart 1989). The main idea of IS is to change the probability distributions driving the system, introducing a bias that is “counter-balanced” by a weight in the estimator, called the *likelihood ratio*, in order to recover an unbiased estimator. The choice of the IS distributions is the key for producing an important variance reduction.

The literature on HRMS performance analysis (mainly by IS) has been quite extensive. Several types of measures have been studied. We can mention among the main ones the mean time to failure (MTTF) (Shahabuddin et al. 1988) or any steady-state performance measures (Goyal, Heidelberger, and Shahabuddin 1987, Shahabuddin 1994b, Shahabuddin 1994a, Nakayama 1996, Cancela, Rubino, and Tuffin 2002) thanks to regenerative simulation combined with IS. Transient measures have been studied by Shahabuddin (1994a), Nakayama and Shahabuddin (2004), with a particular attention paid to the system unreliability (the probability that it fails before a time horizon) or the interval availability. A known performing method by Lewis and Böhm (1984), Shahabuddin (1994a) is called *failure biasing plus forcing*. The principle is to accelerate failures with respect to repairs (the failure biasing part) and to force the first component-failure transition before the time horizon (the forcing part).

978-1-4799-7486-3/14/$31.00 ©2014 IEEE

Here we focus on a specific transient metric: the average interval availability, of interest in many different contexts, especially when there is a predefined mission time. It can be simulated using failure biasing plus forcing. Indeed, forcing the first transition plus failure biasing was shown to produce bounded relative error (Shahabuddin 1994a), i.e., to ensure that the sample size required to get a specified relative error will not increase as failures go rarer, when t is small with respect to the MTTF. But our main contribution will be the design and application of a (specific) approximation of the zero-variance IS (or ZVA for zero-variance approximation), i.e., an approximation of the change of probability distribution that always produces the exact value (L'Ecuyer and Tuffin 2008). ZVA has been successfully applied to HRMS for estimating steady-state measures (L'Ecuyer and Tuffin 2011), and to estimate the probability that the Markov chain hits a given set of states before a time limit (De Boer et al. 2007). We borrow the general model from De Boer et al. (2007), but ZVA needs a new development because applied to a different metric. Actually the average interval availability make it a bit more difficult to implement ZVA than in De Boer et al. (2007). For illustration purposes, we will make use of a very simple (the simplest) running example made of two states. To highlight the potential of our new method we will also implement on this example the failure biasing plus forcing and an IS that just changes the failure rates (i.e., the rates of rare events) by making them occurring more often; it will highlight the gains obtained from our method.

The paper is organized as follows. Section 2 describes the model, mainly based on De Boer et al. (2007) and the metric we wish to compute. Section 3 describes IS in a general way applied to the model, as well as existing algorithms. Section 4 is devoted to the ZVA procedure ad describes the algorithm. Its efficiency is illustrated on the small example and compared with previous methods. Finally, Section 5 concludes and provides the next steps we wish to develop.

2 MODEL

Consider a stochastic process $(X_t)_{t\geq 0}$ (which is Markov chain when using exponential distributions) such that X_t is the state of the system at time t, with finite state space $\mathscr{X}$, partitioned into $\mathscr{U}$, set of up states, and $\mathscr{D}$, set of down states. For example, a state $x \in \mathscr{X}$ can be a multi-dimensional state $x = (x_1, \ldots, x_c)$ with $0 \leq x_i \leq n_i$ non-negative integer representing the number of operational type-i components when we have c types of components (n_i is then the total number of components of this type). The system is in an operational state $x \in \mathscr{U}$ if for some combinations of a sufficient number of components of each type. The state U denotes the state where every single component is operational.

Denote by $\mathbb{E}_x[\cdot]$ the expectation when starting from state x and let $u(x,t)$ be the (total) unavailability over an interval $[0,t]$ when starting from x, i.e.,

$$u(x,t) = \mathbb{E}_x\left[\int_0^t 1_{\mathscr{D}}(X_v)dv\right].$$

Our aim is to compute $u(U,T)$ for some T representing the *mission time*. We add for convenience the convention $u(x,t) = 0$ if $t \leq 0$.

Following the lines developed by De Boer et al. (2007), define the process $(X_j, T_j)_{j\geq 0}$ with

- $(X_j)_j$ the state of $(X_t)_{t\geq 0}$ right after of j-th event (change of state), and
- T_j the time *remaining* before the end of simulation when entering in X_j.

Let τ be the last change of state before the simulation ends and set the convention $T_{\tau+1} = 0$ (and $T_0 = T$).

