ESAIM: PROCEEDINGS, January 2014, Vol. 44, p. 239-259
SMAI Groupe MAS – Journées MAS 2012 – Sessions thématiques

SOME RECENT RESULTS IN RARE EVENT ESTIMATION

Virgile Caron[1], Arnaud Guyader[2], Miguel Munoz Zuniga[3] and Bruno Tuffin[4]

Abstract. This article presents several state-of-the-art Monte Carlo methods for simulating and estimating rare events. A rare event occurs with a very small probability, but its occurrence is important enough to justify an accurate study. Rare event simulation calls for specific techniques to speed up standard Monte Carlo sampling, which requires unacceptably large sample sizes to observe the event a sufficient number of times. Among these variance reduction methods, the most prominent ones are *Importance Sampling* (IS) and *Multilevel Splitting*, also known as Subset Simulation. This paper offers some recent results on both aspects, motivated by theoretical issues as well as by applied problems.

Résumé. Cet article propose un état de l'art de plusieurs méthodes Monte Carlo pour l'estimation d'événements rares. Un événement rare est par définition un événement de probabilité très faible, mais d'importance pratique cruciale, ce qui justifie une étude précise. La méthode Monte Carlo classique s'avérant prohibitivement coûteuse, il importe d'appliquer des techniques spécifiques pour leur estimation. Celles-ci se divisent en deux grandes catégories : échantillonnage préférentiel d'un côté, méthodes multi-niveaux de l'autre. Nous présentons ici quelques résultats récents dans ces domaines, motivés par des considérations tant pratiques que théoriques.

Introduction

Reliability analysis is a current important topic, aiming at verifying if energy, telecommunication or transportation systems (for example) satisfy the societal needs. A similar type of analysis is required in many other important fields, including finance, physics, biology, etc. But the constantly evolving and increasingly sophisticated systems, due to technology changes, induce an increasing complexity of the underlying mathematical models. As a consequence, the traditional analysis methodologies become non-applicable to solve the considered models, due to too stringent assumptions that would not allow to represent reality sufficiently well and/or because they require a too long computational time to give a meaningful result. Monte Carlo simulation then becomes a relevant tool, just requiring to generate independent copies of the system, from which an estimator and confidence intervals can easily be derived [1, 45].

A difficulty is nevertheless that many problems usually mean investigating the probability of rare events. In that case, standard Monte Carlo requires in average a very long time to observe the rare event only once, for instance 10^9 independent copies on average for an event of probability 10^{-9}, a typical target. Specific techniques have therefore to be implemented. Among the most prominent ones are Importance Sampling and Multilevel Splitting (sometimes called Subset Simulation).

[1] LSTA, UPMC-Paris VI, France; e-mail: virgile.caron@upmc.fr
[2] Université Rennes 2, IRMAR, Inria Rennes, Campus de Villejean, 35043 Rennes Cedex, France; e-mail: arnaud.guyader@uhb.fr
[3] Institut de Radioprotection et de Sûreté Nucléaire, Fontenay-aux-roses, France; e-mail: mmunozzu@hotmail.fr
[4] Inria Rennes, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France; e-mail: bruno.tuffin@inria.fr

© EDP Sciences, SMAI 2013

Article published online by EDP Sciences and available at http://www.esaim-proc.org or http://dx.doi.org/10.1051/proc/201444015

Importance Sampling (IS) has emerged in the Monte Carlo literature as a general and powerful tool to reduce the variance of an estimator, which, in the case of rare event estimation, also means increasing the occurrence of the rare event. The generic idea of IS is to change the probability laws of the system under study to sample more frequently the events that are more "important" for the simulation. Of course, using a new distribution results in a biased estimator if no correction is applied. Therefore the simulation output needs to be translated in terms of the original measure. This is done by multiplication with a so-called likelihood ratio. IS has received substantial theoretical attention. We refer to Robert and Casella [38] for a discussion on IS from a general point of view, and to Bucklew [9], L'Ecuyer, Mandjes and Tuffin [27], and Rubino and Tuffin [39], for the application of IS in the context of rare event estimation. Recent contributions to IS are presented in Sections 2 and 3 of the present article.

Multilevel splitting (also called splitting, importance splitting, or subset simulation) is an alternative technique for accelerating the rate of occurrence of the rare event of interest. Here, we do not change the probability laws driving the model. Instead, we use a selection mechanism to favor the trajectories deemed likely to lead to the rare event. The main idea is to decompose the paths to the rare event of interest into shorter subpaths whose probability is not so small, encourage the realizations that take these subpaths (leading to the event of interest) by giving them a chance to reproduce (a bit like in selective evolution), and discourage the realizations that go in the wrong direction by killing them with some positive probability. In the end, an unbiased estimator can be recovered by multiplying the contribution of each trajectory by the appropriate weight. Multilevel splitting was introduced by Kahn and Harris [24] in the early fifties in the field of particle transmission. We refer the reader to the tutorial of L'Ecuyer, Le Gland, Lezaud and Tuffin [26] for an in-depth review and a thorough list of references, and to Section 1, for more details.

This paper is divided into four independent sections. Section 1, authored by Arnaud Guyader via a collaboration with Nick Hengartner and Eric Matzner-Løber, proposes a new multilevel splitting type algorithm when the rare event is described by the tail distribution of a function of a random vector. In Section 2, Bruno Tuffin outlines an approximation of the optimal IS estimator in a general Markovian context (joint works with Pierre L'Ecuyer, Gerardo Rubino and Samira Saggadi). Section 3, by Virgile Caron and based on his PhD thesis under the supervision of Michel Broniatowski, presents an approximation of the optimal IS density when dealing with a long random walk conditioned to an average of its summands. Finally, Miguel Munoz Zuniga describes a new algorithm for rare event estimation in a challenging context, namely for a high computational time model and a relatively large input dimension. This latter is based on works with Josselin Garnier, Emmanuel Remy and Etienne de Rocquigny.

1. Multilevel splitting in a static context

Suppose X is a random vector in $\mathbb{R}^d$ with law μ that we can simulate, and Φ is a mapping from $\mathbb{R}^d$ to $\mathbb{R}$, also called a score function. Because of the complexity of the underlying process, we view Φ as black box, that is, we do not have an analytic expression for Φ but we can readily evaluate $\Phi(X)$ for any given instance X. Then given a threshold q which lies far out in the right hand tail of the distribution of $\Phi(X)$, we seek to estimate the very low probability $p = \mathbb{P}(\Phi(X) > q)$.

A Standard Monte Carlo that uses an i.i.d. N-sample $X_1, \ldots, X_N$ to estimate p by the fraction $\hat{p}_{mc} = \#\{i : \Phi(X_i) > q\}/N$ is not practical when p is very small. Indeed, in order to obtain a reasonable precision of the estimate given by the relative variance $\mathbb{V}(\hat{p}_{mc})/p^2$, which is equal to $(1-p)/(Np)$, one needs to select a sample size N of order p^{-1}. For instance, a random sample of one billion observations is needed to estimate a target probability of 10^{-9}.

Importance sampling, which draws samples according to an auxiliary law π and weights each observation $X = x$ by $w(x) = d\mu(x)/d\pi(x)$ can decrease the variance of the estimated probability which in turn greatly reduces the need for such large sample sizes. Unfortunately, when Φ is a black box, these weights cannot be computed, and hence importance sampling is not available to us.

