Ridder, Ad; Tuffin, Bruno

Working Paper

Probabilistic Bounded Relative Error Property for Learning Rare Event Simulation Techniques

Tinbergen Institute Discussion Paper, No. 12-103/III

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Ridder, Ad; Tuffin, Bruno (2012) : Probabilistic Bounded Relative Error Property for Learning Rare Event Simulation Techniques, Tinbergen Institute Discussion Paper, No. 12-103/III, Tinbergen Institute, Amsterdam and Rotterdam, https://nbn-resolving.de/urn:nbn:nl:ui:31-1871/38762

This Version is available at:
https://hdl.handle.net/10419/87274

TI 2012-103/III
Tinbergen Institute Discussion Paper

Probabilistic Bounded Relative Error Property for Learning Rare Event Simulation Techniques

Ad Ridder[1]
Bruno Tuffin[2]

[1] *Faculty of Economics and Business Administration, VU University Amsterdam, and Tinbergen Institute;*
[2] *Inria Rennes Bretagne Atlantique.*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Probabilistic Bounded Relative Error Property for Learning Rare Event Simulation Techniques

Ad Ridder
Vrije Universiteit & Tinbergen Institute
De Boelelaan 1105
1081 HV Amsterdam
THE NETHERLANDS

Bruno Tuffin
Inria Rennes Bretagne Atlantique
Campus Universitaire de Beaulieu
35042 Rennes Cedex
FRANCE

October 2, 2012

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 have been defined in the literature, that an estimator is expected to satisfy. Though, those properties are not adapted to estimators for which the estimators come from a parametric family and the optimal parameter is learned and random. For this reason, we motivate in this paper the need to define probabilistic robustness properties, because the accuracy of the resulting estimator is therefore random. 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, illustrate them and simple but standard examples, and hope that it will foster discussions and new works in the area.

Keywords: Rare-event probability, Cross-entropy method, Markov chains, Probabilistic robustness properties

JEL Classifiacation Number: C6

1 Introduction

Rare event simulation has been the topic of an extensive research during the past thirty years; see [5, 6, 13] 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, the outcome when it happens 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, thus to analyze. To circumvent this problem, sophisticated techniques have been designed, are can be organized in two main families: *importance sampling* (IS), which consists in 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 of interest, and *splitting*, which basically favors "successful" trajectories, i.e., those that get closer to the rare event, by replicating them in the form of a number of offsprings, and proceeds successively that way until the event is reached.

A key issue when designing a rare event probability estimator is to determine wether 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 small value of the probability. There exist several definitions of such a 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, or 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. Remark though, there exist numerous other properties, dealing with moments of order larger than 2, the normal approximation, 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 [4, 7].

But many rare event estimators come from adaptive techniques where parameters leading to a valid estimator are learned and as a consequence random, since depending on sample values. A typical example is the *cross-entropy*-based (CE) rare event estimation, part of the broader class of adaptive IS. In adaptive IS, the idea is to determine from empirical values the parameters minimizing the estimator's variance. In CE the optimal parameters are more specifically derived from the minimization of the Kullback-Leibler distance between the considered parametric family and the zero-variance IS change of measure [14] obtained from empirical results. When the optimal parameters are approached, the rare event can be estimated thanks too a long(er) simulation. Since the parameters are random, it is then hardly possible to guarantee that a robustness property is satisfied. 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 threefolds:

- we aim first at illustrating that getting a robustness property may be itself random for adap-

tive methods, and that an adaptive algorithm can yield estimators with a wide diversity of results. We limit ourselves in this paper to importance sampling.

- 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*, but the other existing robustness properties can easily be extended to the probabilistic case in a similar way and without complication.

- Our third goal is to provide sufficient conditions under which those probabilistic properties are verified, and to give some illustrations.

The rest of this paper is organized as follows. Section 2 describes the basic ideas and steps of adaptive IS, and more specifically 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 respectively described in Section 5 and Section 6. Finally, Section 7 concludes the paper and gives some directions for future research.

2 Rare event simulation, adaptive techniques and related robustness issue

As described in the previous section and to simplify the presentation, we limit ourselves to IS, but a similar analysis can be performed for other rare event simulation techniques. We therefore start with a brief description of IS, before an introduction to adaptive and cross-entropy techniques.

Consider the estimation of

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

where X is a random variable 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 $L(X) = d\mathbb{P}(x)/d\mathbb{P}_\theta(x)$ likelihood ratio.

