PROBABILISTIC BOUNDED RELATIVE ERROR FOR RARE EVENT SIMULATION LEARNING TECHNIQUES

Bruno Tuffin

Inria Rennes Bretagne Atlantique
Campus Universitaire de Beaulieu
35042 Rennes Cedex, FRANCE

Ad Ridder

Vrije Universiteit
De Boelelaan 1105
1081 HV Amsterdam, THE NETHERLANDS

ABSTRACT

In rare event simulation, we look for estimators such that the relative accuracy of the output is "controlled" when the rarity is getting more and more critical. Different robustness properties of estimators have been defined in the literature. However, these properties are not adapted to estimators coming from a parametric family for which the optimal parameter is random due to a learning algorithm. These estimators have random accuracy. For this reason, we motivate in this paper the need to define probabilistic robustness properties. We especially focus on the so-called probabilistic bounded relative error property. We additionally provide sufficient conditions, both in general and Markov settings, to satisfy such a property, and hope that it will foster discussions and new works in the area.

1 INTRODUCTION

Rare event simulation has been the topic of an extensive research during the past thirty years; see Heidelberger (1995), Juneja and Shahabuddin (2006), Rubino and Tuffin (2009), and the references therein. Rare events are indeed important in many fields, from failures in transportation systems or nuclear plants, to losses and bankruptcy of financial companies, as well as losses of information in telecommunication systems. Even if the event is rare, its outcome may be so catastrophic in terms of money losses or human lives that it cannot be neglected and has to be carefully studied. Computing rare event probabilities has been proved to be a difficult task because the rarity makes the event difficult or impossible to observe.To circumvent this problem, sophisticated techniques have been designed. The two main methods are: *importance sampling* (IS), which consists of changing the probability laws driving the considered model in order to increase the occurrence of the event but keeps an unbiased estimator by also changing the random variable (rv) of interest; and *splitting*, which basically favors "successful" trajectories, i.e., those that get closer to the rare event, by replicating them in a number of offspring, and proceeds successively until the event is reached.

A key issue when designing a rare event probability estimator is to determine whether or not its accuracy does not deteriorate as the probability goes to zero (that is, when the event becomes rarer). By accuracy, we mean *relative* accuracy, because what counts is the error relative to the value of the probability. There exist several definitions of robustness in the literature, more or less strict in terms of accuracy and more or less easy to satisfy. The two main definitions are the so-called *bounded relative error* property, stating that the relative error given by the standard deviation of the estimator divided by the probability of interest is kept bounded whatever the rarity of the event; and the weaker *asymptotic optimality* (or *logarithmic efficiency*) which means that the second moment and the square of the mean go to zero at the same *exponential* rate. There exist numerous other properties, for instance dealing with moments of order larger than 2, or requiring the relative error to decrease to zero with the rarity (the so-called *vanishing relative error* property). For a description and an exhaustive list of references, as well as the relations between all those properties, the reader is advised to go to Rubino and Tuffin (2009), L'Ecuyer et al. (2010).

But many rare event estimators come from adaptive techniques where parameters leading to a valid estimator are learned during a presimulation stage, such as in the *cross-entropy method* (CE). Consequently, the parameters are random. In CE the optimal parameters are derived from the minimization of the Kullback-Leibler distance between the considered parametric family and the zero-variance IS change of measure (Rubinstein 1999). After this learning phase, the rare event can be estimated thanks to a long(er) simulation. Since the parameters are random, it is hard to *guarantee* robustness property for a given choice of parameters. This randomness has to be addressed though when discussing robustness, and to our knowledge it has never been considered in the literature. The goals of this paper are therefore threefold:

- We aim first at illustrating that getting a robustness property may be itself random for adaptive methods, and that an adaptive algorithm can yield estimators with a wide diversity of results. We limit ourselves in this paper to IS.
- Our main goal is then to define *probabilistic robustness properties*, describing that a robustness property is verified with a given probability. We here focus on *probabilistic bounded relative error*, looking at the evolution of the relative error produced by the confidence interval as the event goes rarer, for a given choice of the (random) IS parameters. We also discuss briefly the property that the overall estimator should satisfy. Remark that the other existing robustness properties can easily be extended to the probabilistic case in a similar way and without complication.
- We then provide sufficient conditions under which those probabilistic robustness properties are verified.

The rest of this paper is organized as follows. Section 2 describes the basic ideas of adaptive IS, and more specifically the CE algorithm. It also illustrates why the classical robustness analysis can hardly be applied in this context. Section 3 introduces the definitions of probabilistic bounded relative error that we feel relevant for this kind of problem. Section 4 presents a general sufficient condition in order to verify this property, and the specific contexts of highly reliable Markovian systems and of the simple and toy M/M/1/b queue are described in Section 5 and Section 6.

2 RARE EVENT SIMULATION AND RELATED ROBUSTNESS ISSUE

Consider the estimation of

$$\mu = \mathbb{E}[g(X)] = \int g(x)d\mathbb{P}(x)$$

where X is a rv distributed according to probability measure $\mathbb{P}$. Parametric IS makes use of a family of measures $\{\mathbb{P}_\theta : \theta \in \Theta\}$ on a set Θ of parameters. Note that we do not assume that the default or original probability measure $\mathbb{P}$ is a member of this family. If $d\mathbb{P}_\theta(x) > 0$ when $g(x)d\mathbb{P}(x) \neq 0$, then

$$\mathbb{E}[g(X)] = \int g(x)\frac{d\mathbb{P}(x)}{d\mathbb{P}_\theta(x)}d\mathbb{P}_\theta(x) = \mathbb{E}_\theta[g(X)L(X)]$$

with $L(X) = d\mathbb{P}(x)/d\mathbb{P}_\theta(x)$ the likelihood ratio.

Let us denote by $Z(\theta) = g(X)L(X) = g(X)d\mathbb{P}/d\mathbb{P}_\theta(X)$ the single run unbiased IS estimator of μ when using parameter θ. Applying IS is of particular interest when trying to estimate probabilities of rare events A; i.e., $\mu = p = \mathbb{P}[A] \ll 1$. Thus, $g(X) = I\{A\}$ is the Bernoulli rv, whose value is 1 if A is reached and 0 otherwise. Typically we sample n independent copies of $Z(\theta)$ and use their sample mean $\widehat{Z}_n(\theta) = (1/n)\sum_{i=1}^n Z_i(\theta)$ as unbiased estimator. This estimator satisfies a δ-α accuracy if

$$\mathbb{P}(p(1-\delta) < \widehat{Z}_n(\theta) < p(1+\delta)) > 1 - \alpha.$$

Note that this is equivalent to require that the $100(1-\alpha)\%$ confidence interval has δ relative precision. Applying Chebyshev's inequality or a CLT approximation we obtain that $\mathbb{V}ar[\widehat{Z}_n(\theta)]/(p^2\delta^2) < \alpha$ suffices.

Hence, the required sample size is

$$n \approx \frac{\text{RE}^2[Z(\theta)]}{\delta^2 \alpha}, \tag{1}$$

where $\text{RE}[Z(\theta)] = \sqrt{\mathbb{V}ar[Z(\theta)]}/p$ is the *relative error* of the single run estimator. Equation (1) relates samples size and relative error. We have a similar relation for the crude Monte Carlo estimator, $Z = g(X) = I\{A\}$, which has relative error $\text{RE}[Z] = \sqrt{1-p}/\sqrt{p}$. Then equation (1) says that the required sample size is of order $(1-p)/p \sim 1/p \to \infty$ as $p \to 0$. In other words, the rarer the event, the larger the sample size required to get a confidence interval with a fixed relative accuracy.