Multilevel splitting (also called splitting, importance splitting or subset simulation) was introduced by Kahn and Harris [24] and is another powerful algorithm for rare event estimation. We refer the reader to the tutorial

of L'Ecuyer, Le Gland, Lezaud and Tuffin [26] for an in-depth review and a detailed list of references on splitting in the specific case of stochastic processes. Indeed, it turns out that since the pioneering work of Kahn and Harris in the early fifties and until the 2000's, most of the applications of multilevel splitting methods were dedicated to the estimation of rare events in a dynamic context, typically for Markovian processes. In the present contribution, we will rather focus on splitting methods in a static context, which means the estimation of a rare event for a random vector, not for a random process.

The basic idea of multilevel splitting, adapted to our static context, is to fix a set of increasing levels $-\infty = L_0 < L_1 < L_2 < \cdots < L_{n_0} = q$, and to decompose the tail probability as follows

$$\mathbb{P}(\Phi(X) > q) = \prod_{m=0}^{n_0-1} \mathbb{P}(\Phi(X) > L_{m+1} | \Phi(X) > L_m).$$

Each conditional probability $p_m = \mathbb{P}(\Phi(X) > L_{m+1} | \Phi(X) > L_m)$ is then estimated separately. Two practical issues associated with the implementation of multilevel splitting are the need for computationally efficient algorithms for estimating the successive conditional probabilities, and the optimal selection of the sequence of levels.

To our knowledge, the first instance in which static rare event simulation using multilevel splitting was proposed is a paper by Au and Beck [2]. But these authors call it "Subset Simulation" and do not make any connection with multilevel splitting, which is why people in the rare event community do not systematically mention this work afterwards. The next work where a reversible transition kernel was introduced to deal with such static rare events is due to Del Moral, Doucet and Jasra [16].

The paper of Cérou, Del Moral, Furon and Guyader [13] proposes to adaptively select the levels using the $(1-p_0)$-quantiles of the conditional distributions of $\Phi(X)$ given that $\Phi(X) > L_m$, where p_0 is held fixed (for example $p_0 = 3/4$). Considering the following decomposition of the probability p

$$p = r_0\, p_0^{n_0} \quad \text{where} \quad n_0 = \lfloor \log p / \log p_0 \rfloor \quad \text{and} \quad r_0 = p\, p_0^{-n_0} \in (p_0, 1]$$

their tail probability estimate writes $\tilde{p} = \tilde{r}_0\, p_0^{\tilde{n}_0}$ and it turns out that when the number of particles N tends to infinity, the best precision that one can achieve by applying this technique is

$$\mathbb{V}(\tilde{p}) \underset{N\to\infty}{\sim} \frac{p^2}{N} \left(\frac{(1-p_0)\cdot \log p}{p_0 \cdot \log p_0} \right). \tag{1}$$

It is noteworthy that a very similar approach has been independently proposed by Rubinstein [40] and Botev and Kroese [6] in the context of combinatorial optimization, counting and sampling, demonstrating the performance of this algorithm via an extensive simulation study.

Interestingly, since the mapping $\psi : p_0 \mapsto (1-p_0)/(-p_0 \log p_0)$ is nonincreasing on $(0,1)$, one can deduce from equation (1) that the larger p_0, the lower the variance, with

$$\lim_{p_0 \to 1^-} N \times \mathbb{V}(\tilde{p}) = -p^2 \log p. \tag{2}$$

Hence the idea to choose p_0 as large as possible. However, with an adaptive method, the largest possible value is clearly $p_0 = 1 - 1/N$. This is the main idea of the so-called Last Particle Algorithm presented hereafter. We refer the reader to the paper by Guyader, Hengartner and Matzner-Løber [22] for details and proofs, and to Cérou, Guyader, Lelièvre and Pommier [14] for the application of this algorithm in the context of molecular dynamics. Before proceeding, let us mention that this algorithm bears a resemblance to the "Nested Sampling" approach which was proposed by Skilling in the context of sampling from general distributions and estimating their normalising constants [15, 44].

1.1. The Last Particle Algorithm

Consider the following **idealized** algorithm:

- Start with an i.i.d. sample $(X_1, X_2, \ldots, X_N)$ from μ and initialize $L_0 = -\infty$ and
$$X_1^1 = X_1, \ldots, X_N^1 = X_N.$$
- For $m = 1, 2, \ldots$, set
$$L_m = \min(\Phi(X_1^m), \ldots, \Phi(X_N^m)),$$
and define for all $i = 1, 2, \ldots, N$:
$$X_i^{m+1} = \begin{cases} X_i^m & \text{if } \Phi(X_i^m) > L_m \\ X^\star \sim \mathcal{L}(X|\Phi(X) > L_m) & \text{if } \Phi(X_i^m) = L_m, \end{cases} \tag{3}$$
where $X^\star$ is independent of $\{X_1^m, \ldots, X_N^m\}$.
- Stopping rules:
  (a) To estimate a tail probability p given a quantile q, continue until $m = M$ where $M = \max\{m : L_m \leq q\}$ and set $\hat{p} = \left(1 - \frac{1}{N}\right)^M$. Note that the number M of iterations is a random variable.
  (b) To estimate a quantile q given a tail probability p, continue until iteration $m = \left\lceil \frac{\log(p)}{\log(1-N^{-1})} \right\rceil$, and set $\hat{q} = L_m$. Note that this time, the number m of iterations is deterministic.

Simulating exactly according to $\mathcal{L}(X|\Phi(X) > L_m)$ at step (3) is impossible in general and we propose to do so approximately using Markov Chain Monte Carlo techniques. However, for the theoretical analysis of Section 1.2, we will consider only the case where that simulation could be done perfectly, and this is the reason why we call it the **Idealized** Last Particle Algorithm.

For the practical implementation, to draw $X^\star$ at step (3), we run a Markov chain with a suitable μ-symmetric and one-step μ-irreducible kernel K. That is: K satisfies the detailed balance property with μ; and from any initial point x, the Radon-Nikodym derivative $dK(x, dx')/d\mu(dx')$ is strictly positive. Either, one knows such a kernel K or otherwise could use a Metropolis-Hasting kernel K based on a one-step μ-irreducible instrumental kernel $Q(x, dx')$ (see for example [38]). This latter is possible when the law μ is known up to a normalizing constant, that is $\mu(dx) \propto f(x)dx$, which is usually the case in Bayesian statistics (among others).

Example: Let us suppose that X has a standard Gaussian distribution on $\mathbb{R}$. Then let us present two ways to get such a transition kernel K:

(1) Direct construction: fix $\sigma > 0$ and denote K the transition kernel defined by
$$K(x, dx') = \left(\frac{1+\sigma^2}{2\pi\sigma^2}\right)^{\frac{d}{2}} \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\left\|x' - \frac{x}{\sqrt{1+\sigma^2}}\right\|^2\right)\lambda(dx'),$$
where λ stands for Lebesgue's measure on $\mathbb{R}^d$. Denoting W a Gaussian standard variable on $\mathbb{R}^d$ independent of X, the transition $X \rightsquigarrow X'$ proposed by K is thus $X' = (X + \sigma W)/\sqrt{1+\sigma^2}$.