Applying IS is of particular interest when trying to estimate the probability p of a rare event A, $p = \mathbb{P}[A] \ll 1$. To estimate this probability by a crude Monte Carlo simulation, we just sample n independent copies $(X_i)_{1 \leq i \leq n}$ of the Bernoulli random variable $X = 1(A)$, whose value is 1 of A is reached and 0 otherwise, and use as an unbiased estimator the proportion of times A is reached over the sample, $(1/n)\sum_{i=1}^{n} X_i$. But for rare events, it is unlikely that the rare event occurs only once if n is not large: for example, to estimate a probability of 10^{-9}, we need in average a

sample of size 10^9 to get the event once, and much more if we want to get a confidence interval. Moreover, if applying the central limit theorem (assuming it relevant for n l arge enough), one can note that the *relative* half-width of the confidence interval, given by this half-width divided by the expected value we are computing, i.e., $n^{-1/2}c_\alpha\sigma/p$, is $c_\alpha\sqrt{1-p}/\sqrt{np} \to \infty$ as $p \to 0$, with c_α the $1-\alpha/2$ quantile of the standard normal distribution (with mean 0 and variance 1). 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 [8] and reduce the *relative error*. The relative error of an estimator X of p and with variance σ^2 is defined as $\mathrm{RE}[X] = \sigma/p$. An estimator X will be said to verify *bounded relative error* (BRE) if $\mathrm{RE}[X]$ remains bounded as $\mathbb{E}[X] = p \to 0$. In that case, the sample size needed to get a specified relative accuracy is bounded whatever the rarity of the event.

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 $g(X)L(X)$ under IS. While this value can be learnt 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_k of independent copies on X generated according to $\mathbb{P}_{\theta_{j-1}}$
   (b) determine the value θ_j minimizing the variance of $g(X)L(X)$.
3. Simulation:
   (a) use a sample n of independent copies on X generated according to $\mathbb{P}_{\theta_k}$
   (b) provide an estimator of $\mathbb{E}_{\theta_k}[g(X)L(X)]$ 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 $\mathbb{P}^{(\mathrm{ZV})}$ and $\mathbb{P}_\theta$:

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

When estimating $\mathbb{E}[g(X)]$, it is known that the optimal change of measure is $d\mathbb{P}^{(\mathrm{ZV})} = \frac{|g(X)|}{\mathbb{E}[|g(X)|]} d\mathbb{P}$ [1]. This gives after straightforward simplifications

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

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

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

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 learnt values of θ yielding estimators experiencing large variations in their variance.

Example 1. *Consider a random variable X exponentially distributed with rate λ, and assume that we want to compute $\mathbb{P}[X<\varepsilon]=\mathbb{E}[g(X)]=1-e^{-\lambda\varepsilon}$ with $g(x)=1_{[0,\varepsilon]}$. Suppose that we use importance sampling and still sample from an exponential density, but with a different rate θ, The second moment of that IS estimator is*

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

The estimator satisfies bounded relative error as $\varepsilon\to 0$ if $\mathbb{E}_\theta[g(X)^2L^2]/(\mathbb{E}[g(X)])^2$ 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.

Let $\lambda=1$, $\varepsilon=10^{-2}$, leading to $\mathbb{P}[X<\varepsilon]\approx 9.95\,10^{-3}$ and consider $k=1$ in the CE technique with $n_1=1000$ and $\theta_0=\lambda$, and later $n=10^6$ for the final estimation. An estimator of $\theta^{(\min)}$ during the presimulation is given from (1) by

$$\hat\theta_{n_1}=\frac{(1/n_1)\sum_{i=1}^{n_1}1_{[X_i\le\varepsilon]}}{(1/n_1)\sum_{i=1}^{n_1}X_i1_{[X_i\le\varepsilon]}}.$$

We also get

$$\hat\theta_{n_1}\longrightarrow_{n_1\to\infty}\frac{\mathbb{P}[X\le\varepsilon]}{\mathbb{E}[X1_{[X\le\varepsilon]}]}=\frac{1-e^{-\lambda\varepsilon}}{(1-e^{-\lambda\varepsilon})/\lambda-\varepsilon e^{-\lambda\varepsilon}}=\frac{2}{\varepsilon}+o(1), \qquad (3)$$

this limit yielding a parameter θ^* with BRE from our verification using right after (2).

Considering $n_1=100$ independent experiments like this, we have observed important variations in the results. The maximal variance of an experiment was $4.52\,10^{-4}$ obtained when $\theta^{(\min)}$ was estimated as $\hat\theta_{n_1}=492.25$ while the minimal variance was $5.34\,10^{-5}$ obtained when the estimated parameters was $\hat\theta_{n_1}=158.00$, hence a (large) relative ratio of 12.

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 performant, 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, due to the random nature of the parameters. Our goal is to define a *probabilistic* robustness property in the next section and, afterwards, to characterize sufficient conditions to satisfy it.