Define in general the probability density that, from (x,t) the next state is (y,t') (with $t' < t$), as $\pi(y,t'|x,t)$. Recall that for a Markov chain with transition rates $\lambda_{x,y}$ defined for $x,y \in \mathscr{X}$, and with $\lambda_x = \sum_{y\in\mathscr{X}} \lambda_{x,y}$, we have

$$\pi(y,t'|x,t) = \lambda_{x,y}e^{-\lambda_x(t-t')}.$$

Example 1 In the whole paper, for illustration purposes, we will make use of the simplest possible example, made of a single item with a time to failure following an exponential distribution with rate λ. Upon failure, the item is repaired. The time to restoration also follows an exponential distribution, with rate μ. We thus have only two states U (the up state) and D (there down state). The operational life of the item is $T = 25$. Typical values for the parameters in our numerical experiments will be $\lambda = 10^{-6}$ and $\mu = 2$. If we directly simulate this system, we have a time interval availability $u(U,25) \approx$1.2E-5 and a variance (for a single run) of approximately 1.2E-5 too.

3 GENERAL APPLICATION OF IMPORTANCE SAMPLING

The *likelihood* of a sample path $((X_0,T_0),\ldots,(X_\tau,T_\tau))$ is

$$\left(\prod_{i=0}^{\tau-1} \pi(X_{i+1},T_{i+1}|X_i,T_i)\right) \bar{\Pi}_{X_\tau}(T_\tau),$$

the last term being the probability that no transition occurs while T_τ is remaining, $\bar{\Pi}_{X_\tau}(t) = \sum_y \int_{-\infty}^{0} \pi(y,t'|x,t)dt'$.

The general idea of IS is to change for all transition from $x \in \mathscr{X}$ to $y \in \mathscr{X}$ the (conditional) density $\pi(y,t'|x,t)$ by another density $g(y,t'|x,t)$. We define analogously to $\bar{\Pi}_{X_\tau}(t)$ the function $\bar{G}_{X_\tau}(t) = \sum_y \int_\infty^t g(y,t'|x,t)dt$. The *likelihood ratio* of the path is defined as

$$L = \left(\prod_{i=0}^{\tau-1} \frac{\pi(X_{i+1},T_{i+1}|X_i,T_i)}{g(X_{i+1},T_{i+1}|X_i,T_i)}\right) \frac{\bar{\Pi}_{X_\tau}(T_\tau)}{\bar{G}_{X_\tau}(T_\tau)}.$$

An unbiased IS estimator (if $g(y,t'|x,t) > 0$ as soon as $\pi(y,t'|x,t) > 0$) is then

$$U_{IS}^{(L)} = \left(\sum_{i=0}^{\tau} (T_i - T_{i+1}) 1_{\mathscr{D}}(X_i)\right) L,$$

provided that the expected number of steps τ is finite under the IS distribution (Glynn and Iglehart 1989).

Another (unbiased) way to proceed with importance sampling is to rather multiply at each change of state the cumulated time spent down by the likelihood ratio up to that change of state. In other words, the estimator is

$$U_{IS} = \left(\sum_{i=0}^{\tau} (T_i - T_{i+1}) 1_{\mathscr{D}}(X_i) L_i\right) \tag{1}$$

with

$$L_i = \left(\prod_{j=0}^{i} \frac{\pi(X_{j+1},T_{j+1}|X_j,T_j)}{g(X_{j+1},T_{j+1}|X_j,T_j)}\right) \quad \text{for } i < \tau$$

and

$$L_i = \left(\prod_{j=0}^{\tau-1} \frac{\pi(X_{j+1},T_{j+1}|X_j,T_j)}{g(X_{j+1},T_{j+1}|X_j,T_j)}\right) \frac{\bar{\Pi}_{X_\tau}(T_\tau)}{\bar{G}_{X_\tau}(T_\tau)} \quad \text{for } i = \tau.$$

Using this estimator reduces the variance with respect to the above IS estimator (since "avoiding" at each step the useless noise due to the likelihood of the rest of the path), see Asmussen and Glynn (2007).

Example 2 Coming back to Example 1, we can look at the gains in terms of variance with respect to standard Monte Carlo simulation for $U_{IS}^{(L)}$ and U_{IS} when simply replacing the failure rate λ of by another rate λ'. Figure 1 displays this evolution of the variance of $U_{IS}^{(L)}$ an U_{IS} in terms of λ'. Recall that without IS, the variance the estimator is approximately 1.2E-5. Estimator $U_{IS}^{(L)}$ with $\lambda' = 10^{-4}$ yields a variance

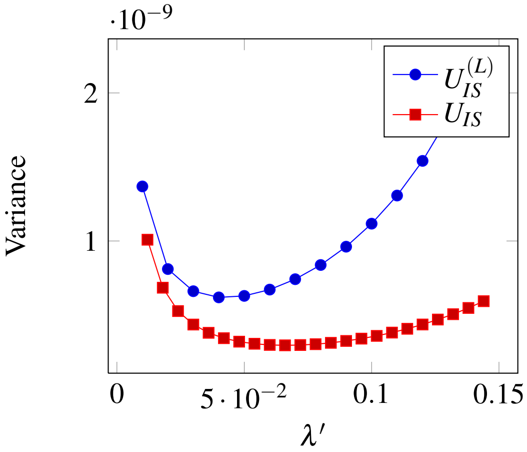

Figure 1: Evolution of the variances of $U_{IS}^{(L)}$ and U_{IS} when chaining the failure rate to λ', on Example 2.

of 1.2E-7. For 0.04, we get a variance of 6.2E-10, hence a huge reduction factor. Remark though that the simulation is a bit longer than with standard Monte Carlo since more events are generated. As expected, the variance of U_{IS} is smaller than that of $U_{IS}^{(L)}$. With $\lambda' = 0.04$ yields a variance of 3.5E-10, but it is even 2.95E-10 for $\lambda' = 0.066$. Displaying the results obtained with this IS scheme (as well as failure biasing plus forcing in the next example) will help to illustrate the power of the zero-variance approximation that will be defined in Section 4.