(2) Metropolis-Hastings kernel: fix $\sigma > 0$ and denote Q the transition kernel defined by
$$Q(x, dx') = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{d}{2}} \exp\left(-\frac{\|x' - x\|^2}{2\sigma^2}\right)\lambda(dx').$$
Denoting W a Gaussian standard variable on $\mathbb{R}^d$ independent of X, the transition $X \rightsquigarrow X'$ proposed by Q is this time $X' = X + \sigma W$. Then, starting from Q, the transition kernel K constructed by

Metropolis-Hastings is μ-symmetric and one-step μ-irreducible.

With this kernel K at disposal, let us now detail how one can approximately satisfy step (3) of the idealized algorithm. For this, we fix an integer T and we assume that L_m is reached at a unique point of the sample, which is always the case when $\Phi(X)$ has a continuous cdf. To simplify the writings, let us suppose without loss of generality that $L_m = \Phi(X_1^m)$. Then it suffices to apply the following procedure to get the **Practical** Last Particle Algorithm:

Pick an integer R randomly between 2 and N
Let $X_1^{m+1} = X_R^m$
For $t = 1 : T$
  From X_1^{m+1}, draw a new particle $X^\star \sim K(X_1^{m+1}, .)$
  If $\Phi(X^\star) > L_m$, then let $X_1^{m+1} = X^\star$
Endfor

If $T = +\infty$, then the Practical and the Idealized Last Particle Algorithms coincide. But this is of course unrealistic, and, as for any Markov Chain Monte Carlo method, the tuning of T depends on the choice of the kernel K. To conclude with these practical considerations, let us emphasize that the choice of the kernel K is crucial in order to make this method efficient.

1.2. **Properties of the Idealized Last Particle Algorithm**

In what follows, we consider the case of the Idealized Last Particle Algorithm, which means that step (3) of the previous algorithm is perfectly matched at each iteration. Again, even if this is somehow unrealistic, this gives us some insights into the properties of the proposed algorithm. Besides this very strong assumption, we only suppose that the random variable $\Phi(X)$ has a continuous cdf. This ensures that there are no ties for the "last particle" at each iteration.

1.2.1. *Estimation of a low probability*

Considering a quantile q such that $\mathbb{P}(\Phi(X) > q) = p$, we can prove that the random number of steps M specified by stopping rule (a) of the Idealized Last Particle Algorithm has a Poisson distribution, namely that $M \sim \mathcal{P}(-N \log p)$.

This being noted, it is readily seen that the estimator $\hat{p} = (1 - 1/N)^M$ is a discrete random variable with distribution

$$\mathbb{P}\left[\hat{p} = \left(1 - \frac{1}{N}\right)^m\right] = \frac{p^N(-N \log p)^m}{m!}, \quad m = 0, 1, 2, \ldots.$$

It follows that $\hat{p}$ is an unbiased estimator of p with variance:

$$\mathbb{V}(\hat{p}) = p^2 \left(p^{-\frac{1}{N}} - 1\right). \tag{4}$$

Notice that in practice, the number N of particles is typically larger than 100, so that $\mathbb{V}(\hat{p}) \approx -p^2 \log p/N$, which is according to (2) the best achievable precision when applying a multilevel technique.

From the law of $\hat{p}$, one can deduce some confidence intervals as well. Let α be a fixed number between 0 and 1 and denote by $Z_{1-\alpha/2}$ the quantile of order $1 - \alpha/2$ of the standard Gaussian distribution. Let us denote

$$\hat{p}_\pm = \hat{p} \exp\left(\pm \frac{Z_{1-\alpha/2}}{\sqrt{N}} \sqrt{-\log \hat{p} + \frac{Z_{1-\alpha/2}^2}{4N}} - \frac{Z_{1-\alpha/2}^2}{2N}\right),$$

then $I_{1-\alpha}(p) = [\hat{p}_-, \hat{p}_+]$ is a $100(1 - \alpha)\%$ confidence interval for p.

Comparing our estimator with the one obtained through Standard Monte Carlo is instructive. Recall that the Standard Monte Carlo estimate $\hat{p}_{mc}$ for the tail probability has relative variance $\frac{\mathbb{V}(\hat{p}_{mc})}{p^2} = \frac{1-p}{Np}$. Comparing the latter with the relative variance of our estimator $\hat{p}$ as specified by (4) reveals that, for the same precision, Standard Monte Carlo requires about $(-p\log p)^{-1}$ more particles than the method presented here.

However, considering that the simulation of $\Phi(X)$ has an algorithmic cost equal to 1, the Standard Monte Carlo estimator has a lower complexity, namely $C_{mc} = N$, than our algorithm whose expected complexity value is $C(\hat{p}) = -kN\log N\log p$. Indeed, it requires:

- A sorting of the initial sample, whose cost is (in expectation) in $\mathcal{O}(N\log N)$ via a quicksort algorithm;
- Around $-N\log p$ steps, whose cost is decomposed in:
  - T proposed kernel transitions,
  - the dichotomic search and the insertion of the new particle at the right place in the ordered sample, whose cost is in $\mathcal{O}(\log N)$ via a min-heap algorithm.

To take into account both computational complexity and variance, Hammersley and Handscomb have proposed to define the efficiency of a Monte Carlo procedure as "inversely proportional to the product of the sampling variance and the amount of labour expended in obtaining this estimate" [23]. In this respect, $\hat{p}$ is more efficient than $\hat{p}_{mc}$ whenever $\mathbb{V}(\hat{p})\cdot C(\hat{p}) \leq \mathbb{V}(\hat{p}_{mc})\cdot C_{mc}$, that is

$$-\frac{p^2\log p}{N}\cdot(-kN\log N\log p) \leq \frac{p(1-p)}{N}\cdot N \quad \Leftrightarrow \quad k\log N \quad \leq \quad \frac{1-p}{p(\log p)^2}.$$

That inequality is satisfied when p goes to zero since the right-hand side goes then to infinity. For example, let us fix $N = 200$ and $k = 10$, then one can check numerically that the condition

$$10\log(200) \quad \leq \quad \frac{1-p}{p(\log p)^2}$$

is true as soon as $p \leq 1.0\times 10^{-4}$. The take-home message here is that our estimator is useful compare to Standard Monte Carlo only if the probability to estimate is low enough. Finally, we refer the reader to [22] for a discussion on asymptotic efficiency properties.

1.2.2. *Estimation of an extreme quantile*

Consider this time the problem of estimating the quantile q for a given p such that $\mathbb{P}(\Phi(X) > q) = p$. Using stopping rule (b) of the Idealized Last Particle Algorithm, a natural estimator for the quantile q is $\hat{q} = L_m$, where $m = \left\lceil \frac{\log(p)}{\log(1-N^{-1})} \right\rceil$.

Given sufficient smoothness of the distribution at the quantile q, we obtain an asymptotic normality result for our estimator. Specifically, if the cdf F of $Y = \Phi(X)$ is differentiable at point q, with density $f(q) \neq 0$, then

$$\sqrt{N}(\hat{q} - q) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{-p^2\log p}{f(q)^2}\right). \tag{5}$$

Under the same assumptions, the Standard Monte Carlo estimator defined as $\hat{q}_{mc} = Y_{(\lfloor(1-p)N\rfloor)}$, where $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the order statistics of $\Phi(X_1), \ldots, \Phi(X_N)$, satisfies the following central limit theorem (see for example [42], Theorem 7.25)

$$\sqrt{N}(\hat{q}_{mc} - q) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{p(1-p)}{f(q)^2}\right).$$

This proves again that in order to achieve the same precision in terms of variance of the estimator, Standard Monte Carlo requires about $(-p\log p)^{-1}$ more particles than the estimator proposed here. Nonetheless, to be

completely fair, one shall also take into account the respective algorithmic complexities of both estimators. The conclusion is then just the same as in Section 1.2.1.