Similar difficulties have been observed in the literature:

Example 2. *Another example is taken from [2], Section 3: consider the sum of n i.i.d. random variables X_i ($1 \leq i \leq n$) where each X_i follows an exponential distribution with parameter 1. The goal is to estimate $p = \mathbb{P}[X_1 + \cdots + X_{\geq}\gamma]$ via IS, i 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 [2]): "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 importance sampling estimator unreliable".*

Remark 1. *The principle of adaptive techniques is then to first estimate the optimal parameter θ^* by $\hat{\theta}_{n_1}$ with a sample size n_1 and then to use it in the IS change of measure to get an estimator $\hat{g}_{\hat{\theta}_{n_1},n}$ of the searched random variable. The question is (again) to understand the robustness properties of the estimator with this (then fixed) parameter $\hat{\theta}_{n_1}$. There would be an alternative way to proceed to in order to look for a* deterministic *robustness property: it would consist in repeating the full experience, parameter estimation plus simulation, m times independently, and to apply the central limit theorem to those m trials. The estimator we consider in this case is then $\hat{g}_{\hat{\theta}_{n_1},n}$ with $\hat{\theta}_{n_1}$ as a random variable. The variance of this estimator is*

$$\begin{aligned}\sigma^2\left[\hat{g}_{\hat{\theta}_{n_1},n}\right] &= \mathbb{E}\left[\sigma^2\left[\hat{g}_{\hat{\theta}_{n_1},n}\middle|\hat{\theta}_{n_1}\right]\right] + \sigma^2\left[\mathbb{E}\left[\hat{g}_{\hat{\theta}_{n_1},n}\middle|\hat{\theta}_{n_1}\right]\right] \\ &= \mathbb{E}\left[\sigma^2\left[\hat{g}_{\hat{\theta}_{n_1},n}\middle|\hat{\theta}_{n_1}\right]\right]\end{aligned}$$

where the second term on the first line is zero from the unbiasedness of the estimator (assuming the support of the change of measure not affected by the parameter choice), and the remaining term is the expected variance. But again, this means that the considered estimator is one for which the IS parameter search (and the resulting simulation) is repeated m times independently, which does not correspond to what is done in practice.

3 Probabilistic Bounded Relative Error Definitions

We formalize now several notions of Probabilistic Bounded Relative Error (P-BRE). We start by introducing few useful notations. similarly in a full version of this draft.

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 random variable, 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 importance sampling measure if $d\mathbb{P}_\theta(x) > 0$ whenever $g_\varepsilon(x)d\mathbb{P}(x) \neq 0$. The zero-variance probability measure $\mathbb{P}_\varepsilon^{(\mathrm{ZV})}$ for the ε-problem is defined by

$$d\mathbb{P}_\varepsilon^{(\mathrm{ZV})} = \frac{|g_\varepsilon(X)|}{\mathbb{E}[|g_\varepsilon(X)|]} d\mathbb{P},$$

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 $\mathbb{P}_\varepsilon^{(\mathrm{ZV})}$. 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 random variable for each ε-problem, resulting in a collection of random variables $\{\widehat{\theta}(\varepsilon) : \varepsilon > 0\}$. The realizations $\theta(\varepsilon)$ of $\widehat{\theta}(\varepsilon)$ are used for estimating $\mu(\varepsilon)$ by the importance sampling measure $\mathbb{P}_{\theta(\varepsilon)}$.

Suppose that, in this way, we have constructed for each $\varepsilon > 0$ an importance sampling measure $\mathbb{P}_{\theta(\varepsilon)}$, and their associated importance sampling estimators $Z_{\theta(\varepsilon)}(\varepsilon)$ being unbiased estimators of $\mu(\varepsilon)$.

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*

$$\exists K < \infty \text{ that does not depend on } \varepsilon \text{ and for which } \sup_{\varepsilon > 0} \mathrm{RE}(\theta(\varepsilon), \varepsilon) \leq K. \tag{4}$$

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 (4) 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*

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

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). 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.

Remark 2. *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 3. *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 importance sampling probability measures $\{\mathbb{P}_{\theta^*(\varepsilon)} : \varepsilon > 0\}$ for which the associated importance sampling estimators $Z_{\theta^*(\varepsilon)}(\varepsilon)$ of $\mu(\varepsilon)$ show bounded relative error; and suppose that we have estimated these probability measures by $\mathbb{P}_{\widehat{\theta}(\varepsilon)}$.

The next propositions show that if the targeted probability distributions are sufficiently approximated (through the likelihood ratio), then P-BRE properties can be satisfied.

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

$$\overline{\mathbb{P}}\left(\sup_{x:g_\varepsilon(x)\neq 0} \frac{d\mathbb{P}_{\theta^*(\varepsilon)}(x)}{d\mathbb{P}_{\widehat{\theta}(\varepsilon)}(x)} \leq K \right) > \alpha. \tag{5}$$

Then the importance sampling estimators $Z_{\widehat{\theta}(\varepsilon)}(\varepsilon)$ show weak probabilistic bounded relative error.