An IS method known to be robust to rare failures is failure biasing + forcing (Lewis and Böhm 1984, Shahabuddin 1994a). The basic idea of failure biasing when in state x is to force the next event to be a failure with a fixed probability p (a repair being highly likely otherwise), and the sojourn time in the state is kept the same (an exponential random variable with rate the sum of rates of transitions). One issue with failure biasing when in a state where only failures can happen (mainly the fully operational state) is that the time to first component failure may be long. Forcing ensures this initial failure to occur by sampling it conditionally on the fact that it is less than T (i.e., it becomes a truncated exponential distribution). Two cases can be considered: forcing is applied only to the first failure, or each time we are in (or come back to) a state where only failures can happen.

Example 3 Still on our toy Example 1, failure biasing has here no effect since only one transition is possible in any state. Forcing (meaning using a truncated exponential distribution between 0 and the remaining mission time) when applied each time the chain is in state 1 or just the first failure both yield a variance of approximately 1.5E-10 (with estimator U_{IS}). It is about half the optimal variance obtained when changing the failure rate, with the advantage of not having to perform an optimization on the choice of failure rate. The reason why forcing each time we are in state U or just the first one has a low impact on the variance is because paths with several failures have a low contribution to the estimator and to the second moment.

4 ZERO-VARIANCE IS APPROXIMATION

4.1 Zero-variance change of measure

Let us now define a specific density for each transition from any state $x \in \mathscr{X}$ to any other state $y \in \mathscr{X}$, jumping from a remaining mission time t to $t' < t$, as

$$g^*(y,t'|x,t) = \frac{u(y,t') + 1(x \text{ failed}) \min(t-t',t)}{u(x,t)} \pi(y,t'|x,t). \tag{2}$$

Remark that the transitions are not exponential anymore, not even truncated exponentials.

The following proposition states that when using those densities, the resulting estimator is "perfect".

Proposition 4 The change of measure using densities defined in (2) leads to an estimator U_{IS} defined in (1) with variance zero, i.e., an estimator always producing the exact solution $u(U,T)$.

Proof. The estimator is

$$\begin{aligned}
U_{IS} &= \left(\sum_{i=0}^{\tau}(T_i - T_{i+1})1_{\mathscr{D}}(X_i)L_i\right)\\
&= \sum_{i=0}^{\tau-1}(T_i - T_{i+1})1_{\mathscr{D}}(X_i)\left(\prod_{j=0}^{i}\frac{u(X_j,T_j)}{u(X_{j+1},T_{j+1}) + 1_{\mathscr{D}}(X_j)(T_j - T_{j+1})}\right)\\
&+ (T_\tau - T_{\tau+1})1_{\mathscr{D}}(X_\tau)\left(\prod_{j=0}^{\tau-1}\frac{u(X_j,T_j)}{u(X_{j+1},T_{j+1}) + 1_{\mathscr{D}}(X_j)(T_j - T_{j+1})}\right)\frac{\bar{\Pi}_{X_\tau}(T_\tau)}{\bar{G}_{X_\tau}(T_\tau)}.
\end{aligned}$$

We first note that

$$\begin{aligned}
\bar{G}_{X_\tau}(T_\tau) &= \int_{-\infty}^{0}\sum_y g^*(y,t'|X_\tau,T_\tau)dt' = \int_{-\infty}^{0}\sum_y \frac{u(y,t') + 1_{\mathscr{D}}(X_\tau)\min(T_\tau - t', T_\tau)}{u(X_\tau,T_\tau)}\pi(y,t'|X_\tau,T_\tau)dt'\\
&= 1_{\mathscr{D}}(X_\tau)\frac{T_\tau}{u(X_\tau,T_\tau)}\int_{-\infty}^{0}\sum_y \pi(y,t'|X_\tau,T_\tau)dt'\\
&= 1_{\mathscr{D}}(X_\tau)\frac{T_\tau}{u(X_\tau,T_\tau)}\bar{\Pi}_{X_\tau}(T_\tau).
\end{aligned}$$