Using IS makes sense here in order to increase the occurrence of the event and reduce the relative error (Rubino and Tuffin 2009, Chapter 2). For instance, suppose that $\text{RE}[Z(\theta)]$ remains bounded as $p \to 0$. In that case, the sample size needed to get a specified relative accuracy is bounded whatever the rarity of the event. The estimator $Z(\theta)$ will be said to verify *bounded relative error* (BRE).

But finding out an IS change of measure yielding efficient results (that is an optimal parameter θ in the set Θ), and robust to rarity, is not an easy task in general. Adaptive IS tries to learn an optimal θ, that is a θ minimizing the variance of $Z(\theta)$ under IS. While this value can be learned during the simulation (i.e., updated at each step, but then data are correlated, complicating the output analysis), we consider here the case where parameters are learned during a presimulation, potentially in a sequential way during k steps where IS parameters determined at step $j-1$ are used at step j. More precisely, the algorithm is as follows:

1. Define $\theta_0 \in \Theta$
2. Presimulation: For $j = 1$ to k
   (a) use a sample of size n_j of independent copies of X generated according to $\mathbb{P}_{\theta_{j-1}}$;
   (b) determine the value θ_j minimizing the variance of $Z(\theta_j)$.
3. Simulation:
   (a) use a sample n of independent copies of $Z(\theta_k)$ generated according to $\mathbb{P}_{\theta_k}$;
   (b) provide an estimate of μ and an associated confidence interval.

A typical adaptive technique is the so-called Cross-Entropy (CE) method where the minimization procedure is realized to determine the $\theta \in \Theta$ minimizing the Kullback-Leibler (or Cross-Entropy) distance between the zero-variance change of measure $\mathbb{P}^{(ZV)}$ and $\mathbb{P}_\theta$: $\mathscr{D}(\mathbb{P}^{(ZV)}, \mathbb{P}_\theta) = \mathbb{E}^{(ZV)}\left[\log \frac{d\mathbb{P}^{(ZV)}}{d\mathbb{P}_\theta}\right]$. When estimating $\mathbb{E}[g(X)]$, it is known that the optimal change of measure is $d\mathbb{P}^{(ZV)} = \frac{|g(X)|}{\mathbb{E}[|g(X)|]} d\mathbb{P}$ (Asmussen and Glynn 2007). This gives after straightforward simplifications

$$\mathscr{D}(\mathbb{P}^{(ZV)}, \mathbb{P}_\theta) = \mathbb{E}\left[\frac{|g(X)|}{\mathbb{E}[|g(X)|]} \log\left(\frac{|g(X)|}{\mathbb{E}[|g(X)|]} d\mathbb{P}\right)\right] - \frac{1}{\mathbb{E}[|g(X)|]} \mathbb{E}\left[|g(X)| \log d\mathbb{P}_\theta\right].$$

The minimization problem is then equivalent to solving at each step j

$$\max_\theta \mathbb{E}\left[|g(X)| \log d\mathbb{P}_\theta\right] = \max_\theta \mathbb{E}_{\theta_{j-1}}\left[\frac{d\mathbb{P}}{d\mathbb{P}_{\theta_{j-1}}} |g(X)| \log d\mathbb{P}_\theta\right] \approx \max_\theta \frac{1}{n_j} \sum_{i=1}^{n_j} |g(X_i)| \frac{d\mathbb{P}(X_i)}{d\mathbb{P}_{\theta_{j-1}}(X_i)} \log d\mathbb{P}_\theta(X_i) \tag{2}$$

with $(X_i)_i$ sequence of independent copies of r.v. X.

The next example illustrates the difficulty to ensure robustness properties when using adaptive techniques. More exactly, it shows that the algorithm, if re-run independently, can lead to learned values of θ yielding estimators experiencing large variations in their variance.

Example 1 Consider a rv X exponentially distributed with rate λ, and we want to compute $\mathbb{P}[X < \varepsilon] = \mathbb{E}[g(X)] = 1 - e^{-\lambda\varepsilon}$ with $g(x) = I\{x \in [0, \varepsilon]\}$. Suppose that we use IS and still sample from an exponential

density, but with a different rate θ, The second moment of that IS estimator is

$$\mathbb{E}_\theta[g(X)^2 L^2] = \int_0^\varepsilon \left(\frac{\lambda e^{-\lambda y}}{\theta e^{-\theta y}}\right)^2 \theta e^{-\theta y} dy = \frac{\lambda^2}{\theta(2\lambda - \theta)}(1 - e^{-(2\lambda - \theta)\varepsilon}). \tag{3}$$

The estimator satisfies bounded relative error as $\varepsilon \to 0$ if $\mathbb{E}_\theta[g(X)^2 L^2]/(\mathbb{E}[g(X)])^2$ remains bounded as $\varepsilon \to 0$. One can easily check that it is the case when $\theta = \alpha/\varepsilon$ for any $\alpha > 0$. As a consequence, the value of θ minimizing the variance, say $\theta^{(\min)}$, satisfies BRE too.

We can solve the CE-program (2) by considering the first order condition and interchanging differentiation and expectation. It is an easy exercise to show that

$$\theta^* = \arg\max_\theta \mathbb{E}\left[I\{X \leq \varepsilon\} \log d\mathbb{P}_\theta\right] = \frac{\mathbb{P}[X \leq \varepsilon]}{\mathbb{E}[XI\{X \leq \varepsilon\}]} = \frac{1 - e^{-\lambda \varepsilon}}{(1 - e^{-\lambda \varepsilon})/\lambda - \varepsilon e^{-\lambda \varepsilon}} = \frac{2}{\varepsilon} + o(1)$$

(as $\varepsilon \to 0$). The associated IS estimator with θ^* shows BRE from our verification right after (3).

Let $\lambda = 1$, $\varepsilon = 10^{-2}$, leading to $\mathbb{P}[X < \varepsilon] \approx 9.95\ 10^{-3}$, and (by numerical computations) $\theta^{(\min)} = \arg\min_\theta \mathbb{V}ar_\theta[g(X)L] \approx 159.68$; $\theta^* \approx 200.33$. Consider $k = 1$ in the CE technique with $\theta_0 = \lambda$; an estimator of $\theta^{(\min)}$ during the presimulation is given from (2) by

$$\hat{\theta}_{n_1} = \frac{(1/n_1)\sum_{i=1}^{n_1} I\{X_i \leq \varepsilon\}}{(1/n_1)\sum_{i=1}^{n_1} X_i I\{X_i \leq \varepsilon\}} \longrightarrow \frac{\mathbb{P}[X \leq \varepsilon]}{\mathbb{E}[XI\{X \leq \varepsilon\}]} = \theta^* \quad \text{a.s.}(n_1 \to \infty) \tag{4}$$

We executed 100 experiments with presimulation sample size $n_1 = 1000$, and later with $n = 10^6$ for the final estimation. We observed important variations in the results. The max variance of an experiment was $4.52\ 10^{-4}$ obtained when $\theta^{(\min)}$ was estimated as $\hat{\theta}_{n_1} = 492.25$ while the min variance was $5.34\ 10^{-5}$ obtained when $\hat{\theta}_{n_1} = 158.00$, hence a (large) relative ratio of 12. $\qquad\Box$

The disparity obtained when estimating the optimal parameter $\theta^{(\min)}$ in the last example illustrates that we may end up with a parameter selection yielding bad robustness properties. This can happen even if the parameter selection algorithm is performing, just because of (statistical) bad luck. As a consequence, it seems difficult to ensure a *strict* robustness property for the final estimator in adaptive IS simulation for a given choice of parameters, due to tits random nature. Our goal is to define a *probabilistic* robustness property in the next section, describing a satisfactory probabilistic evolution of the relative error in terms of the rarity parameter, and, afterwards, to characterize sufficient conditions to satisfy it. We also briefly look at the overall estimator (parameters plus final estimation) and present a property that it should verify.