Proof. According to (4), 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 (5) 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,

$$\begin{aligned}
&\mathbb{E}_\theta[Z^2_\theta(\varepsilon)] = \int g(x)^2 \left(\frac{d\mathbb{P}(x)}{d\mathbb{P}_\theta(x)}\right)^2 d\mathbb{P}_\theta(x) \\
&= \int g(x)^2 \left(\frac{d\mathbb{P}(x)}{d\mathbb{P}_{\theta^*(\varepsilon)}(x)}\right)^2 \left(\frac{d\mathbb{P}_{\theta^*(\varepsilon)}(x)}{d\mathbb{P}_\theta(x)}\right) d\mathbb{P}_{\theta^*(\varepsilon)}[x] \\
&\leq K \int g(x)^2 \left(\frac{d\mathbb{P}(x)}{d\mathbb{P}_{\theta^*(\varepsilon)}(x)}\right)^2 d\mathbb{P}_{\theta^*(\varepsilon)}(x) = K\mathbb{E}_{\theta^*(\varepsilon)}\left[g(X)^2 \left(\frac{d\mathbb{P}(X)}{d\mathbb{P}_{\theta^*(\varepsilon)}(X)}\right)^2 \right] \\
&= K\mathbb{E}_{\theta^*(\varepsilon)}[Z^2_{\theta^*(\varepsilon)}(\varepsilon)] \leq KK'\mu^2(\varepsilon).
\end{aligned}$$

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

$$\forall \alpha \in (0,1)\ \exists \tilde{K} < \infty \text{ such that } \forall \varepsilon > 0\ \overline{\mathbb{P}}\left(\frac{\sigma_{\widehat{\theta}(\varepsilon)}(\varepsilon)}{\mu(\varepsilon)} \leq \tilde{K} \right) > \alpha.$$

□

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

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

Then the importance sampling estimators $Z_{\hat{\theta}(\varepsilon)}(\varepsilon)$ 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 [1] that, if we note $\hat{d}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i 1_{[X_i\leq\varepsilon]}$, the law of*

$$\frac{\sqrt{n_1}\left(\hat{\theta}_{n_1} - \frac{\mathbb{P}[X\leq\varepsilon]}{\mathbb{E}[X1_{[X\leq\varepsilon]}]}\right)}{\sigma/\hat{d}_{n_1}}$$

converges when $n_1 \to \infty$ to a Normal law with mean 0 and variance 1, where

$$\begin{aligned}
\sigma^2 &= \sigma^2[1_{[X\leq\varepsilon]}] - 2\frac{\mathbb{P}[X\leq\varepsilon]}{\mathbb{E}[X1_{[X\leq\varepsilon]}]}Cov[1_{[X\leq\varepsilon]}, X1_{[X\leq\varepsilon]}] + \left(\frac{\mathbb{P}[X\leq\varepsilon]}{\mathbb{E}[X1_{[X\leq\varepsilon]}]}\right)^2\sigma^2[X1_{[X\leq\varepsilon]}] \\
&= \mathbb{P}[X\leq\varepsilon](1-\mathbb{P}[X\leq\varepsilon]) - 2\frac{\mathbb{P}[X\leq\varepsilon]}{\mathbb{E}[X1_{[X\leq\varepsilon]}]}\mathbb{E}[X1[X_{\leq}\varepsilon]](1-\mathbb{P}[X\leq\varepsilon]) \\
&\quad + \left(\frac{\mathbb{P}[X\leq\varepsilon]}{\mathbb{E}[X1_{[X\leq\varepsilon]}]}\right)^2(\mathbb{E}[X^2 1_{[X\leq\varepsilon]}] - (\mathbb{E}[X1_{[X\leq\varepsilon]}])^2) \\
&= -e^{-\lambda\varepsilon}(1-e^{-\lambda\varepsilon}) \\
&\quad + \left(\frac{1-e^{-\lambda\varepsilon}}{(1-e^{-\lambda\varepsilon})/\lambda - \varepsilon e^{-\lambda\varepsilon}}\right)^2\left(-\varepsilon^2 e^{-\lambda\varepsilon} - 2\frac{\varepsilon e^{-\lambda\varepsilon}}{\lambda} + 2\frac{1-e^{-\lambda\varepsilon}}{\lambda^2} - ((1-e^{-\lambda\varepsilon})/\lambda - \varepsilon e^{-\lambda\varepsilon})^2\right).
\end{aligned}$$

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,

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

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

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 $\varepsilon > 0$.

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 importance sampling estimators of these performance measures. More on this type of model and specific IS schemes can be found in [9, 10, 12, 15].