In other words, a state $x \notin \mathscr{D}$ will not be the last state of a sample path. For $X_\tau \in \mathscr{D}$,

$$\bar{G}_{X_\tau}(T_\tau) = \frac{T_\tau}{u(X_\tau,T_\tau)}\bar{\Pi}_{X_\tau}(T_\tau),$$

thus

$$\frac{\bar{\Pi}_{X_\tau}(T_\tau)}{\bar{G}_{X_\tau}(T_\tau)} = \frac{u(X_\tau,T_\tau)}{T_\tau} = \frac{u(X_\tau,T_\tau)}{u(X_{\tau+1},T_{\tau+1}) + 1_{\mathscr{D}}(X_\tau)(T_\tau - T_{\tau+1})}.$$

This gives

$$U_{IS} = \sum_{i=0}^{\tau}(T_i - T_{i+1})1_{\mathscr{D}}(X_i)\left(\prod_{j=0}^{i}\frac{u(X_j,T_j)}{u(X_{j+1},T_{j+1}) + 1_{\mathscr{D}}(X_j)(T_j - T_{j+1})}\right).$$

Next, by recursion on the value of τ, $U_{IS} = u(X_0,T)$. This shows the proposition. □

Example 5 Still on the two-states Example 1, the densities become

$$\begin{aligned}
g^*(D,t'|U,t) &= \frac{u(D,t')}{u(U,t)}\lambda e^{-\lambda(t-t')}\\
g^*(U,t'|D,t) &= \frac{u(U,t') + \min(t-t',t)}{u(D,t)}\mu e^{-\mu(t-t')}.
\end{aligned}$$

4.2 Zero-variance approximation

Unfortunately, the zero-variance change of measure (2) requires the knowledge of $u(\cdot,\cdot)$, i.e., what we are trying to compute. It is therefore not applicable directly because if we knew $u(\cdot,\cdot)$, there would be no interest to use simulation. However, the expression (2) provides a good idea of the type of densities we should try to approach.

What we propose here is to make use of (2), by just replacing the unknown $u(\cdot,\cdot)$ by a rough approximation $\hat{u}(\cdot,\cdot)$ easy to compute.

Notice that since we use an approximation of $u(\cdot,\cdot)$, the integral (over the next remaining time t') of $\sum_y g(y,t'|x,t) = \sum_y \frac{\hat{u}(y,t')+1(x \text{ failed})\min(t-t',t)}{\hat{u}(x,t)}\pi(y,t'|x,t)$ may not be 1, hence we need to introduce a constant of normalization $c(x,t)$ and use the density

$$g(y,t'|x,t) = c(x,t)(\hat{u}(y,t') + 1(x \text{ failed})\min(t-t',t))\pi(y,t'|x,t). \qquad (3)$$

What type of approximation do we suggest? Since failures are rare with respect to repairs, when a state in $\mathscr{D}$ is reached, and from it $\mathscr{U}$ is reached again, it is unlikely to come back to $\mathscr{D}$ a second time, and those additional trips to $\mathscr{D}$ can be neglected. This corresponds to a first order approximation (a single visit to failure) of $u(\cdot,\cdot)$ (L'Ecuyer and Tuffin 2011). In other words, a simple principle to approach $u(x,t)$ by:

1. $\forall x \in \mathscr{D}$, $\hat{u}(x,t) = \min(t, \mathbb{E}[S_x])$, with S_x the sojourn time in $\mathscr{D}$ before leaving it for the first time when starting in $x \in \mathscr{D}$. S_x can also be simplified as the duration of the most probable path leaving $\mathscr{D}$, so that $\mathbb{E}[S_x]$ is the expectation of this path, therefore the expected value of a hypo-exponential distribution: $\sum_{i=1}^{k_{\mathscr{U}}(x)} 1/\lambda_{x'_{i-1},x'_i}$ if this path is $(x'_0 = x, x'_1, \dots, x'_{k_{\mathscr{U}}(x)})$, made of $k_{\mathscr{U}}(x)$ steps.
2. $\forall x \in \mathscr{U}$,
$$\hat{u}(x,t) = \int_0^t \min(t-v, \mathbb{E}[S_x]) f_x(v)dv \qquad (4)$$
where we consider the most likely path from x to $\mathscr{D}$, with hypo-exponential density of duration f_x (v is the time to reach $\mathscr{D}$), and neglect the possibility of returning to $\mathscr{D}$ if leaving it, by taking $\hat{u}()$ as the time $\min(t-v, \mathbb{E}[S_x])$ spent in $\mathscr{D}$ during the first visit. S_x is here the sojourn time in $\mathscr{D}$ during a single visit to $\mathscr{D}$ when starting from x (but in $\mathscr{U}$); this can again be simplified using the most likely path from x to $\mathscr{D}$, with entrance state $x_{\mathscr{D}}$, and S_X is the time to leave $\mathscr{D}$ from $x_{\mathscr{D}}$. This again requires the identification of most likely paths on the embedded discrete time Markov chain (again, see (L'Ecuyer and Tuffin 2011)).

We now describe the computation of the IS densities and the generation of random variables, but the reader may go first to Section 4.3 where the simpler application to the two-states chain is detailed.