Note that similar difficulties have been observed in the literature.

Example 2 Another example is taken from Section 3 of Chan et al. (2011): consider the sum of n i.i.d. rv X_i ($1 \leq i \leq n$) where each X_i follows an exponential distribution with parameter 1. The goal is to estimate $p = \mathbb{P}[X_1 + \cdots + X_n \geq \gamma]$ via IS, with the parametric family of exponential distributions with rate $\theta > 0$. It is shown that the optimal $\theta^{(CE)}$ when using the CE technique is such that (quoting Chan et al. (2011)): "when γ is sufficiently large, the estimation error in obtaining $\theta^{(CE)}$ in the multi-level CE procedure might be so substantial that it renders the resulting IS estimator unreliable". $\qquad\Box$

3 PROBABILISTIC BOUNDED RELATIVE ERROR DEFINITIONS

We formalize now several notions of Probabilistic Bounded Relative Error (P-BRE). We consider a measurable space $(\Omega, \mathscr{F})$ on which is defined a default probability measure $\mathbb{P}$, and a family of parameterized probability measures $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Let $X : \Omega \to \mathbb{R}$ be a rv, and $g : \mathbb{R} \to \mathbb{R}$ be the output function. We can assume that the output function depends on a rarity parameter ε, thus we denote $g_\varepsilon(\cdot)$; for instance $g_\varepsilon(x) = I\{x \in A(\varepsilon)\}$ for some sequence of Borel subsets $\{A(\varepsilon) : \varepsilon > 0\}$. The ε-problem is to estimate $\mu(\varepsilon) = \mathbb{E}[g_\varepsilon(X)]$, where $\mu(\varepsilon) \to 0$ as the rarity parameter $\varepsilon \to 0$. Another framework is when it is the probability measure that depends on ε, leading to a family $(\mathbb{P}_\varepsilon)_\varepsilon$ defined on $(\Omega, \mathscr{F})$, and $\mu(\varepsilon) = \mathbb{E}_\varepsilon[g(X)]$ with $g(x) = I\{x \in A\}$ for a given set A.

A member of the parameterized family can be used as an IS measure if $d\mathbb{P}_\theta(x) > 0$ whenever $g_\varepsilon(x)d\mathbb{P}(x) \neq 0$. The zero-variance probability measure $\mathbb{P}_\varepsilon^{(ZV)}$ for the ε-problem is defined by $d\mathbb{P}_\varepsilon^{(ZV)} = \frac{|g_\varepsilon(X)|}{\mathbb{E}[|g_\varepsilon(X)|]} d\mathbb{P}$, which is not necessarily a member of the parameterized family. However, suppose that there is some probability space $(\overline{\Omega}, \overline{\mathscr{F}}, \overline{\mathbb{P}})$ on which we can do experiments, for instance simulations, that allow us to learn $\mathbb{P}_\varepsilon^{(ZV)}$. More formally, equip Θ with a sigma-algebra $\mathscr{F}_\Theta$, and let $\widehat{\theta} : \overline{\Omega} \to \Theta$ be a measurable mapping. Suppose that we can construct such a rv for each ε-problem, resulting in a collection of rv $\{\widehat{\theta}(\varepsilon) : \varepsilon > 0\}$. The realizations $\theta(\varepsilon)$ of $\widehat{\theta}(\varepsilon)$ are used for estimating $\mu(\varepsilon)$ by the IS measure $\mathbb{P}_{\theta(\varepsilon)}$.

Suppose that, in this way, we have constructed for each $\varepsilon > 0$ an IS measure $\mathbb{P}_{\theta(\varepsilon)}$, and its associated IS estimator $Z(\theta(\varepsilon), \varepsilon)$ being an unbiased estimator of $\mu(\varepsilon)$. The relative error of this estimator is

$$\mathrm{RE}(\theta(\varepsilon), \varepsilon) = \frac{\sqrt{\mathbb{V}ar_{\theta(\varepsilon)}[Z(\theta(\varepsilon), \varepsilon)]}}{\mu(\varepsilon)}.$$

In the line of our observations in the previous section we recall the classical BRE property to a *realization* $(\theta(\varepsilon))_\varepsilon$ of $(\widehat{\theta}(\varepsilon))_\varepsilon$.

Definition 1 We say that the IS estimators $Z(\theta(\varepsilon), \varepsilon)$ show bounded relative errors (BRE) if

$$\exists K < \infty \text{ that does not depend on } \varepsilon \text{ and for which } \sup_{\varepsilon > 0} \mathrm{RE}(\theta(\varepsilon), \varepsilon) \leq K. \tag{5}$$

Equivalent conditions are $\sup_{\varepsilon > 0} \mathrm{RE}(\theta(\varepsilon), \varepsilon) < \infty$; $\sup_{\varepsilon > 0} \frac{\mathbb{E}_{\theta(\varepsilon)}[Z^2(\theta(\varepsilon), \varepsilon)]}{\mu^2(\varepsilon)} < \infty$.; or $\mathrm{RE}(\theta(\varepsilon), \varepsilon) = O(1)$ as $\varepsilon \to 0$.

However, as discussed before, the IS measure $\mathbb{P}_{\theta(\varepsilon)}$ for the ε-problem is chosen randomly according to some learning algorithm. Thus we deal with (random) IS estimators $Z(\widehat{\theta}(\varepsilon), \varepsilon)$ with *random relative errors* $\mathrm{RE}(\widehat{\theta}(\varepsilon), \varepsilon)$. In other words, these relative errors are rv $\overline{\Omega} \to \mathbb{R}$. Based on this observation, it is a natural step to cast the BRE property (5) in a probabilistic framework.

Definition 2

A. We say that the IS estimators $Z(\widehat{\theta}(\varepsilon), \varepsilon)$ show bounded square relative error in expectation, if their relative errors satisfy
$$\overline{\mathbb{E}}[(\mathrm{RE}(\widehat{\theta}(\varepsilon), \varepsilon))^2] = O(1) \quad (\varepsilon \to 0).$$

B. We say that the IS estimators $Z(\widehat{\theta}(\varepsilon), \varepsilon)$ show weak probabilistic bounded relative error, if

$$\forall \alpha \in (0, 1) \ \exists \text{ a constant } K < \infty \text{ such that } \inf_{\varepsilon > 0} \overline{\mathbb{P}}\left(\mathrm{RE}(\widehat{\theta}(\varepsilon), \varepsilon) \leq K\right) > \alpha.$$

C.   We say that the IS estimators $Z(\widehat{\theta}(\varepsilon), \varepsilon)$ show strong probabilistic bounded relative error, if

$$\forall \alpha \in (0,1) \; \exists \text{ a constant } K < \infty \text{ such that } \overline{\mathbb{P}}\left(\sup_{\varepsilon > 0} \text{RE}(\widehat{\theta}(\varepsilon), \varepsilon) \leq K\right) > \alpha.$$

Instead of ensuring BRE almost surely, we give a probabilistic guarantee to have BRE with any specified probability, after an estimation procedure of the parameter(s). Strong and weak probabilistic bounded relative error look at the evolution of the relative error as the rarity parameter goes to 0. The strong sense means that the upper-bound on the relative error has to be verified for every ε, while in the weak sense it is not needed, it has to be verified at least with a given probability.