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 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 absorbtion in state b when the chain starts in state i. Specifically, consider $\mu_1(\varepsilon)$; since the birth rates tend to zero with ε, we deal with a rare-event probability.*

Suppose that for each $\varepsilon > 0$ there is besides the given 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 induces an unbiased importance sampling estimator of $\mu(\varepsilon)$ by

$$Z_{\theta(\varepsilon)}(\varepsilon) \stackrel{\text{def}}{=} L(\mathbf{X}; \overline{\theta}(\varepsilon), \theta(\varepsilon)) I\{X_{T(\mathbf{X})} \in B\}$$

with likelihood ratio

$$L(\mathbf{x}; \overline{\theta}(\varepsilon), \theta(\varepsilon)) \stackrel{\text{def}}{=} \frac{d\mathbb{P}_{\overline{\theta}(\varepsilon)}(\mathbf{x})}{d\mathbb{P}_{\theta(\varepsilon)}(\mathbf{x})} = \prod_{n=0}^{T(\mathbf{x})-1} \frac{\overline{\theta}(\varepsilon)_{x_n,x_{n+1}}}{\theta(\varepsilon)_{x_n,x_{x+1}}}, \quad \mathbf{x} \in \Omega_X.$$

Denote its standard deviation by

$$\sigma_{\theta(\varepsilon)}(\varepsilon) \stackrel{\text{def}}{=} \sqrt{\mathbb{V}ar_{\theta(\varepsilon)}[Z_{\theta(\varepsilon)}(\varepsilon)]}.$$

Recall from Section 3 the concepts of bounded relative error and probabilistic bounded relative error. Furthermore, recall that the zero-variance probability measure is induced by taking 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 [9] that a matrix $\theta(\varepsilon)^{(\mathrm{ZV})}$ with $\forall i, j \in \mathscr{S}$, $i \neq j$,

$$\theta(\varepsilon)^{(\mathrm{ZV})}_{i,j} = \frac{\overline{\theta}(\varepsilon)_{i,j}\mu_j(\varepsilon)}{\mu_i(\varepsilon)}. \tag{8}$$

For instance, in the setting of Example 4, it gives

$$\theta(\varepsilon)^{(\mathrm{ZV})}_{i,i+1} = \frac{\overline{\theta}(\varepsilon)_{i,i+1}\mu_{i+1}(\varepsilon)}{\mu_i(\varepsilon)}. \tag{9}$$

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 random variable $\widehat{\theta}(\varepsilon) : \overline{\Omega} \to \Theta$.

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

$$\overline{\mathbb{P}}\left(\sup_{i,j \in \mathscr{S}, \theta^{(\mathrm{ZV})}(\varepsilon)_{ij} \neq 0} k \leq \frac{\theta^{(\mathrm{ZV})}(\varepsilon)_{ij}}{\widehat{\theta}(\varepsilon)_{ij}} \leq K \right) > \alpha. \tag{10}$$

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

Proof. Choose $\varepsilon > 0$ and a transition matrix $\theta = \theta(\varepsilon)$ realization of $\widehat{\theta}(\varepsilon)$ such that $\sup_{i,j \in \mathscr{S}, \theta^{(\mathrm{ZV})}(\varepsilon)_{ij} \neq 0} k \leq \frac{\theta^{(\mathrm{ZV})}(\varepsilon)_{ij}}{\theta} \leq K$. 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 $\sup_{i,j \in \mathscr{S}} \frac{\theta^{(\mathrm{ZV})}(\varepsilon)_{ij}}{\theta} \leq K$ 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

$$\frac{I\{X_{T(\mathbf{X})} \in B\}\, L^2(\mathbf{X}; \overline{\theta}(\varepsilon); \theta^{(\mathrm{ZV})}(\varepsilon))}{\mu^2(\varepsilon)} = 1 \quad \text{a.s.},$$

and then bound it using that θ in the set given in (10):

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

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 we assumed a finite statespace $\mathscr{S}$, 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 transiton 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 [9]). Hence, $T(\mathbf{X})/m$ is stochastically smaller than a geometric random variable 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

$$\mathbb{E}_\theta[K^{2T(\mathbf{X})}] = \mathbb{E}_\theta[(K^{2m})^{T(\mathbf{X})/m}] \leq \mathbb{E}_\theta[(K^{2m})^Y] = \frac{p_0(\varepsilon)}{1 - (1 - p_0(\varepsilon))K^{2m}}, \tag{11}$$

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 [9], thanks to our assumptions, we can show that $p_0(\varepsilon) \to 1$ as $\varepsilon \to 0$, hence the result. This basically comes from for any path $\mathbf{x}$ are 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 $p_0(\varepsilon) \to 1$. □

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

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

Let

$$\rho(\varepsilon) \stackrel{\text{def}}{=} \frac{\overline{\theta}(\varepsilon)_{i,i-1}}{\overline{\theta}(\varepsilon)_{i,i+1}} = \frac{1-\varepsilon}{\varepsilon}.$$

(Note that this is not the traditional 'load $\rho = \lambda/\mu$'.) Then it is well-known that

$$\mu_i(\varepsilon) = \frac{\rho(\varepsilon)^i - 1}{\rho(\varepsilon)^b - 1}, \ i = 0, \ldots, b. \tag{12}$$

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

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

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 associated importance sampling estimators show BRE in the context of estimating large population sizes, i.e., in the case of $b \to \infty$ while keeping the transition probabilities $\overline{\theta}(\varepsilon) \equiv \overline{\theta}$ not dependent on the rarity parameter, see for instance Section 2.3.3 and Section 5.3.1 in [13]. The same holds true in our study of reliability probabilities for systems with constant state space.