4.2.1 Computation of $\hat{u}(x,t)$

Considering that the most likely path to $\mathscr{D}$ $(x_0 = x, x_1, \dots, x_{k_{\mathscr{D}}(x)})$ is made of $k_{\mathscr{D}}(x)$ failures only, meaning that the rates λ_{x_{i-1},x_i} are all close to 0, an asymptotic development of the hypo-exponential distribution gives

$$f_x(v) = \prod_{i=1}^{k_{\mathscr{D}}(x)} \lambda_{x_{i-1},x_i} \frac{v^{k_{\mathscr{D}}(x)-1}}{(k_{\mathscr{D}}(x)-1)!} + o((\max_i \lambda_{x_{i-1},x_i})^{k_{\mathscr{D}}(x)}).$$

Replacing f_x by this first order approximation in (4), the computations can be simplified. It gives:

- $\forall x \in \mathscr{D}$, $\hat{u}(x,t) = \min(t, \mathbb{E}[S_x])$ where $\mathbb{E}[S_x] \approx \sum_{i=1}^{k_{\mathscr{U}}(x)} 1/\lambda_{x'_{i-1},x'_i}$ (highlighting the dependence of the most likely path $(x'_0 = x, x'_1, \dots, x'_{k_{\mathscr{U}}(x)})$ to $\mathscr{U}$ from x),
- $\forall x \in \mathscr{U}$, we need to integrate $\hat{u}(x,t) = \left(\prod_{i=1}^{k_{\mathscr{D}}(x)} \lambda_{x_{i-1},x_i}\right) \int_0^t \min(t-v, \mathbb{E}[S_x]) \frac{v^{k_{\mathscr{D}}(x)-1}}{(k_{\mathscr{D}}(x)-1)!} dv.$

- If $t \leq \mathbb{E}[S_x]$,

$$\hat{u}(x,t) = \left(\prod_{i=1}^{k_{\mathscr{D}}(x)} \lambda_{x_{i-1},x_i}\right) \frac{t^{k_{\mathscr{D}}(x)+1}}{(k_{\mathscr{D}}(x)+1)!}.$$

- If $t > \mathbb{E}[S_x]$,

$$\begin{aligned}
\hat{u}(x,t) &= \frac{\left(\prod_{i=1}^{k_{\mathscr{D}}(x)} \lambda_{x_{i-1},x_i}\right)}{(k_{\mathscr{D}}(x)+1)!} \left[\mathbb{E}[S_x] \int_0^{t-\mathbb{E}[S_x]} v^{k_{\mathscr{D}}(x)-1} dv + \int_{t-\mathbb{E}[S_x]}^{t} (t-v) v^{k_{\mathscr{D}}(x)-1} dv\right] \\
&= \frac{\left(\prod_{i=1}^{k_{\mathscr{D}}(x)} \lambda_{x_{i-1},x_i}\right)}{(k_{\mathscr{D}}(x)-1)!} \left[\mathbb{E}[S_x] \frac{(t-\mathbb{E}[S_x])^{k_{\mathscr{D}}(x)}}{k_{\mathscr{D}}(x)} + \frac{t^{k_{\mathscr{D}}(x)+1}}{k_{\mathscr{D}}(x)(k_{\mathscr{D}}(x)+1)} \right. \\
&\qquad \left. - \frac{t(t-\mathbb{E}[S_x])^{k_{\mathscr{D}}(x)}}{k_{\mathscr{D}}(x)} + \frac{(t-\mathbb{E}[S_x])^{k_{\mathscr{D}}(x)+1}}{k_{\mathscr{D}}(x)+1}\right].
\end{aligned}$$

4.2.2 Computation of densities and random variate generation

Using the above expressions $\hat{u}(x,t)$, the approximate zero-variance densities can be computed from (3) as:

- $\forall x \in \mathscr{D}$,
  - If $y \in \mathscr{D}$

$$\hat{g}(y,t'|x,t) \quad \approx \quad c(x,t)(\min(t', \mathbb{E}[S_y]) + \min(t-t',t))\lambda_{x,y} e^{-\lambda_x(t-t')}.$$

  - If $y \in \mathscr{U}$ (we necessarily have a repair),

$$\hat{g}(y,t'|x,t) \quad = \quad c(x,t)(\hat{u}(y,t') + \min(t-t',t))\lambda_{x,y} e^{-\lambda_x(t-t')},$$

  leading to the two cases, denoting by $(y_0 = y, y_1, \ldots, y_{k_{\mathscr{D}}(y)})$ the most likely path from y to $\mathscr{D}$:
    - ∗ If $t \leq \mathbb{E}[S_y]$,

$$\hat{g}(y,t'|x,t) \quad = \quad c(x,y,t)\left(\left(\prod_{i=1}^{k_{\mathscr{D}}(y)} \lambda_{y_{i-1},y_i}\right) \frac{(t')^{k_{\mathscr{D}}(y)+1}}{(k_{\mathscr{D}}(y)+1)!} + \min(t-t',t)\right) \lambda_{x,y} e^{-\lambda_x(t-t')}$$