It can be noted that bounded relative error in expectation is linked with the quality of the overall estimator (parameters plus final estimation). Indeed, from Chebyshev's inequality,

$$\mathbb{P}\left(\left|\frac{\widehat{Z}_n(\widehat{\theta}(\varepsilon), \varepsilon)}{\mu(\varepsilon)} - 1\right| \geq \gamma\right) \leq \frac{\mathbb{V}ar[Z(\widehat{\theta}(\varepsilon), \varepsilon)]}{\gamma^2(\mu(\varepsilon))^2 n} = \frac{\bar{\mathbb{E}}\left[(\text{RE}(\widehat{\theta}(\varepsilon), \varepsilon))^2\right]}{\gamma^2 n}.$$

The last equality comes from $\mathbb{V}ar[Z(\widehat{\theta}(\varepsilon), \varepsilon)] = \bar{\mathbb{E}}\left[\mathbb{V}ar[Z(\widehat{\theta}(\varepsilon), \varepsilon)]|\widehat{\theta}(\varepsilon)\right] = (\mu(\varepsilon)^2)\bar{\mathbb{E}}\left[(\text{RE}(\widehat{\theta}(\varepsilon), \varepsilon))^2\right].$
As a consequence, a bounded square relative error in expectation ensures an overall estimator with a bounded number n of replications to provide a sufficient accuracy.

From now we rather focus on illustrations and conditions for the strong and weak probabilistic bounded relative error properties, conditions for the property in expectation and the impact on the overall estimator being left for further research. We therefore focus on the evolution of the relative error in terms of ε, for a randomly chosen choice of parameters.

Remark 1 Our set-up and definition of randomization of the statistical performance of estimators follow the practice of the learning methods which we mentioned in the introduction. That is, typically one executes a two-stage approach, where the first stage determines parameters by a (pre)simulation procedure. The second stage is for estimating the wished value after having chosen and fixed the parameters. Another view on such randomization would be to select a parameter θ at random and execute the simulation using it. Although this seems to be a more natural approach, we do not follow it because we wanted to stay close to the practice mentioned above. Moreover, there are clearly two different probability mechanisms involved: selection of the parameter; running the simulations. Thus, to define properly a statistical property of the estimator, one needs to be unambiguous about these probability mechanisms.

Remark 2 As an important remark, one can note that in our definitions of robustness, the parameter $\hat{\theta}$ depends on ε – we do not mention here the required time/presimulation runs to obtain it. It might be the case that the required number of runs is $n_1(\varepsilon)$ and increases with ε. A truly robust algorithm can be said to be one for which the presimulation effort is also kept bounded as ε goes to zero. This would lead to definitions of robustness stronger than the one above, but our goal here is not to *define* or put in place such algorithms, it is just to highlight that robustness properties can only be probabilistic for properly chosen presimulation algorithms.

4   SUFFICIENT CONDITIONS FOR P-BRE

Suppose that there exists a collection of IS probability measures $\{\mathbb{P}_{\theta^*(\varepsilon)} : \varepsilon > 0\}$ for which the associated IS estimators $Z(\theta^*(\varepsilon), \varepsilon)$ of $\mu(\varepsilon)$ show bounded relative error; and suppose that we have estimated these probability measures by $\mathbb{P}_{\widehat{\theta}(\varepsilon)}$.

Proposition 1 Assume that for all $\alpha \in (0,1)$ there exists $K > 0$ such that for all $\varepsilon > 0$

$$\overline{\mathbb{P}}\left(\sup_{x:g_\varepsilon(x)\neq 0} \frac{d\mathbb{P}_{\theta^*(\varepsilon)}(x)}{d\mathbb{P}_{\widehat{\theta}(\varepsilon)}(x)} \leq K\right) > \alpha. \tag{6}$$

Then the IS estimators $Z(\widehat{\theta}(\varepsilon), \varepsilon)$ show weak probabilistic bounded relative error.

Proof. According to (5), as the estimators $Z(\theta^*(\varepsilon), \varepsilon)$ show bounded relative error, there is a constant K' such that for all $\varepsilon > 0$, $\mathbb{E}_{\theta^*(\varepsilon)}[Z^2(\theta^*(\varepsilon), \varepsilon)] \leq K'\mu^2(\varepsilon)$. Choose $\varepsilon > 0$ arbitrary, and let $\omega \in \overline{\Omega}$ such that for $\theta \stackrel{\text{def}}{=} \widehat{\theta}(\varepsilon)(\omega)$ the inequality in (6) holds; i.e., $\sup_{x:g_\varepsilon(x)\neq 0} \frac{d\mathbb{P}_{\theta^*(\varepsilon)}(x)}{d\mathbb{P}_\theta(x)} \leq K$. Then,

$$\mathbb{E}_\theta[Z^2(\theta,\varepsilon)] = \int g(x)^2 \left(\frac{d\mathbb{P}(x)}{d\mathbb{P}_\theta(x)}\right)^2 d\mathbb{P}_\theta(x) = \int g(x)^2 \left(\frac{d\mathbb{P}(x)}{d\mathbb{P}_{\theta^*(\varepsilon)}(x)}\right)^2 \left(\frac{d\mathbb{P}_{\theta^*(\varepsilon)}(x)}{d\mathbb{P}_\theta(x)}\right) d\mathbb{P}_{\theta^*(\varepsilon)}[x]$$

$$\leq K \int g(x)^2 \left(\frac{d\mathbb{P}(x)}{d\mathbb{P}_{\theta^*(\varepsilon)}(x)}\right)^2 d\mathbb{P}_{\theta^*(\varepsilon)}(x) = K\mathbb{E}_{\theta^*(\varepsilon)}\left[g(X)^2 \left(\frac{d\mathbb{P}(X)}{d\mathbb{P}_{\theta^*(\varepsilon)}(X)}\right)^2\right]$$

$$= K\mathbb{E}_{\theta^*(\varepsilon)}[Z^2(\theta^*(\varepsilon),\varepsilon)] \leq KK'\mu^2(\varepsilon).$$

Thus, choose $\tilde{K} = KK'$, then for all $\varepsilon > 0$ $\left\{\sup_{x:g(x)\neq 0} \frac{d\mathbb{P}_{\theta^*(\varepsilon)}(x)}{d\mathbb{P}_{\widehat{\theta}(\varepsilon)}(x)} \leq K\right\} \subset \left\{\frac{\mathbb{E}_{\widehat{\theta}(\varepsilon)}[Z^2(\widehat{\theta}(\varepsilon),\varepsilon)]}{\mu^2(\varepsilon)}, \leq \tilde{K}\right\}$. and $\forall \alpha \in (0,1) \exists \tilde{K} < \infty$ such that $\forall \varepsilon > 0 \, \overline{\mathbb{P}}\left(\text{RE}(\widehat{\theta}(\varepsilon),\varepsilon) \leq \tilde{K}\right) > \alpha.$ □

Proposition 2 Assume that for all $\alpha \in (0,1)$ there exists a finite constant K such that

$$\overline{\mathbb{P}}\left(\sup_{\varepsilon>0} \sup_{x:g_\varepsilon(x)\neq 0} \frac{d\mathbb{P}_{\theta^*(\varepsilon)}(x)}{d\mathbb{P}_{\widehat{\theta}(\varepsilon)}(x)} \leq K\right) > \alpha.$$

Then the IS estimators $Z(\widehat{\theta}(\varepsilon), \varepsilon)$ show strong probabilistic bounded relative error.