Proposition 4. *Suppose that we apply importance sampling simulation with these $\theta^*(\varepsilon)$ matrices to estimate $\mu_1(\varepsilon)$. Then the associated importance sampling 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 importance sampling transition probabilities $\theta^*(\varepsilon)_{ij}$, this cycle product equals 1. Thus

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

Note that similar to (12)

$$\mathbb{P}_{\theta^*(\varepsilon)}(X_{T(\mathbf{X})} = b) = \mathbb{P}_{\theta^*(\varepsilon)}(X_{T(\mathbf{X})} = b | X_0 = 1) = \frac{1-\rho(\varepsilon)^{-1}}{1-\rho(\varepsilon)^{-b}},$$

where $\rho(\varepsilon)^{-1} = \varepsilon/(1-\varepsilon) \to 0$ as $\varepsilon \to 0$. Hence,

$$\begin{aligned}
\frac{\mathbb{E}_{\theta^*(\varepsilon)}[Z^2_{\theta^*(\varepsilon)}(\varepsilon)]}{\mu_1(\varepsilon)^2} &= \rho(\varepsilon)^{-2(b-1)} \frac{1-\rho(\varepsilon)^{-1}}{1-\rho(\varepsilon)^{-b}} \frac{(\rho(\varepsilon)^b - 1)^2}{(\rho(\varepsilon)-1)^2} \\
&= \cdots (\text{calculus}) \cdots = \frac{1-\rho(\varepsilon)^{-b}}{1-\rho(\varepsilon)^{-1}} \leq 2(1-\rho(\varepsilon)^{-b}) < 2,
\end{aligned}$$

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

6.2 The zero-variance probability measure

Using (9) and (12) we find

$$\theta(\varepsilon)^{(\mathrm{ZV})}_{i,i+1} = \varepsilon \frac{\rho(\varepsilon)^{i+1} - 1}{\rho(\varepsilon)^i - 1}, \;\; \theta(\varepsilon)^{(\mathrm{ZV})}_{i,i-1} = (1-\varepsilon) \frac{\rho(\varepsilon)^{i-1} - 1}{\rho(\varepsilon)^i - 1},$$

for $k = 1, 2, \ldots$. States 0 and b are again absorbing. Note that $\theta(\varepsilon)^{(\mathrm{ZV})}_{1,0} = 0$, thus

$$\mathbb{P}_{\theta(\varepsilon)^{(\mathrm{ZV})}}(X_{T(\mathbf{X})} = b) = 1.$$

6.3 Cross-entropy estimators

The cross-entropy method is an adaptive simulation technique that estimates the zero-variance transition probabilities by minimizing the Kullback-Leibler divergence between the zero-variance probability measure and the importance sampling change of measure [3] Apply the cross-entropy method with a sample of size R, then it results in estimates

$$\widehat{\theta}(\varepsilon)_{i,i+1}[R] = \frac{\sum_{r=1}^{R} I\{\mathbf{X}^{(r)}_{T(\mathbf{X}^{(r)})} = b\} N_{i,i+1}(\mathbf{X}^{(r)})}{\sum_{r=1}^{R} I\{\mathbf{X}^{(r)}_{T(\mathbf{X}^{(r)})} = b\} \left(N_{i,i+1}(\mathbf{X}^{(r)}) + N_{i,i-1}(\mathbf{X}^{(r)})\right)}, \tag{13}$$

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 $\mathbf{X}$. Note that the simulations have been executed under the original measure $\mathbb{P}_{\overline{\theta}(\varepsilon)}$. Alternatively, we might have executed these simulations under a change of measure $\mathbb{P}_{\theta(\varepsilon)}$. In that case the numerator and denominator in (13) include also the likelihood ratio $L(\mathbf{X}^{(r)}; \overline{\theta}(\varepsilon), \theta(\varepsilon))$. In both cases, when we let $R \to \infty$, we obtain

$$\begin{aligned} \lim_{R\to\infty} \widehat{\theta}(\varepsilon)_{i,i+1}[R] &= \frac{\mathbb{E}_{\overline{\theta}(\varepsilon)}[I\{\mathbf{X}_{T(\mathbf{X})} = b\} N_{i,i+1}(\mathbf{X})]}{\mathbb{E}_{\overline{\theta}(\varepsilon)}[I\{\mathbf{X}_{T(\mathbf{X})} = b\} (N_{i,i+1}(\mathbf{X}) + N_{i,i-1}(\mathbf{X})]} \\ &= \theta(\varepsilon)^{(\mathrm{ZV})}_{i,i+1} \qquad \text{a.s..} \end{aligned} \tag{14}$$

The last equality in (14) has been shown in [11].