    - ∗ If $t > \mathbb{E}[S_y]$,
      - · If $t' > \mathbb{E}[S_y]$

$$\begin{aligned}
\hat{g}(y,t'|x,t) \quad = \quad c(x,y,t) &\left[\frac{\left(\prod_{i=1}^{k_{\mathscr{D}}(y)} \lambda_{y_{i-1},y_i}\right)}{(k_{\mathscr{D}}(y)-1)!} \left[\mathbb{E}[S_y] \frac{(t'-\mathbb{E}[S_y])^{k_{\mathscr{D}}(y)}}{k_{\mathscr{D}}(y)} + \frac{(t')^{k_{\mathscr{D}}(y)+1}}{k_{\mathscr{D}}(y)(k_{\mathscr{D}}(y)+1)} \right.\right. \\
&\left.\left. - \frac{t'(t'-\mathbb{E}[S_y])^{k_{\mathscr{D}}(y)}}{k_{\mathscr{D}}(y)} + \frac{(t'-\mathbb{E}[S_y])^{k_{\mathscr{D}}(y)+1}}{k_{\mathscr{D}}(y)+1}\right] + \min(t-t',t)\right] \lambda_{x,y} e^{-\lambda_x(t-t')},
\end{aligned}$$

      - · If $t' \leq \mathbb{E}[S_y]$

$$\hat{g}(y,t'|x,t) \quad = \quad c(x,y,t) \frac{(t')^{k_{\mathscr{D}}(y)+1}}{(k_{\mathscr{D}}(y)+1)!} \lambda_{x,y} e^{-\lambda_x(t-t')}.$$

- $\forall x \in \mathscr{U}$, and a transition (x,y),

- If $y \in \mathscr{D}$,

$$\hat{g}(y,t'|x,t) \quad \approx \quad c(x,t)\min(t',\mathbb{E}[S_y])\lambda_{x,y}e^{-\lambda_x(t-t')}$$

- If $y \in \mathscr{U}$, let again the most likely path from y to $\mathscr{D}$ be $(y_0 = y, y_1, \ldots, y_{k_{\mathscr{D}}(y)})$.
  - If $t \leq \mathbb{E}[S_y]$,

$$\hat{g}(y,t'|x,t) \quad = \quad c(x,t)\left(\prod_{i=1}^{k_{\mathscr{D}}(y)} \lambda_{y_{i-1},y_i}\right)\frac{(t')^{k_{\mathscr{D}}(y)+1}}{(k_{\mathscr{D}}(y)+1)!}\lambda_{x,y}e^{-\lambda_x(t-t')}$$

  - If $t > \mathbb{E}[S_y]$,
    - If $t' > \mathbb{E}[S_y]$

$$\begin{aligned}\hat{g}(y,t'|x,t) \quad = \quad & c(x,t)\frac{\left(\prod_{i=1}^{k_{\mathscr{D}}(y)} \lambda_{y_{i-1},y_i}\right)}{(k_{\mathscr{D}}(y)-1)!}\left[\mathbb{E}[S_y]\frac{(t'-\mathbb{E}[S_y])^{k_{\mathscr{D}}(y)}}{k_{\mathscr{D}}(y)} + \frac{(t')^{k_{\mathscr{D}}(y)+1}}{k_{\mathscr{D}}(y)(k_{\mathscr{D}}(y)+1)}\right. \\ & \left. - \frac{t'(t'-\mathbb{E}[S_y])^{k_{\mathscr{D}}(y)}}{k_{\mathscr{D}}(y)} + \frac{(t'-\mathbb{E}[S_y])^{k_{\mathscr{D}}(y)+1}}{k_{\mathscr{D}}(y)+1}\right]\lambda_{x,y}e^{-\lambda_x(t-t')},\end{aligned}$$

    - If $t' \leq \mathbb{E}[S_y]$

$$\hat{g}(y,t'|x,t) \quad = \quad c(x,t)\frac{(t')^{k_{\mathscr{D}}(y)+1}}{(k_{\mathscr{D}}(y)+1)!}\lambda_{x,y}e^{-\lambda_x(t-t')}.$$

How to generate the next state and next time? We follow the same type of method as De Boer et al. (2007): we compute the normalization constant by integrating $\sum_y \hat{g}(y,t'|x,t)$ with respect to t' over $[0,t]$. If there is no repair from x, $\lambda_x \ll 1$ and the exponential term can be neglected. Otherwise, an integration by part is realized. Notice that the relative contribution of each integral of $\hat{g}(y,t'|x,t)$ gives the next state. Given this state, t' is sampled by inversion (if we have a monomial) or by acceptance-rejection (for example with the exponential distribution for the proposed t', accepting the value with probability taken as the polynomial divided by its maximum value, using an indicator function to separate the two cases for t' when we have two).

4.2.3 General algorithm

The general algorithm can be decomposed as described in Algorithm 1.