Yet, expression (5) is of limited use to construct confidence intervals as it requires the knowledge of $f(q)$. But it turns out that exploiting a connection with Poisson processes allows us to derive non asymptotic confidence intervals for q without having to estimate the density at the quantile q. Indeed, fix $\alpha \in (0,1)$, denote by $Z_{1-\alpha/2}$ the quantile of order $1-\alpha/2$ of the standard Gaussian distribution, define

$$\begin{aligned} m_- &= \left\lfloor -N\log p - Z_{1-\alpha/2}\sqrt{-N\log p} \right\rfloor \\ m^+ &= \left\lceil -N\log p + Z_{1-\alpha/2}\sqrt{-N\log p} \right\rceil \end{aligned}$$

and consider L_{m_-}, L_{m^+} the associate levels. Then it can be established that a $100(1-\alpha)\%$ confidence interval for the quantile q is $I_{1-\alpha}(q) = [L_{m_-}, L_{m^+}]$.

Two important remarks are in order. First, the computational price to pay to obtain the confidence interval is the cost of running the algorithm until step m^+ in order to get the upper confidence bound L_{m^+}. This requires the algorithm to run around $Z_{1-\alpha/2}\sqrt{-N\log p}$ additional steps. Secondly, let us emphasize that compared to the CLT type result given by (5), the great interest of this property lies in the fact that it does not require any estimation of the probability density function f.

2. Approximation of Zero-Variance Importance Sampling Estimators and Application to Reliability Models

We recall in this section the general principles of Importance Sampling (IS) variance reduction technique, one of the two major methods to deal with rare event simulation. We describe the characterization of the (non-implementable) zero-variance IS estimator, and how it can be approached. An efficient implementation approximating the zero-variance change of measure is obtained under a general and a Markovian framework, and applied to reliability-oriented problems, outperforming the other existing methods. Specifically, the resulting estimators are shown to have a bounded relative error as the system becomes more and more reliable (in a specified way), and can even have a vanishing relative error in some cases.

2.1. **Importance Sampling**

2.1.1. *General framework*

Consider in general the computation of an expectation $\mu = \mathbb{E}[h(Y)] = \int h(y)d\mathbb{P}(y)$ by simulation, with h a function and a Y random variable obeying some probability law $\mathbb{P}$. IS basically consists in a change of variable in the integral, rewriting

$$\mu = \mathbb{E}[X] = \int h(y)d\mathbb{P}(y) = \int h(y)\frac{d\mathbb{P}(y)}{d\tilde{\mathbb{P}}(y)}d\tilde{\mathbb{P}}(y) = \tilde{\mathbb{E}}\left[h(Y)L(Y)\right],$$

with $\tilde{\mathbb{P}}$ another probability measure and $L = d\mathbb{P}/d\tilde{\mathbb{P}}$ called the *likelihood ratio*. The above equality holds under the condition $d\tilde{\mathbb{P}}(y) \neq 0$ when $h(y)d\mathbb{P}(y) \neq 0$. By this transformation, we therefore consider the random variable $h(Y)L(Y)$ instead of $h(Y)$, and the probability law $\tilde{\mathbb{P}}$ instead of $\mathbb{P}$ to generate Y, but the expectation does not change. Using this new framework, an unbiased estimator of μ is

$$\frac{1}{n}\sum_{i=1}^{n} h(Y_i)L(Y_i)$$

with $(Y_i,\, 1 \leq i \leq n)$ i.i.d. copies of Y, generated according to $\tilde{\mathbb{P}}$. How to select the probability law $\tilde{\mathbb{P}}$? The idea is to reduce the variance of the estimator with respect to the initial one, i.e., choose $\tilde{\mathbb{P}}$ so that

$$\tilde{\sigma}^2[h(Y)L(Y)] = \tilde{\mathbb{E}}[(h(Y)L(Y))^2] - \mu^2 < \sigma^2[h(Y)].$$

Note that determining a probability law leading to a variance reduction is not an easy task in general: some counter-intuitive results can be obtained [27].

But it is known that the change of measure with

$$d\tilde{\mathbb{P}} = \frac{|h(Y)|}{\mathbb{E}[|h(Y)|]} d\mathbb{P}$$

is actually optimal in the sense that it is the one minimizing the variance among all possible IS choices [1, 27]. If $h \geq 0$, it actually leads to $\tilde{\mathbb{E}}[(h(Y)L(Y))^2] = (\mathbb{E}[h(Y)])^2$, i.e., $\tilde{\sigma}^2(h(Y)L(Y)) = 0$. We therefore have the *zero-variance* change of measure, always providing as an output the exact result μ. Unfortunately, implementing the estimator based on this change of measure requires to know $\mathbb{E}[|h(Y)|]$ in the computation of the likelihood ratio, but it is what we are trying to compute. If we knew it, there would be no need to resort to simulation. All this is not bad though, the optimal form provides us with an idea on what a good IS should look like.

2.1.2. *Importance Sampling applied to Markov chains*

Consider the more specific framework of a (discrete time) Markov chain $(Y_j)_{j\geq 0}$ defined on a state space $\mathcal{S}$, with transition matrix $P = (P(y,z))_{y,z\in\mathcal{S}}$ and initial probabilities $\pi_0(y) = \mathbb{P}[Y_0 = y]$. Let $X = h(Y_0, \ldots, Y_\tau)$ function of the sample path up to a stopping time τ and define $\mu(y) = \mathbb{E}_y[X]$ the expectation of X given that we start in $Y_0 = y$.

In this context, IS replaces the probabilities of paths $\mathbb{P}[(Y_0, \ldots, Y_\tau) = (y_0, \ldots, y_\tau)] = \pi_0(y_0) \prod_{j=1}^{\tau-1} P(y_{j-1}, y_j)$, by $\tilde{\mathbb{P}}[(Y_0, \ldots, Y_\tau) = (y_0, \ldots, y_\tau)]$ such that $\tilde{\mathbb{E}}[\tau] < \infty$. For computational simplicity, it is very often considered an IS measure such that we still have a discrete time Markov chain, replacing $P(y,z)$ by $\tilde{P}(y,z)$ and $\pi_0(y)$ by $\tilde{\pi}_0(y)$. The likelihood ratio is then

$$L(Y_0, \ldots, Y_\tau) = \frac{\pi_0(Y_0)}{\tilde{\pi}_0(Y_0)} \prod_{j=1}^{\tau} \frac{P(Y_{j-1}, Y_j)}{\tilde{P}(Y_{j-1}, Y_j)}$$

and can be updated at each time step.