The estimator (13) satisfies the central limit theorem, see for instance [1, page 107]:

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

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 R large enough

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

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

Proof. Apply Proposition 3 while using (14) and the constant size of state space for all ε. □

6.4 Numerical/simulation results

Data: $b = 20$, $\varepsilon = k^{-\delta}$ with $\delta = 1.0$, and $k = 5, 6, \ldots, 100$. The rare event probabilities $\mu_1(\varepsilon)$ run from 2.7285e-12 for $k = 5$ to 1.1982e-38 for $k = 100$. We have applied the cross-entropy method as follows: in the first iteration we generated $R = 100000$ sample paths using the uniform transition probabilities; that is

$$\theta(\varepsilon)^{(0)}_{i,i+1} = \theta(\varepsilon)^{(0)}_{i,i-1} = 0.5.$$

We obtained estimates $\theta(\varepsilon)_{ij}^{(1)}$ by adopting the rule (13) to cope with this change of measure:

$$\theta(\varepsilon)_{i,i+1}^{(1)} = \frac{\sum_{r=1}^{R} L(\mathbf{X}^{(r)};\overline{\theta}(\varepsilon);\theta(\varepsilon)^{(0)}) I\{\mathbf{X}^{(r)}_{T(\mathbf{X}^{(r)})} = b\} N_{i,i+1}(\mathbf{X}^{(r)})}{\sum_{r=1}^{R} L(\mathbf{X}^{(r)};\overline{\theta}(\varepsilon);\theta(\varepsilon)^{(0)}) I\{\mathbf{X}^{(r)}_{T(\mathbf{X}^{(r)})} = b\} \left(N_{i,i+1}(\mathbf{X}^{(r)}) + N_{i,i-1}(\mathbf{X}^{(r)})\right)}. \quad (15)$$

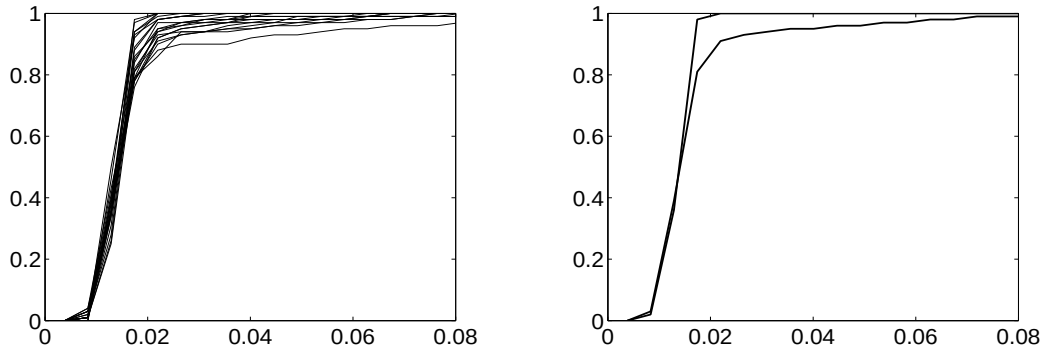
Then, in the second ($k = 2$) and third iteration ($k = 3$) we generated again $R = 100000$ sample paths using the $\theta(\varepsilon)^{(k-1)}$ transition probabilities, to get estimates $\theta(\varepsilon)^{(k)}$ by the (15) rule. With the $\theta(\varepsilon)^{(\mathrm{CE})} = \theta(\varepsilon)^{(3)}$ transition probabilities we generated $n = 1000$ sample paths for estimating the rare event probability $\mu_1(\varepsilon)$. This gave us a relative error $\sigma_{\theta(\varepsilon)^{(\mathrm{CE})}}(\varepsilon)/\mu_1(\varepsilon)$ of the rare event estimator. This whole procedure is repeated $M = 100$ times. In this way we have generated M observations of a cross-entropy estimator $\widehat{\theta}(\varepsilon)$ and M observations of the relative errors of the associated importance sampling estimators. From these M relative error data we have constructed the empirical distribution function

$$F^{(\mathrm{emp})}_{\widehat{\theta}(\varepsilon)}(x;\varepsilon) = \frac{1}{M}\sum_{m=1}^{M} I\{\sigma_{\theta(\varepsilon)^{(\mathrm{CE})}_m}(\varepsilon)/\mu_1(\varepsilon) \le x\}$$

as an approximation to the CDF $F_{\widehat{\theta}(\varepsilon)}(x;\varepsilon)$:

$$F_{\widehat{\theta}(\varepsilon)}(x;\varepsilon) \stackrel{\mathrm{def}}{=} \overline{\mathbb{P}}_{\widehat{\theta}(\varepsilon)}\left(\left\{\theta \in \Theta : \frac{\sigma_\theta(\varepsilon)}{\mu(\varepsilon)} \le x\right\}\right), \quad x \ge 0.$$