4.3 Application to the two-states chain

Coming back to our illustrative two-states example, we can rewrite the approximation of the unreliability as:

$$\begin{aligned}\hat{u}(D,t) &= \min(t, 1/\mu) \\ \hat{u}(U,t) &= \int_0^t \min(t-v, 1/\mu)\lambda e^{-\lambda v}dv.\end{aligned}$$

This last integral gives

- $\hat{u}(U,t) = t - (1-e^{-\lambda t})/\lambda$ if $t < 1/\mu$ and
- $\hat{u}(U,t) = 1/\mu + e^{-\lambda t}(1-e^{\lambda/\mu})/\lambda$ otherwise.

Algorithm 1 Algorithm to simulate a single run

```
$t := T$; $L := 1$; $Cumul := 0$;
Select an initial state $x_0$;
$x := x_0$
repeat
    Compute the normalization constant $c(x,t)$;
    Generate the next state $y$ and value $t'$;
    Update the likelihood ratio $L$;
    If $t' < 0$ then $t' = 0$;
    If $x \in \mathscr{D}$, $Cumul += L \times (t - t')$;
    $x := y$; $t := t'$;
until $t' = 0$;
return $Cumul$.
```

We thus get (for $t' \in [0,t]$) the densities (using simplifications given that $\lambda \ll 1$):

- $g(D,t'|U,t) = c(U,t)\min(t',1/\mu)\lambda e^{-\lambda(t-t')}$. This density can be generated by inverting the cumulative distribution function.
  - If $t < 1/\mu$, $g(D,t'|U,t) = c(U,t)t'\lambda e^{\lambda t'}$ with $1/c(U,t) = te^{-\lambda t} + (1-e^{-\lambda t})/\lambda$. To generate the next *remaining* time t', the cumulative distribution function is $c(U,t)(t'e^{\lambda t'} + (1-e^{\lambda t'})/\lambda)$. But we can use a further approximation using the fact that λ is close to 0. From the Taylor expansion in terms of λ, we get approximately $t'/(2t^2)$ as the density over $[0,t]$.
  - If $t \geq 1/\mu$, $g(D,t'|U,t) = c(U,t)t'\lambda e^{-\lambda(t-t')}$ if $0 \leq t' < 1/\mu$ and $g(D,t'|U,t) = c(U,t)(\lambda/\mu)e^{-\lambda(t-t')}$ if $t' \in [1/\mu,t]$, with
  $$1/c(U,t) = \frac{e^{-\lambda t}\mu - e^{-\lambda(\mu t-1)/\mu}\mu + \lambda}{\lambda\mu}.$$

    We can again use the additional approximation due to λ close to 0, leading to
  $$\hat{g}(D,t'|U,t) = \begin{cases} t'/(t/\mu - 1/(2\mu^2)) \text{ if } t' < 1/\mu \\ (1/\mu)\cdot 1/(t/\mu - 1/(2\mu^2)) \text{ if } t' \geq 1/\mu. \end{cases}$$

    This density (in terms of t') is easy to sample by inversion in all the mentioned cases.
- From $x = D$,
  $$\begin{aligned} g(U,t'|D,t) &= c(D,t)(t - (1-e^{-\lambda t})/\lambda + \min(t-t',t))\mu e^{-\mu(t-t')} \quad \text{if } t < 1/\mu \\ g(U,t'|D,t) &= c(D,t)(1/\mu + e^{-\lambda t}(1-e^{\lambda/\mu})/\lambda + \min(t-t',t))\mu e^{-\mu(t-t')} \quad \text{otherwise.} \end{aligned}$$

  Again, the expressions can be simplified by using the Taylor expansion in terns of λ, giving, if $t < 1/\mu$ first, a density (for $t - t' \geq 0$, i.e., $t' \in (-\infty, t]$)
  $$\hat{g}(U,t'|D,t) = \min(t-t',t)e^{-\mu(t-t')}\frac{\mu^2}{1-e^{-\mu t}}$$

  and actually the same asymptotic density when $t \geq 1/\mu$. This density (still in terms of t') can be sampled using for example the acceptance-rejection method with an exponential distribution as the proposal for $t-t'$ and accepting it with probability $\min(t-t',t)/t$. Remark also that $\bar{G}_D(t) = \frac{\mu t e^{-\mu t}}{1-e^{-\mu t}}$.

Example 6 If we use this estimator with the data of Example 1, we obtain an empirical variance of 3.1E-13. This is a variance 10^3 times smaller than the one obtained with the best IS parameter when changing the exponential failure rate, and 500 times smaller than the one using failure biasing plus forcing. The numerical results are summarized in Table 1 with the last column representing the required sample size for standard Monte Carlo to obtain the same accuracy as the presented method.

Table 1: Summary of numerical results on the 2-states example, with a sample size $n = 10^6$.