If we restrict ourselves to the case when $X = \sum_{j=1}^{\tau} c(Y_{j-1}, Y_j)$ is an additive and positive cost (but this can be generalized) on transitions, there is actually a Markov chain change of measure yielding a zero variance:

$$\tilde{P}(y,z) = \frac{P(y,z)(c(y,z) + \mu(z))}{\sum_w P(y,w)(c(y,w) + \mu(w))} = \frac{P(y,z)(c(y,z) + \mu(z))}{\mu(y)}. \tag{6}$$

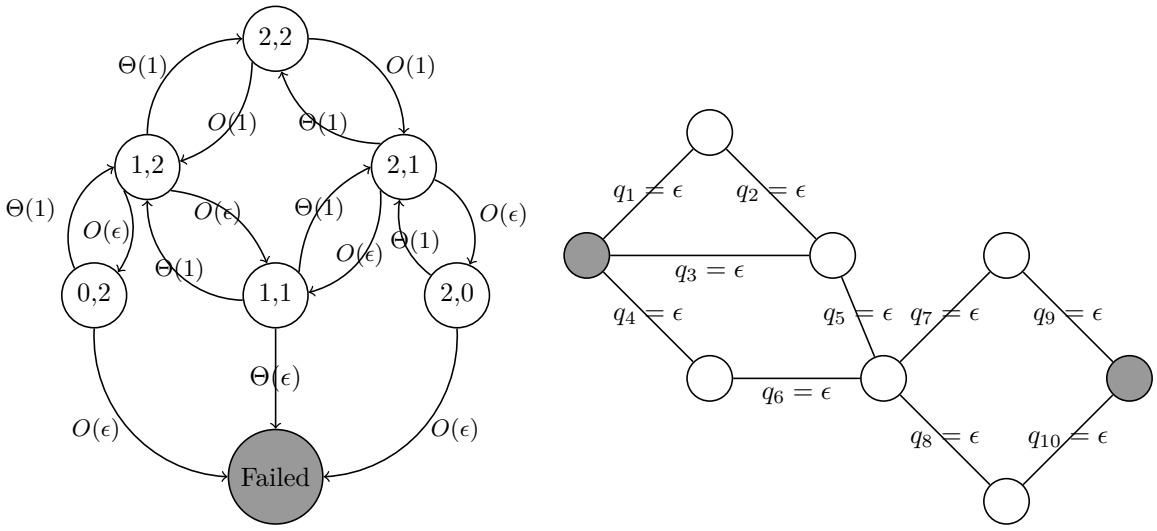
Here again though, implementing this IS scheme requires the knowledge of the quantities $\mu(y)\ \forall y \in \mathcal{S}$, i.e., what we are looking for.

2.1.3. *Zero-Variance Approximation*

Even if (6) is not implementable, it gives us some insight on the IS transition matrices of interest. One suggestion that we are going to consider in the next sections is to use a rough approximation $\hat{\mu}(y)$ of all $\mu(y)$ and to plug them into the zero-variance IS expression (6). The approximation of the zero-variance change of measure is then

$$\tilde{P}(y,z) = \frac{P(y,z)(c(y,z) + \hat{\mu}(z))}{\sum_w P(y,w)(c(y,w) + \hat{\mu}(w))}.$$

Note that there exists other ways to try to approach the zero-variance IS estimator; see also [4, 5, 17, 18].



(a) Highly Reliable Markovian System (b) Static Reliability Graph

FIGURE 1. Examples of reliability problems

2.2. **Application to Highly Reliable Markovian Systems**

The first type of reliability models we are interested in are the so-called *highly reliable Markovian systems* (HRMS) often used to represent the evolution of multicomponent systems in reliability settings [30] and whose state space increases exponentially with the number of component classes. Assume that we have a number c of types of components, and n_i components of type i. Those components may fail or be repaired (in both cases possibly in a grouped way), so that the model is usually represented by a continuous time Markov chain, but we can consider the embedded discrete time Markov chain at the instants of change of state, $Y = (Y^{(j)})_{j\geq 0}$, with $Y^{(j)} = (Y_1^{(j)}, \ldots, Y_c^{(j)})$ and where $Y_i^{(j)}$ is the number of operational components of type i at the j-th change of state. Figure 1(a) shows an example of such a problem with $c = 2$ and two components of each type. The goal is to compute the probability that the system reaches a failed state (represented in grey in Figure 1(a)) before coming back to the fully-operational one. Denote by $\mu(y)$ this probability given that we start in state y. Usually failures are rare compared with repairs, so that we can introduce a *rarity parameter* ϵ such that failures have a probability $O(\epsilon)$ (except from the fully-operational state: a transition is in this case necessarily a failure) while repair probabilities are $\Theta(1)$ (a function f is said to be $\Theta(\epsilon^d)$ if there are constants c_1 and c_2 such that $c_1\epsilon^d \leq f(\epsilon) \leq c_2\epsilon^d$ for ϵ sufficiently small). The previously designed IS methods applied to this problem were trying to increase the probability of component failures [35,43]. We rather chose to implement a zero-variance IS estimator based on Equation (6). In this context, $c(y, z) = 0$ (no cost associated to a transition), but $\mu(y) = 1$ if y is a failed state while $\mu(y) = 0$ if y is the state with all components operational. For all states y that are not failed or fully operational, we proposed in [30] to replace $\mu(y)$ by $\hat{\mu}(y)$, the probability of the most likely path from y to a failed state (which can be computed in general in a polynomial time). To get the intuition behind our approximation, remark that as $\epsilon \to 0$, $\hat{\mu}(y) = \Theta(\mu(y))$, i.e., the probability of the path is of the same order in terms of ϵ as $\mu(y)$.

It can be proved that the produced estimator satisfies bounded relative error (BRE) under mild conditions, and can even have vanishing relative error (VRE) under slightly stronger conditions [30]. BRE (also called strong efficiency) means that the standard deviation of the estimator divided by its mean is kept bounded as $\epsilon \to 0$, while for VRE it converges to 0 when $\epsilon \to 0$ [25]. In other words, the sample size required to get a

confidence interval with a predefined relative accuracy is not sensitive to the rarity of the event when BRE is satisfied; when VRE is satisfied this relative accuracy gets better with the rarity.

To briefly illustrate the efficiency of the method, consider a system made of $c = 3$ types of components, $n_i = 6$ components of type i, and such that the system is down whenever fewer than two components of any one type are operational. Assume that a component of type i fails with rate λ_i with $\lambda_1 = \varepsilon$, $\lambda_2 = 1.5\varepsilon$, and $\lambda_3 = 2\varepsilon^2$, while any failed component gets repaired with rate $\mu = 1$. The transition probabilities of the embedded discrete time Markov chain, from a given state, are then the rate of the considered transition divided by the sum of rates of transitions from this state. For $\epsilon = 10^{-2}$, the relative error is 0.79, and reducing to 0.204 for $\epsilon = 10^{-3}$. Actually, VRE can be shown to be satisfied.

2.3. **Application to Static Reliability Analysis Models**

The second class of problem is the static network reliability estimation, where we want to compute the probability that a set of nodes $\mathcal{K}$ of a graph is connected, when links fail randomly. This class of problems belongs to the NP-hard family and Monte Carlo simulation is therefore a relevant tool to provide an estimation. For example, looking at Figure 1(b), we want to investigate the probability that the set $\mathcal{K}$ of grey nodes are *not* connected (which is in most cases the rare event) given that for each link e, there is a probability $q_e = O(\epsilon)$ that the link is failed (and thus can be removed from the graph). Failures of links are considered independent. If we have m links on the graph, a (random) *configuration* is given by the vector $Y = (Y_1, \ldots, Y_m)$ with $Y_e = 1$ if link e works, 0 otherwise. The state of the system is then $\phi(Y)$, where $\phi(Y) = 1$ iff $\mathcal{K}$ not connected in configuration Y, so that

$$\mu = \mathbb{E}[\phi(Y)] = \sum_{y \in \{0,1\}^m} \phi(y)\mathbb{P}[Y = y]$$

represents the *unreliability* of the graph that we wish to compute.