Proof. The proof follows the same line of reasoning as the proof of Proposition 1. □

Example 3 Coming back to Example 1, we know from the central limit theorem for a ratio of estimators (Asmussen and Glynn 2007) that, if we note $\hat{d}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i I\{X_i \leq \varepsilon\}$, the law of $\frac{\sqrt{n_1}(\hat{\theta}_{n_1} - \theta^*)}{\sigma/\hat{d}_{n_1}}$ converges when $n_1 \to \infty$ to a Normal law with mean 0 and variance 1, where

$$\sigma^2 = \sigma^2[I\{X \leq \varepsilon\}] - 2\theta^*\text{Cov}[I\{X \leq \varepsilon\}, XI\{X \leq \varepsilon\}] + (\theta^*)^2\sigma^2[XI\{X \leq \varepsilon\}]$$

$$= \mathbb{P}[X \leq \varepsilon](1 - \mathbb{P}[X \leq \varepsilon]) - 2\theta^*\mathbb{E}[XI\{X_{\leq}\varepsilon\}](1 - \mathbb{P}[X \leq \varepsilon]) + (\theta^*)^2 \left(\mathbb{E}[X^2 I\{X \leq \varepsilon\}] - (\mathbb{E}[XI\{X \leq \varepsilon\}])^2\right)$$

$$= -e^{-\lambda\varepsilon}(1 - e^{-\lambda\varepsilon})$$

$$+ \left(\frac{1 - e^{-\lambda\varepsilon}}{(1 - e^{-\lambda\varepsilon})/\lambda - \varepsilon e^{-\lambda\varepsilon}}\right)^2 \left(-\varepsilon^2 e^{-\lambda\varepsilon} - 2\frac{\varepsilon e^{-\lambda\varepsilon}}{\lambda} + 2\frac{1 - e^{-\lambda\varepsilon}}{\lambda^2} - ((1 - e^{-\lambda\varepsilon})/\lambda - \varepsilon e^{-\lambda\varepsilon})^2\right).$$

Then $|\hat{\theta}_{n_1} - \theta^*| \leq c_\alpha \frac{\sigma}{\hat{d}_{n_1}\sqrt{n_1}}$ with probability α where c_α is the $(1+\alpha)/2$ quantile of the standard normal distribution. Let $n_1 = n_1(\varepsilon)$ be large enough so the relative error is bounded by $\delta < 1$ independent of ε, i.e.,

the right hand-side of the above equation upper bounded by $\delta\theta^*$, or more exactly $n_1 \geq \left((c_\alpha\sigma)/(\mathbb{E}[\hat{d}]\delta\theta^*)\right)^2$. Then, with probability α,

$$\sup_{x:g(x)\neq 0} \frac{d\mathbb{P}_{\theta^*}[x]}{d\mathbb{P}_{\hat{\theta}_{n_1}}[x]} = \sup_{0\leq x\leq\varepsilon} \frac{\theta^* e^{-\theta^* x}}{\hat{\theta}_{n_1} e^{-\hat{\theta}_{n_1} x}} = \sup_{0\leq x\leq\varepsilon} \frac{\theta^*}{\hat{\theta}_{n_1}} e^{(\hat{\theta}_{n_1}-\theta^*)x} \leq \frac{1}{1-\delta} e^{\delta\theta^*\varepsilon}.$$

As, for ε small enough, $\theta^*\varepsilon \leq 3$ from (4), one can apply Proposition 2 with $K = \frac{1}{1-\delta}e^{3\delta}$. $\qquad\square$

5 HIGHLY RELIABLE MARKOVIAN SYSTEMS

Our previous sections were dealing with general probability distributions, but many models used in simulation are Markov chains. We therefore consider rare events in the context of discrete-time Markov chains $\{X_n, n = 0, 1, \ldots\}$ on a constant finite state space $\mathscr{S}$. Let Ω_X be the space of all sample paths of finite lengths $\mathbf{x} = (x_0, x_1, \ldots, x_{T(\mathbf{x})})$, $T(\mathbf{x}) \in \{1, 2, \ldots\}$, with sigma-algebra $\mathscr{F}_X$. Denote a matrix of transition probabilities by $\theta = (\theta_{ij})_{i,j\in\mathscr{S}}$. Associate with each matrix θ a probability measure $\mathbb{P}_\theta$ on $(\Omega_X, \mathscr{F}_X)$. Let Θ be the family of all matrices of transition probabilities; specifically, we consider a given collection of matrices $\{\overline{\theta}(\varepsilon)\} \subset \Theta$, parameterized by $\varepsilon > 0$. We call these the default or nominal matrices.

Furthermore, we assume here that the state space contains a perfect state 0, a set B of failed states, and the set U of remaining 'up' states, i.e., $\mathscr{S} = \{0\} \cup U \cup B$. For any $\mathbf{x} \in \Omega_X$ we define $T(\mathbf{x}) = \inf\{n : x_n \in \{0\} \cup B\}$, the stopping time of "return" to 0 or reaching failure, and for any state $i \in \mathscr{S}$ we let $\mu_i(\varepsilon) = \mathbb{P}_{\overline{\theta}(\varepsilon)}(X_{T(\mathbf{X})} \in B | X_0 = i) = \mathbb{E}_{\overline{\theta}(\varepsilon)}[I\{X_{T(\mathbf{X})} \in B\}|X_0 = i]$ the probability of reaching a failed state before the perfect state given that we start in i. Note that $\mu_0(\varepsilon) = 0$ and $\mu_i(\varepsilon) = 1$ whenever $i \in B$. The purpose is to estimate $\mu(\varepsilon) = \sum_{i\in\mathscr{S}} \overline{\theta}(\varepsilon)_{0i}\mu_i(\varepsilon)$, the probability that after leaving the perfect state 0, the chain will hit the failure set B before returning to 0. We assume that $\mu(\varepsilon) \to 0$ as $\varepsilon \to 0$. We are interested in efficient IS estimators of these performance measures. More on this type of model and specific IS schemes can be found in L'Ecuyer and Tuffin (2012), Nakayama (1996), Rubino and Tuffin (2009), Shahabuddin (1994).