Method	estimation	95% Confidence interval	Variance	Variance gain / SMC
Standard Monte Carlo (SMC)	1.166E-5	(5.6531E-6, 1.7658E-5)	1.2E-5	1
$U_{IS}^{(L)}$ (changing failure rate)	1.224E-5	(1.2212E-5, 1.2310E-5)	6.2E-10	19000
U_{IS} (changing failure rate)	1.225E-5	(1.2212E-5, 1.2279E-5)	2.95E-10	39000
Forcing	1.225E-5	(1.2223E-5, 1.2271E-5)	1.5E-10	77000
Zero-Variance Approximation	1.225E-5	(1.2249E-5, 1.2251E-5)	3.1E-13	3.7E+7

Remark 7 Remark that the expected number of steps τ before stopping the simulation is finite. Indeed, note first that starting from U, this number of steps is necessarily even. Moreover, if $\tilde{\mathbb{P}}$ denotes the IS probability measure, $\tilde{\mathbb{P}}[\tau > 2k] \leq (1 - \bar{G}_D(T))^k$ because at each step the probability of "failure" to reach the remaining time t is smaller than $1 - \bar{G}_D(T)$ when in D (and from U we necessarily go to D). Therefore the expected value of tau is finite because $\tilde{\mathbb{E}}[\tau] = \sum_k \tilde{\mathbb{P}}[\tau \geq k] < \infty$.

5 CONCLUSIONS AND FUTURE WORK

We have described in this paper a new importance sampling procedure to estimate the interval unavailability of a highly reliable Markovian system. This procedure approaches the zero-variance change of measure that we have characterized. We have illustrated on a simple two-states system the significant gains that can be reached thanks to this method with respect to previously known methods.

The next steps of the work will be first to prove the robustness of the method when individual failures become more rare in a spirit similar to what was done by De Boer et al. (2007) and L'Ecuyer and Tuffin (2011). Second, we want to perform extensive testing on large-scale systems.

ACKNOWLEDGMENTS

The author would like to thank Pierre Dersin and René Valenzuela from ALSTOM Transport for valuable discussions highlighting the need for such an analysis, and the anonymous reviewers for their helpful comments.

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation*. New York: Springer-Verlag.

Blanchet, J. H., and H. Lam. 2011. "Rare Event Simulation Techniques". In *Proceedings of the 2011 Winter Simulation Conference*, 146–160: IEEE Press.

Cancela, H., G. Rubino, and B. Tuffin. 2002. "MTTF Estimation by Monte Carlo Methods Using Markov Models". *Monte Carlo Methods and Applications* 8 (4): 312–341.

De Boer, P. T., P. L'Ecuyer, G. Rubino, and B. Tuffin. 2007. "Estimating the Probability of a Rare Event Over a Finite Horizon". In *Proceedings of the 2007 Winter Simulation Conference*, 403–411: IEEE Press.

Glynn, P. W., and D. L. Iglehart. 1989. "Importance Sampling for Stochastic Simulations". *Management Science* 35:1367–1392.

Goyal, A., P. Heidelberger, and P. Shahabuddin. 1987. "Measure Specific Dynamic Importance Sampling for Availability Simulations". In *Proceedings of the 1987 Winter Simulation Conference*, edited by I. press, 351–357.

L'Ecuyer, P., and B. Tuffin. 2008. "Approximate Zero-Variance Simulation". In *Proceedings of the 2008 Winter Simulation Conference*, 170–181: IEEE Press.

L'Ecuyer, P., and B. Tuffin. 2011. "Approximating Zero-Variance Importance Sampling in a Reliability Setting". *Annals of Operations Research* 189:277–297.

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

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

Nakayama, M. K., and P. Shahabuddin. 2004. "Quick Simulation Methods for Estimating the Unreliability of Regenerative Models of Large Highly Reliable Systems". *Probability in the Engineering and Information Sciences* 18:339–368.

Nicola, V. F., M. K. Nakayama, P. Heidelberger, and A. Goyal. 1993, December. "Fast Simulation of Highly Dependable Systems with General Failure and Repair Processes". *IEEE Transactions on Computers* 42 (12): 1440–1452.

Nicola, V. F., P. Shahabuddin, P. Heidelberger, and P. W. Glynn. 1993. "Fast Simulation of Steady-State Availability in Non-Markovian Highly Dependable Systems". In *Proceedings of the 23rd International Symposium on Fault-Tolerant Computing*, 38–47: IEEE Computer Society Press.

Rubino, G., and B. Tuffin. (Eds.) 2009. *Rare Event Simulation using Monte Carlo Methods*. Wiley.

Shahabuddin, P. 1994a. "Fast Transient Simulation of Markovian Models of Highly Dependable Systems". *Performance Evaluation* 20:267–286.

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

Shahabuddin, P., V. F. Nicola, P. Heidelberger, A. Goyal, and P. W. Glynn. 1988. "Variance Reduction in Mean Time to Failure Simulations". In *Proceedings of the 1988 Winter Simulation Conference*, 491–499: IEEE Press.

AUTHOR BIOGRAPHY

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