The general principle if we want to apply the (Markovian) zero-variance context of Equation (6) is to sample the links one after the other, with an IS probability that *depends on the state of previously sampled links*. We thus define an order on the m links and define $\mu_k(y_1, \ldots, y_{k-1})$, with $y_i \in \{0, 1\}$, as the unreliability of the graph given the states of the links 1 to $k-1$: if $y_i = 1$ the link i is operational, otherwise it is failed. With those notations, $\mu = \mu_1()$. Under such a sequential construction, it is easy to show that sampling the i-th link as failed using IS probability

$$\tilde{q}_i = \frac{q_i \mu_{i+1}(y_1, \ldots, y_{i-1}, 0)}{q_i \mu_{i+1}(y_1, \ldots, y_{i-1}, 0) + (1 - q_i)\mu_{i+1}(y_1, \ldots, y_{m-1}, 1)} \tag{7}$$

will yield a zero variance [28].

But again, this IS scheme is not implementable, since it requires to know all the $\mu_i(y_1, \ldots, y_{i-1})$. What is proposed in [28] is to replace $\mu_{i+1}(y_1, \ldots, y_{i-1}, y_i)$ in (7) by a heuristic approximation $\hat{\mu}_{i+1}(y_1, \ldots, y_{i-1}, y_i)$ which is the the probability of the most probable *mincut* on the links $i \geq m + 1$, the state of each link $i \leq m$ being fixed to y_i. Recall that a *cut* is a set of edges such that, if we remove them, the nodes in $\mathcal{K}$ are not in the same connected component, and a *mincut* (minimal cut) is a cut that contains no other cut than itself. The intuition is, somewhat similarly to the HRMS case, that the unreliability is the probability of union of all cuts, the most crucial one(s) being the mincut(s) with highest probability.

With this IS scheme, it can be shown that BRE is satisfied in general by the estimator when the unreliability of each link e is $q_e = O(\epsilon)$ and more exactly it is a polynomial in ϵ for a rarity parameter ϵ. Here again, VRE is satisfied for graphs verifying some properties identified in [28]. VRE was never obtained by the variance reduction schemes previously proposed in the literature.

As a last remark, this IS technique can be combined with other existing variance reduction techniques or graph reduction techniques, for an additional efficiency improvement [11, 29].

For many numerical illustrations of the efficiency of the methods, we advise the reader to go to [11, 28, 29].

2.4. **Conclusion**

We have highlighted in this section how the zero-variance IS estimator can be characterized and then efficiently approached if we have at our disposal a heuristic, even a rough one, of the value(s) of interest. We have briefly recalled how this can be applied to two types of reliability models and hope that it will trigger similar applications in other fields.

3. Approximation of zero-variance importance sampling estimators for light-tailed sums

Improving Importance Sampling estimators for rare event probabilities requires sharp approximations of the optimal Importance Sampling density, leading to a nearly zero-variance estimator. This section presents a sharp approximation of the density of long runs of a random walk conditioned by an average of a function of its summands as their number tends to infinity. This is achieved when the event is defined by the fact that the end value of this random walk belongs to some countable union of intervals. Using this approximation in the implementation of an IS scheme can provide strongly efficient estimators.

3.1. **Introduction**

In this section, we consider the problem of efficient estimation of the probability of large deviations for a sum of independent, identically distributed, light-tailed and non-lattice random variables. This is a very specific case, since most of the rare events which are studied in the literature are not of this form; see e.g. [37] for a queuing application or [20] for applications in credit risk-modeling. Nevertheless the problem considered here, besides being of independent importance, may also be considered as a building block for more complex problems.

Consider $\mathbf{X}_1^n = (\mathbf{X}_1, ..., \mathbf{X}_n)$ n i.i.d. random variables with known common density $p_{\mathbf{X}}$ on $\mathbb{R}$, and u a real valued measurable function defined on $\mathbb{R}$. Define $\mathbf{U} := u(\mathbf{X})$ and

$$\mathbf{U}_{1,n} := \sum_{i=1}^{n} \mathbf{U}_i.$$

We intend to estimate

$$P_n := P\left(\mathbf{U}_{1,n} \in nA\right)$$

for large but fixed n and where A is a countable union of intervals of $\mathbb{R}$. The case where $A = A_n := (a_n, \infty)$ where a_n is a convergent sequence is studied in details in [8].

The state-independent IS scheme for rare event estimation (see [10] or [41]), rests on two basic ingredients: the sampling distribution is fitted to the so-called dominating point (which is the point where the quantity to be estimated is mostly captured; see [36]) of the set to be measured; independent and identically distributed replications under this sampling distribution are performed. More recently, a state-dependent algorithm leading to a strongly efficient estimator is provided by [5] when $u(x) = x$ and $A = (a; \infty)$ (or, more generally in $\mathbb{R}^d$, with a smooth boundary and a unique dominating point). Indeed, adaptive tilting defines a sampling density for the $i-$th r.v. in the run which depends both on the target event $(\mathbf{U}_{1,n} \in nA)$ and on the current state of the path up to step $i-1$ for A independent of n. Jointly with an ad hoc stopping rule controlling the excursion of the current state of the path, this algorithm provides an estimate of P_n with a coefficient of variation independent upon n. This result shows that nearly optimal estimators can be obtained without approximating the conditional density.

Our proposal is somehow different since it is based on a sharp approximation result of the conditional density of long runs. The approximation holds for any point conditioning of the form $(\mathbf{U}_{1,n} = nv)$; sampling v in A according to the distribution of $\mathbf{U}_{1,n}$ conditioned upon $(\mathbf{U}_{1,n} \in nA)$ produces the estimator; by its very definition this procedure does not make use of any dominating point, since it randomly explores the set A. Indeed, our proposal hints on two choices: first do not make use of the notion of dominating point and explore

all the target set instead (no part of the set A is neglected); secondly, do not use i.i.d. replications, but merely sample long runs of variables under a proxy of the optimal sampling scheme.

The basic estimate of P_n is defined as follows: generate L i.i.d. samples $X_1^n(l) := (X_1(l), .., X_n(l))$ with underlying density $p_{\mathbf{X}}$ and define

$$P^{(n)}(A) := \frac{1}{L}\sum_{l=1}^{L} \mathbb{1}_{\mathcal{E}_n}\left(X_1^n(l)\right)$$

where

$$\mathcal{E}_n := \{(x_1, ..., x_n) \in \mathbb{R}^n : (u(x_1) + ... + u(x_n)) \in nA\}.$$

The Importance Sampling estimator of P_n with sampling density g on $\mathbb{R}^n$ is

$$P_g^{(n)}(A) := \frac{1}{L}\sum_{l=1}^{L} \hat{P}_n(l)\mathbb{1}_{\mathcal{E}_n}\left(Y_1^n(l)\right) \tag{8}$$

where $\hat{P}_n(l)$ is called "importance factor" and can be written

$$\hat{P}_n(l) := \frac{\prod_{i=1}^{n} p_{\mathbf{X}}\left(Y_i(l)\right)}{g\left(Y_1^n(l)\right)}$$

where the L samples $Y_1^n(l) := (Y_1(l), ..., Y_n(l))$ are i.i.d. with common density g; the coordinates of $Y_1^n(l)$ however need not be i.i.d.