Example 4 As a toy example, consider birth-death type of Markov chains on $\mathscr{S} = \{0, 1, \ldots, b\}$ for a finite constant b; thus we consider only transition probabilities satisfying $\theta_{i,i+1} + \theta_{i,i-1} = 1$; we let the states 0 and b be absorbing. Let $T(\mathbf{x}) = \inf\{n : x_n \in \{0, b\}\}$. For each $\theta \in \Theta$ we assume that $\mathbb{P}_\theta(T(\mathbf{X}) < \infty) = 1$, with $\mathbb{P}_\theta(X_{T(\mathbf{X})} = b | X_0 = i) > 0$ for all $i \neq 0$. Suppose that the nominal matrices $\overline{\theta}(\varepsilon) \in \Theta$ satisfy $\max_{i=1,\ldots,b-1} \overline{\theta}(\varepsilon)_{i,i+1} \leq \varepsilon$. Finally, define $\mu_i(\varepsilon) = \mathbb{P}_{\overline{\theta}(\varepsilon)}(X_{T(\mathbf{X})} = b | X_0 = i)$: the probability of absorption in state b when the chain starts in state i. $\qquad\square$

Suppose that for each $\varepsilon > 0$ there is besides the nominal matrix $\overline{\theta}(\varepsilon)$ some other transition matrix $\theta(\varepsilon) \in \Theta$, such that $\theta(\varepsilon)_{ij} > 0$ whenever $\overline{\theta}(\varepsilon)_{ij} > 0$. This matrix is called the reference matrix, and it induces an unbiased IS estimator of $\mu(\varepsilon)$ by $Z(\theta(\varepsilon), \varepsilon) \overset{\text{def}}{=} L(\mathbf{X}; \overline{\theta}(\varepsilon), \theta(\varepsilon))I\{X_{T(\mathbf{X})} \in B\}$ with likelihood ratio $L(\mathbf{x}; \overline{\theta}(\varepsilon), \theta(\varepsilon)) \overset{\text{def}}{=} \frac{d\mathbb{P}_{\overline{\theta}(\varepsilon)}(\mathbf{x})}{d\mathbb{P}_{\theta(\varepsilon)}(\mathbf{x})} = \prod_{n=0}^{T(\mathbf{x})-1} \frac{\overline{\theta}(\varepsilon)_{x_n,x_{n+1}}}{\theta(\varepsilon)_{x_n,x_{x+1}}}$, $\mathbf{x} \in \Omega_X$.

Recall that the zero-variance probability measure is induced by taking reference matrices $\theta^{(\mathrm{ZV})}(\varepsilon) \in \Theta$ for which

$$d\mathbb{P}_{\theta(\varepsilon)^{(\mathrm{ZV})}}(\mathbf{x}) = \frac{I\{x_{T(\mathbf{x})} \in B\}d\mathbb{P}_{\overline{\theta}(\varepsilon)}(\mathbf{x})}{\mu(\varepsilon)}, \quad \mathbf{x} \in \Omega_X. \tag{7}$$

It is known from L'Ecuyer and Tuffin (2012) that a matrix $\theta(\varepsilon)^{(ZV)}$ with $\forall i, j \in \mathscr{S}$, $i \neq j$, $\theta(\varepsilon)^{(ZV)}_{i,j} = \frac{\overline{\theta}(\varepsilon)_{i,j}\mu_j(\varepsilon)}{\mu_i(\varepsilon)}$. For instance, in the setting of Example 4, it gives

$$\theta(\varepsilon)^{(ZV)}_{i,i+1} = \frac{\overline{\theta}(\varepsilon)_{i,i+1}\mu_{i+1}(\varepsilon)}{\mu_i(\varepsilon)}. \tag{8}$$

We shall consider a sufficient condition for probabilistic bounded relative error. Denote by $\overline{\mathbb{P}}_{\widehat{\theta}(\varepsilon)}$ the probability measure on $(\Theta, \mathscr{F}_\Theta)$ induced by the rv $\widehat{\theta}(\varepsilon) : \overline{\Omega} \to \Theta$.

Proposition 3 Assume that for all $\alpha \in (0,1)$ there exist finite constants $K_1, K_2 > 0$ such that for all $\varepsilon > 0$ the (random) parameters are close enough the to zero-variance ones with probability at least α:

$$\overline{\mathbb{P}}\left(K_1 \leq \sup_{i,j \in \mathscr{S}, \theta^{(ZV)}(\varepsilon)_{ij} \neq 0} \frac{\theta^{(ZV)}(\varepsilon)_{ij}}{\widehat{\theta}(\varepsilon)_{ij}} \leq K_2\right) > \alpha. \tag{9}$$

Then, under the assumption that cycles have $\mathbb{P}_{\overline{\theta}(\varepsilon)}$-probability $O(\varepsilon^\delta)$ for some constant $\delta > 0$, the IS estimators $Z(\widehat{\theta}(\varepsilon), \varepsilon)$ show weak probabilistic bounded relative error.

Proof. Choose $\varepsilon > 0$ and a reference transition matrix $\theta = \theta(\varepsilon)$ from a realization of $\widehat{\theta}(\varepsilon)$ such that

$$K_1 \leq \sup_{i,j \in \mathscr{S}, \theta^{(ZV)}(\varepsilon)_{ij} \neq 0} \leq \frac{\theta^{(ZV)}(\varepsilon)_{ij}}{\theta} \leq K_2.$$

We wish to show that there is a constant $\tilde{K}$ which does not depend on ε, such that $\frac{\mathbb{E}_\theta[Z^2(\theta,\varepsilon)]}{\mu^2(\varepsilon)} \leq \tilde{K}$. Then P-BRE follows from the fact that, by assumption, $\theta = \theta(\varepsilon)$ is a realization of $\widehat{\theta}(\varepsilon)$ such that $\sup_{i,j \in \mathscr{S}} \frac{\theta^{(ZV)}(\varepsilon)_{ij}}{\theta} \leq K_2$ with probability at least α.

Note that we can rewrite this ratio using the expression for the zero-variance probability measure given in (7) which says that $\frac{I\{X_{T(\mathbf{X})} \in B\} L^2(\mathbf{X};\overline{\theta}(\varepsilon);\theta^{(ZV)}(\varepsilon))}{\mu^2(\varepsilon)} = 1$ a.s., and then bound it using that θ in the set given in (9):

$$\begin{aligned}
\frac{\mathbb{E}_\theta[Z^2(\theta,\varepsilon)]}{\mu^2(\varepsilon)} &= \frac{\mathbb{E}_\theta[I\{X_{T(\mathbf{X})} \in B\} L^2(\mathbf{X};\overline{\theta}(\varepsilon);\theta)]}{\mu^2(\varepsilon)} \\
&= \frac{\mathbb{E}_\theta[I\{X_{T(\mathbf{X})} \in B\} L^2(\mathbf{X};\overline{\theta}(\varepsilon);\theta^{(ZV)}(\varepsilon)) L^2(\mathbf{X};\theta^{(ZV)}(\varepsilon);\theta)]}{\mu^2(\varepsilon)} \\
&= \mathbb{E}_\theta[I\{X_{T(\mathbf{X})} \in B\} L^2(\mathbf{X};\theta^{(ZV)}(\varepsilon);\theta)] = \mathbb{E}_\theta\left[I\{X_{T(\mathbf{X})} \in B\}\left(\frac{d\mathbb{P}_{\theta^{(ZV)}(\varepsilon)}}{d\mathbb{P}_\theta}(\mathbf{X})\right)^2\right] \leq \mathbb{E}_\theta[K_2^{2T(\mathbf{X})}].
\end{aligned}$$

Define $\mathscr{A}(i)$ to be the set of sample paths that, starting in state $i \in \mathscr{S}$, reach the failure set before state 0 without cycles: $\mathscr{A}(i) = \{\mathbf{x} \in \Omega_X : x_0 = i; x_{T(\mathbf{x})} \in B; \text{no cycles}\}$. Let $\mathscr{A} = \bigcup_{i \in \mathscr{S}} \mathscr{A}(i)$. Because the state space $\mathscr{S}$ is finite, there is a finite constant m such that for any $\varepsilon > 0$ all sample paths $\mathbf{x} \in \mathscr{A}$ have length $T(\mathbf{x}) < m$. Denote $p_0(\varepsilon) = \mathbb{P}_\theta(\mathscr{A})$. Note that p_0 depends on ε because transition matrix $\theta = \theta(\varepsilon)$ refers to the estimate $\widehat{\theta}(\varepsilon)$ of the zero-variance matrix for the ε-problem.