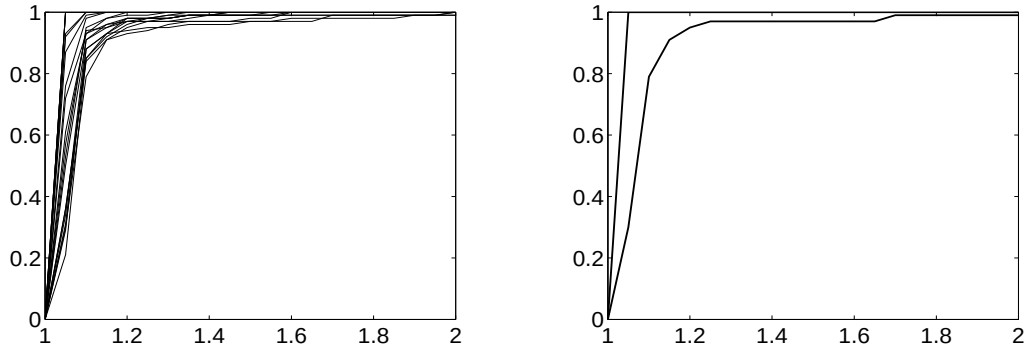
The figure on the left shows the graphs of these empirical distribution functions for $\varepsilon = k^{-1}$ with $k = 5, 10, \ldots, 100$; the figure on the right for $k = 5$ and $k = 100$ only. The figures illustrate the probabilistic bounded relative error property of the cross-entropy method.



As an illustration of Proposition 3, the following figures show the empirical distribution functions of the maximal ratio of transition probabilities

$$\sup_{i,j:\theta(\varepsilon)^{(\mathrm{CE})}_{ij} \neq 0} \frac{\theta^{(\mathrm{ZV})}(\varepsilon)_{ij}}{\theta(\varepsilon)^{(\mathrm{CE})}_{ij}},$$

for $\varepsilon = k^{-1}$ with $k = 5, 10, \ldots, 100$ (figure left) and for $k = 5$ and $k = 100$ (figure right).



7 Conclusions

This paper introduced a new complexity concept for rare-event simulation: probabilistic bounded relative error. The motivation was to formalize a bounded relative error property of a rare-event estimator when its distribution is obtained by a random procedure, such as the cross-entropy method. The concept is supported by a simulation study of a simple queue.

References

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

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

[3] P. T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross entropy method. *Annals of Operations Research*, 134:19–67, 2005.

[4] P.W. Glynn, G. Rubino, and B. Tuffin. Robustness properties and confidence interval reliability issues. In G. Rubino and B. Tuffin, editors, *Rare Event Simulation using Monte Carlo Methods*, pages 63–84. Wiley, 2009. Chapter 4.

[5] P. Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation*, 5(1):43–85, 1995.

[6] S. Juneja and P. Shahabuddin. Rare event simulation techniques: An introduction and recent advances. In S. G. Henderson and B. L. Nelson, editors, *Simulation*, Handbooks in Operations Research and Management Science, pages 291–350. Elsevier, Amsterdam, The Netherlands, 2006. Chapter 11.

[7] P. L'Ecuyer, J. H. Blanchet, B. Tuffin, and P. W. Glynn. Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation*, 20(1):Article 6, 2010.

[8] P. L'Ecuyer, M. Mandjes, and B. Tuffin. Importance sampling and rare event simulation. In G. Rubino and B. Tuffin, editors, *Rare Event Simulation Using Monte Carlo Methods*, pages 17–38. Wiley, 2009. Chapter 2.

[9] P. L'Ecuyer and B. Tuffin. Approximating zero-variance importance sampling in a reliability setting. *Annals of Operations Research*, 2012. To appear.

[10] M. K. Nakayama. General conditions for bounded relative error in simulations of highly reliable Markovian systems. *Advances in Applied Probability*, 28:687–727, 1996.

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

[12] G. Rubino and B. Tuffin. Markovian models for dependability analysis. In G. Rubino and B. Tuffin, editors, *Rare Event Simulation using Monte Carlo Methods*, pages 125–144. John Wiley & Sons, 2009. Chapter 6.

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

[14] R. Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 2:127–190, 1999.

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