The optimal choice for g is the density of $\mathbf{X}_1^n := (\mathbf{X}_1, ..., \mathbf{X}_n)$ conditioned upon $(\mathbf{X}_1^n \in \mathcal{E}_n)$, leading to a zero variance estimator. We will propose an IS sampling density which approximates this conditional density very sharply on its first components $y_1, ..., y_k$ where $k = k_n$ is very large, namely $k/n \to 1$. However, but in the Gaussian case, k should satisfy $(n-k) \to \infty$ by the very construction of the approximation. The IS density on $\mathbb{R}^n$ is obtained multiplying this proxy by a product of a much simpler adaptive IS scheme following [5].

We rely on [7] where the basic approximation used in the present section can be found. All proofs of the results in the present section can be found in [8]. The extension to the d-dimensional case is in [12].

3.2. **Notations and assumptions**

3.2.1. *Conditional densities and their approximations*

Throughout the section the value of a density $p_{\mathbf{Z}}$ of some continuous random vector $\mathbf{Z}$ at point z may be written $p_{\mathbf{Z}}(z)$ or $p(\mathbf{Z} = z)$, which may prove more convenient according to the context.

Let p_{nv} denote the density of $\mathbf{X}_1^k$ under the local condition $(\mathbf{U}_{1,n} = nv)$

$$p_{nv}\left(\mathbf{X}_1^k = Y_1^k\right) := p(\mathbf{X}_1^k = Y_1^k \big| \mathbf{U}_{1,n} = nv) \tag{9}$$

where Y_1^k belongs to $\mathbb{R}^k$.

We will also consider the density p_{nA} of $\mathbf{X}_1^k$ conditioned upon $(\mathbf{U}_{1,n} \in nA)$

$$p_{nA}\left(\mathbf{X}_1^k = Y_1^k\right) := p(\mathbf{X}_1^k = Y_1^k \big| \mathbf{U}_{1,n} \in nA).$$

The approximating density of p_{nv} is denoted g_{nv}; the corresponding approximation of p_{nA} is denoted g_{nA}. Explicit formulas for those densities are presented in the next section. The capital symbols P_{nv} et G_{nv} denote the corresponding probability measures.

3.2.2. *Tilted densities and related quantities*

The real valued measurable function u is assumed to be unbounded; standard transformations show that this assumption is not restrictive. It is assumed that $\mathbf{U} = u(\mathbf{X})$ has a density $p_{\mathbf{U}}$ w.r.t. the Lebesgue measure on $\mathbb{R}$ and that the characteristic function of the random variable $\mathbf{U}$ belongs to L^r for some $r \geq 1$.

The r.v. $\mathbf{U}$ is supposed to fulfill the Cramer condition: its moment generating function $\phi_{\mathbf{U}}$ is finite in a non void neighborhood of 0. Define the functions $m(t), s^2(t)$ and $\mu_3(t)$ as the first, second and third derivatives of $\log \phi_{\mathbf{U}}(t)$, and m^{-1} denotes the reciprocal function of m.

Denote

$$\pi_u^\alpha(x) := \frac{\exp(tu(x))}{\phi_{\mathbf{U}}(t)} p_{\mathbf{X}}(x)$$

with $m(t) = \alpha$ and α belongs to the support of $P_{\mathbf{U}}$, the distribution of $\mathbf{U}$. It is assumed that this implicit definition of t makes sense for all α in the support of $\mathbf{U}$. Conditions on $\phi_{\mathbf{U}}(t)$ which ensure this fact are referred to as *steepness properties*, and are exposed in [3], page 153.

3.3. **Conditioned samples**

The starting point is the approximation of p_{nv} defined in (9) on $\mathbb{R}^k$ for large values of k under the point condition

$$(\mathbf{U}_{1,n} = nv)$$

when v belongs to A. We refer to [7] for this result.

We introduce a positive sequence ϵ_n which satisfies

$$\lim_{n\to\infty} \epsilon_n \sqrt{n-k} = \infty \tag{E1}$$

$$\lim_{n\to\infty} \epsilon_n (\log n)^2 = 0. \tag{E2}$$

Define a density $g_{nv}(y_1^k)$ on $\mathbb{R}^k$ as follows. Set

$$g_0(y_1 | y_0) := \pi_u^v(y_1)$$

with y_0 arbitrary and, for $1 \leq i \leq k-1$, define $g(y_{i+1} | y_1^i)$ recursively.

Set t_i the unique solution of the equation

$$m_i := m(t_i) = \frac{n}{n-i}\left(v - \frac{u_{1,i}}{n}\right)$$

where $u_{1,i} := u(y_1) + ... + u(y_i)$.

Define

$$g(y_{i+1} | y_1^i) = C_i p_{\mathbf{X}}(y_{i+1}) \mathfrak{n}(\alpha\beta + v, \alpha, u(y_{i+1}))$$

where C_i is a normalizing constant and $\mathfrak{n}(\mu, \tau, x)$ is the normal density function on $\mathbb{R}$ with mean μ and variance τ at x. Here

$$\alpha = s^2(t_i)(n-i-1)$$

$$\beta = t_i + \frac{\mu_3(t_i)}{2s^4(t_i)(n-i-1)}.$$

Set

$$g_{nv}(y_1^k) := g_0(y_1 | y_0) \prod_{i=1}^{k-1} g(y_{i+1} | y_1^i). \tag{10}$$

Theorem 1. *Assume (E1) and (E2). Then (i)*

$$p_{nv}\left(\mathbf{X}_1^k = Y_1^k\right) = g_{nv}(Y_1^k)(1 + o_{P_{nv}}(\epsilon_n (\log n)^2))$$

and (ii)

$$p_{nv}\left(\mathbf{X}_1^k = Y_1^k\right) = g_{nv}(Y_1^k)(1 + o_{G_{nv}}(\epsilon_n (\log n)^2)).$$

As stated above the optimal choice for the sampling density is p_{nA}. It holds

$$p_{nA}(x_1^k) = \int_A p_{nv}\left(\mathbf{X}_1^k = x_1^k\right) p(\,\mathbf{U}_{1,n}/n = v|\,\mathbf{U}_{1,n} \in nA)dv \tag{11}$$

so that, in contrast with [5] or [10], we do not consider the dominating point approach but merely realize a sharp approximation of the integrand at any point of A and consider the dominating contribution of all those distributions in the evaluation of the conditional density p_{nA}. When $A = (a, \infty)$, the density $p(\,\mathbf{U}_{1,n}/n = v|\,\mathbf{U}_{1,n} \in nA)$ can be well approximated by an exponential density allowing a simpler expression for (11).

3.4. **Adaptive IS estimator for rare event probability**

The IS scheme produces samples $Y := (Y_1, ..., Y_k)$ distributed under g_{nA}, which is a continuous mixture of densities g_{nv} as in (10) with $p\left(\mathbf{U}_{1,n}/n = v|\mathbf{U}_{1,n} \in nA\right)$.