Clearly, $\mathbb{P}_\theta(T(\mathbf{X}) \geq m) \leq \mathbb{P}_\theta(\mathscr{A}^c) = 1 - p_0(\varepsilon)$. From this we can reason that $\mathbb{P}_\theta(T(\mathbf{X}) \geq km) \leq (1 - p_0(\varepsilon))^k$ for all $k = 0, 1, \ldots$ (see also the proof of Theorem 1 in L'Ecuyer and Tuffin (2012)). Hence,

$T(\mathbf{X})/m$ is stochastically smaller than a geometric rv Y (on $0,1,\ldots$) with parameter $p_0(\varepsilon)$. Using the expression of the generating function $\mathbb{E}_\theta[z^Y]$ of Y, we get

$$\mathbb{E}_\theta[K_2^{2T(\mathbf{X})}] = \mathbb{E}_\theta[(K_2^{2m})^{T(\mathbf{X})/m}] \leq \mathbb{E}_\theta[(K_2^{2m})^Y] = \frac{p_0(\varepsilon)}{1 - (1 - p_0(\varepsilon))K_2^{2m}},$$

if we are able to prove that $(1 - p_0(\varepsilon))K^{2m} < 1$ (for ε small enough). Then BRE will be obtained.

But, following exactly the proof of Theorem 2 in L'Ecuyer and Tuffin (2012), thanks to our assumptions, we can show that $p_0(\varepsilon) \to 1$ as $\varepsilon \to 0$, hence the result. This basically comes from the fact that any path $\mathbf{x}$ is such that $\mathbb{P}_\theta[x] = \Theta(\frac{\mathbb{P}[x]}{\mu(\varepsilon)})$ by expanding the probability of the path as the product of probabilities of individual transitions and using the bounds in terms of the zero-variance change of measure. As a consequence *dominant* paths (those of probability $\Theta(\mu(\varepsilon))$ under the original distribution) have probability $\Theta(1)$ while non-dominant paths have probability $o(1)$ and therefore $p_0(\varepsilon) \to 1$. $\qquad\square$

6 $M/M/1/b$ **MODEL**

As a special case of Section 5 we consider a Markov chain on $\mathscr{S} = \{0, 1, \ldots, b\}$ for a finite constant b with transition probabilities

$$\overline{\theta}(\varepsilon)_{i,i+1} = \varepsilon, \ \ \overline{\theta}(\varepsilon)_{i,i-1} = 1 - \varepsilon, \ \ (i = 1, \ldots, b - 1),$$

for $\varepsilon > 0$. This models the discrete-time Markov chain associated with the continuous-time $M/M/1/b$ Markov chain by embedding at the jump times. States 0 and b are absorbing, $\overline{\theta}(\varepsilon)_{00} = 1$, $\overline{\theta}(\varepsilon)_{bb} = 1$. For instance, when $\varepsilon = k^{-\delta}$ for some $\delta > 0$ and $k \to \infty$, the decaying is polynomially fast. When $\varepsilon = e^{-\delta k}$ for some $\delta > 0$, the decaying is exponentially fast as $k \to \infty$. The output function $I\{x_{T(\mathbf{x})} = b\}$ is absorption in state b, where $T(\mathbf{x})$ is the first entrance time in the absorption set, and $I\{A\}$ denotes the indicator function of event A. Suppose that the chain starts in state i, then we denote

$$\mu_i(\varepsilon) \overset{\text{def}}{=} \mathbb{P}_{\overline{\theta}(\varepsilon)}(X_{T(\mathbf{X})} = b | X_0 = i).$$

Let $\rho(\varepsilon) \overset{\text{def}}{=} (\overline{\theta}(\varepsilon)_{i,i-1})/(\overline{\theta}(\varepsilon)_{i,i+1}) = (1 - \varepsilon)/\varepsilon$. Then it is well-known, e.g., see Chapter XIV in Feller (1968), that

$$\mu_i(\varepsilon) = \frac{\rho(\varepsilon)^i - 1}{\rho(\varepsilon)^b - 1}, \ \ i = 0, \ldots, b. \tag{10}$$

Specifically we are interested in $\mu_1(\varepsilon)$ for $\varepsilon \to 0$:

$$\mu_1(\varepsilon) = \frac{\rho(\varepsilon) - 1}{\rho(\varepsilon)^b - 1} = \frac{(1 - \varepsilon)/\varepsilon - 1}{(1 - \varepsilon)^b/\varepsilon^b - 1} = \frac{(1 - 2\varepsilon)\varepsilon^{b-1}}{(1 - \varepsilon)^b - \varepsilon^b} = \Theta\left(\varepsilon^{b-1}\right), \ \ \varepsilon \to 0.$$

Thus, the rare event probability $\mu_1(\varepsilon)$ decays to 0 polynomially in ε as $\varepsilon \to 0$.

6.1 An Estimator with Bounded Relative Error

Define transition matrices $\{\theta^*(\varepsilon), \varepsilon > 0\}$ by $\theta^*(\varepsilon)_{i,i+1} = \overline{\theta}(\varepsilon)_{i,i-1} = 1 - \varepsilon$, $\theta^*(\varepsilon)_{i,i-1} = \overline{\theta}(\varepsilon)_{i,i+1} = \varepsilon$, $(i = 1, \ldots, b - 1)$, for $\varepsilon > 0$. States 0 and b are again absorbing. It is known that by interchanging arrival and service jump probabilities the IS estimators show BRE when $b \to \infty$ while keeping the transition probabilities $\overline{\theta}(\varepsilon) \equiv \overline{\theta}$ not dependent on ε. The same holds true in our study of absorption probabilities for systems with constant state space as we shall show formally.

Proposition 4 Suppose that we apply IS simulation with these $\theta^*(\varepsilon)$ matrices to estimate $\mu_1(\varepsilon)$. Then the associated IS estimators $Z(\theta^*(\varepsilon), \varepsilon)$ show bounded relative error.

Proof. Recall that $Z(\theta^*(\varepsilon), \varepsilon) = \frac{d\mathbb{P}_{\overline{\theta}(\varepsilon)}(\mathbf{X})}{d\mathbb{P}_{\theta^*(\varepsilon)}(\mathbf{X})} I\{X_{T(\mathbf{X})} = b\}$. The likelihood ratio of a path $\mathbf{X}$ that reaches b equals

$$\frac{d\mathbb{P}_{\overline{\theta}(\varepsilon)}(\mathbf{X})}{d\mathbb{P}_{\theta^*(\varepsilon)}(\mathbf{X})} = \prod_{i=1}^{b-1} \frac{\overline{\theta}(\varepsilon)_{i,i+1}}{\theta^*(\varepsilon)_{i,i+1}} \times \prod_{i=2}^{b-1} \left(\frac{\overline{\theta}(\varepsilon)_{i,i-1}}{\theta^*(\varepsilon)_{i,i-1}} \frac{\overline{\theta}(\varepsilon)_{i-1,i}}{\theta^*(\varepsilon)_{i-1,i}} \right)^{N_i(\mathbf{X})},$$

where the second $\prod$-factor takes into account all cycles: $N_i(\mathbf{X})$ is the number of times the cycle $i \to i-1 \to i$ occurs. By definition of the IS transition probabilities $\theta^*(\varepsilon)_{ij}$, this cycle product equals 1. Thus

$$\mathbb{E}_{\theta^*(\varepsilon)}[Z^2(\theta^*(\varepsilon), \varepsilon)] = \left(\prod_{i=1}^{b-1} \frac{\varepsilon}{1-\varepsilon} \right)^2 \mathbb{E}_{\theta^*(\varepsilon)}[I\{X_{T(\mathbf{X})} = b\}] = \left(\frac{\varepsilon}{1-\varepsilon} \right)^{2(b-1)} \mathbb{P}_{\theta^*(\varepsilon)}(X_{T(\mathbf{X})} = b).$$

But $\mathbb{P}_{\theta^*(\varepsilon)}(X_{T(\mathbf{X})} = b) = \frac{1-\rho(\varepsilon)^{-1}}{1-\rho(\varepsilon)^{-b}}$, where $\rho(\varepsilon)^{-1} = \frac{\varepsilon}{1-\varepsilon} \to 0$ with ε. So,

$$\frac{\mathbb{E}_{\theta^*(\varepsilon)}[Z^2(\theta^*(\varepsilon), \varepsilon)]}{\mu_1(\varepsilon)^2} = \rho(\varepsilon)^{-2(b-1)} \frac{1-\rho(\varepsilon)^{-1}}{1-\rho(\varepsilon)^{-b}} \frac{(\rho(\varepsilon)^b - 1)^2}{(\rho(\varepsilon) - 1)^2} \overset{\text{(calculus)}}{=} \frac{1-\rho(\varepsilon)^{-b}}{1-\rho(\varepsilon)^{-1}} \le 2(1 - \rho(\varepsilon)^{-b}) < 2,$$

for all $\varepsilon < \varepsilon_0$ for which $1 - \rho(\varepsilon_0)^{-1} > \frac{1}{2}$. $\qquad\square$

Using (8) and (10) the zero-variance probability measure gives $\theta(\varepsilon)_{i,i+1}^{(\text{ZV})} = \varepsilon \frac{\rho(\varepsilon)^{i+1} - 1}{\rho(\varepsilon)^i - 1}$, $\theta(\varepsilon)_{i,i-1}^{(\text{ZV})} = (1-\varepsilon) \frac{\rho(\varepsilon)^{i-1} - 1}{\rho(\varepsilon)^i - 1}$, for $1 \le i \le b-1$. Note that $\theta(\varepsilon)_{1,0}^{(\text{ZV})} = 0$, thus $\mathbb{P}_{\theta(\varepsilon)^{(\text{ZV})}}(X_{T(\mathbf{X})} = b) = 1$.

6.2 Cross-Entropy Estimators

Applying the cross-entropy method with a pre-simulation sample size n_1 results in estimates

$$\widehat{\theta}(\varepsilon)_{i,i+1}[n_1] = \frac{\sum_{r=1}^{n_1} I\{\mathbf{X}_{T(\mathbf{X}^{(r)})}^{(r)} = b\} N_{i,i+1}(\mathbf{X}^{(r)})}{\sum_{r=1}^{n_1} I\{\mathbf{X}_{T(\mathbf{X}^{(r)})}^{(r)} = b\} \left(N_{i,i+1}(\mathbf{X}^{(r)}) + N_{i,i-1}(\mathbf{X}^{(r)}) \right)}, \tag{11}$$

where $\mathbf{X}^{(r)}$ is the r-th simulated sample path, and $N_{i,i+1}(\mathbf{X})$ counts the number of transitions $i \to i+1$ on the sample path $\mathbf{X}$. The simulations have been executed under $\mathbb{P}_{\overline{\theta}(\varepsilon)}$ but $\mathbb{P}_{\theta(\varepsilon)}$ could also be used, in which case the numerator and denominator in (11) should include the likelihood ratio $L(\mathbf{X}^{(r)}; \overline{\theta}(\varepsilon), \theta(\varepsilon))$. In both cases, we obtain

$$\lim_{n_1 \to \infty} \widehat{\theta}(\varepsilon)_{i,i+1}[n_1] = \frac{\mathbb{E}_{\overline{\theta}(\varepsilon)}[I\{\mathbf{X}_{T(\mathbf{X})} = b\} N_{i,i+1}(\mathbf{X})]}{\mathbb{E}_{\overline{\theta}(\varepsilon)}[I\{\mathbf{X}_{T(\mathbf{X})} = b\} (N_{i,i+1}(\mathbf{X}) + N_{i,i-1}(\mathbf{X}))]} = \theta(\varepsilon)_{i,i+1}^{(\text{ZV})} \quad \text{a.s.} \tag{12}$$

The last equality in (12) has been shown in Ridder (2010). Estimator (11) satisfies the central limit theorem, see for instance Asmussen and Glynn (2007), page 107):

$$\sqrt{n_1} \left(\widehat{\theta}(\varepsilon)_{i,i+1}[n_1] - \theta(\varepsilon)_{i,i+1}^{(\text{ZV})} \right) \overset{\text{d}}{\to} N(0, \sigma^2(\varepsilon)) \qquad (n_1 \to \infty)$$

for some $\sigma^2(\varepsilon)$. Apply the delta method for obtaining an expression for the variance $\sigma^2(\varepsilon)$. Hence, denoting $z_{1-\alpha/2}$ the $1 - \alpha/2$ quantile of the standard normal distribution, we get for n_1 large enough

$$\overline{\mathbb{P}}_{\widehat{\theta}(\varepsilon)} \left(\left| \widehat{\theta}(\varepsilon)_{i,i+1}[n_1] - \theta(\varepsilon)_{i,i+1}^{(\text{ZV})} \right| < z_{1-\alpha/2} \frac{\sigma(\varepsilon)}{\sqrt{n_1}} \right) \approx 1 - \alpha.$$

Proposition 5 The cross-entropy estimators satisfy probabilistic bounded relative error.

Proof.　　Apply Proposition 3 while using (12) and the constant size of state space for all ε.　　　□

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their insightful comments.

REFERENCES

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

Chan, J., P. Glynn, and D. Kroese. 2011. "A comparison of cross-entropy and variance minimization strategies". *Journal of Applied Probability* 48A:183–194.

Feller, W. 1968. *An Introduction to Probability Theory and its Applications*. Third ed, Volume 1. Wiley.

Heidelberger, P. 1995. "Fast Simulation of Rare Events in Queueing and Reliability Models". *ACM Transactions on Modeling and Computer Simulation* 5 (1): 43–85.

Juneja, S., and P. Shahabuddin. 2006. "Rare Event Simulation Techniques: An Introduction and Recent Advances". In *Simulation*, edited by S. G. Henderson and B. L. Nelson, Handbooks in Operations Research and Management Science, 291–350. Amsterdam, The Netherlands: Elsevier. Chapter 11.

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

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

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

Ridder, A. 2010. "Asymptotic optimality of the cross-entropy method for Markov chain problems". *Procedia Computer Science* 1 (1): 1571 – 1578. ICCS 2010.

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

Rubinstein, R. Y. 1999. "The Cross-Entropy Method for Combinatorial and Continuous Optimization". *Methodology and Computing in Applied Probability* 2:127–190.

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

AUTHOR BIOGRAPHIES

AD RIDDER is an associate professor at Department of Econometrics of the Vrije Universiteit in Amsterdam (the Netherlands), where he teaches courses in simulation, numerical methods, and stochastic models. His main research interests are in the area of applied probability, rare event simulation and performance evaluation of stochastic systems.

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.