Simulation of samples $\mathbf{U}_{1,n}/n$ under this density can be performed through Metropolis-Hastings algorithm, since

$$r(v, v') := \frac{p(\,\mathbf{U}_{1,n}/n = v|\,\mathbf{U}_{1,n} \in nA)}{p(\,\mathbf{U}_{1,n}/n = v'|\,\mathbf{U}_{1,n} \in nA)}$$

turns out to be independent of the event $\{\mathbf{U}_{1,n} \in nA\}$. The proposal distribution of the algorithm should be supported by A.

The density g_{nA} is extended from $\mathbb{R}^k$ onto $\mathbb{R}^n$ by completing the $n - k$ remaining coordinates in two steps, according to the method developed in [5],

$$g_{nA}\left(y_{k+1}^n | y_1^k\right) = \prod_{i=k+1}^{\tau_n} \pi^{m_i}(y_i) \prod_{i=\tau_n+1}^{n} \pi^{m_{\tau_n}}(y_i)$$

where τ_n is defined by

$$\tau_n = \tau_{n,1} \wedge \tau_{n,2} \wedge n$$

with

$$\tau_{n,1} = \inf_{k \le i \le n} \{m_i > \lambda\} \qquad \tau_{n,2} = \inf_{k \le i \le n} \{m_i < (n-i)^{-3/2}\}.$$

where λ is a constant depending on A. The stopping criterion is based on two simple facts. The first stopping time is used, when we are approaching a scaling for which the large deviations asymptotics are no longer applicable. The second stopping time is used when there is no need for applying importance sampling sequentially as the event of interest is not rare anymore, see [5] for details.

We now define our IS estimator of P_n. Let $Y_1^n(l) := Y_1(l), ..., Y_n(l)$ be generated under g_{nA}. Let

$$\widehat{P_n}(l) := \frac{\prod_{i=0}^n p_{\mathbf{X}}(Y_i(l))}{g_{nA}(Y_1^n(l))} \mathbb{1}_{\mathcal{E}_n}\left(Y_1^n(l)\right)$$

and define

$$\widehat{P_n} := \frac{1}{L} \sum_{l=1}^{L} \widehat{P_n}(l),$$

in accordance with (8).

Let us conclude this section with two remarks:

(1) Using arguments from [8] and [5], $\widehat{P_n}$ is strongly efficient when $A = (a, \infty)$ and $\lambda > 2a$. Indeed, on the k first coordinates, the approximation of the zero-variance change of measure coincide with the true density. On the $n - k$ last coordinates, using large deviation arguments, [5] proves the strong efficiency of $\widehat{P_n}$.

(2) The performance of these methods for an ill-behaved set, when this set is non connected, and bears some asymmetry with respect to the underlying probability model has been studied in [8]. As noticed in [21], this case leads to overwhelming loss in relative accuracy for the probability to be estimated. Simulated results enlighten the gain of the present approach over state-dependent Importance Sampling schemes. Using different arguments, Dupuis and Wang [17] are led to a similar result.

4. Adaptive Directional Stratification for the controlled estimation of rare event probabilities

In the context of structural reliability, a small probability to be assessed, a high computational time model and a relatively large input dimension are typical constraints which brought together lead to an interesting challenge. Indeed, in this framework many existing stochastic methods fail in estimating the failure probability with robustness.

Therefore, the aim of this section is to present a method introduced to overcome the specific constraints mentioned above. This new method turns out to be competitive compared with the existing techniques. It is a variant of accelerated Monte Carlo simulation method, named ADS-2 - Adaptive Directional Stratification. It combines, in a two steps adaptive strategy, the stratification into quadrants and the directional simulation techniques. Two ADS-2 estimators and their properties are presented.

4.1. **Introduction**

One way to assess the reliability of a structure from physical considerations is to use a probabilistic approach: it includes the joint probability distribution of the random input variables of a numerical deterministic model representing the physical behavior of the studied structure, for instance its failure margin. We consider a real-valued failure function, $G : \mathbb{D} \subset \mathbb{R}^p \to \mathbb{R}$, whose p-dimensional input vector $\mathbf{X} = (X^1, ..., X^p)$ is random. Then, we assume that the probability density function of the random vector $\mathbf{X}$ exists and is known.

In this context, we want to assess the failure probability P_f:

$$P_f = \mathbb{P}(G(\mathbf{X}) < 0) = \int_{G(\mathbf{x})<0} f(\mathbf{x}) d\mathbf{x} \tag{12}$$

Furthermore, four key features characterize our agenda. (C_0): G may be complex and greedy in computational resources: even when involving high performance computing, industrial constraints limit the number of evaluations of G to a few thousands. (C_1): no continuity or derivability assumptions are considered for G. (C_2): the failure is a rare event, which means that P_f is very small. In this work, we will consider that a small probability is a probability lower than 10^{-3}. (C_3): the results must be robust, i.e. with explicit and trustworthy error control.

The first three constraints correspond to our working hypotheses and the last constraint is the key goal motivating this research. Here, we consider the accelerated Monte Carlo methods and try to develop a specific one, which "converges" as fast as possible and enables to obtain an estimation error control in a reasonable number of simulations. A full analysis of the following method can be found in [33, 34].

4.2. **Preliminary: stratification, directional sampling and optimization**

The idea is to take advantage of the possibilities offered by the stratification and directional simulation methods: optimal allocation result, adaptive strategy, efficient small probability estimation and reasonable calculation time.

We first move the problem into a "spherical space": the Gaussian space, using the Nataf transformation, $\mathbf{U} = T_N(\mathbf{X})$ [31]. In other words, the expectation estimation becomes: $P_f = \mathbb{P}\big(g(\mathbf{U}) < 0\big)$ with $g = G \circ T_N^{-1}$ and $\mathbf{U}$ a random vector with Gaussian and independent components.

Then, $\mathbf{U}$ can be decomposed into the product $R\mathbf{A}$ (R and $\mathbf{A}$ independent) with R^2 a chi-square random variable and $\mathbf{A}$ a random vector uniformly distributed over the unit sphere $S_{p-1} \subset \mathbb{R}^p$ [19]. We will denote $\mathcal{L}(\mathbf{A})$ the distribution of $\mathbf{A}$. Then, Denoting f_R the probability density function of R and using the conditional expectation properties, we have: $P_f = \mathbb{E}\big(\xi(\mathbf{A})\big)$ with $\xi(\mathbf{a}) = \int \mathbb{1}_{g(r\mathbf{a})<0} f_R(r)dr$ a bounded function.

The directional sampling method is based on the fact that we are able to calculate $\xi(\mathbf{a})$ for any $\mathbf{a}$. Indeed, the conditional expectation $\xi(\mathbf{a})$, which is the probability: $\mathbb{P}(g(R\mathbf{a}) < 0)$, is then given as a function of the chi-square cumulative distribution function and $r_{u,i}(\mathbf{a})$ and $r_{l,i}(\mathbf{a})$ respectively the upper and the lower bounds of the i-th interval where $g(r\mathbf{a}) < 0$ [32]. $r_{u,i}(\mathbf{a})$ and $r_{l,i}(\mathbf{a})$ are approximated thanks to root-finding algorithms as the dichotomic or the Brent methods. A stopping criterion for the dichotomic algorithm in the directional sampling method, applied to the estimation of small failure probability, can be found in [34].

At this point, we introduce the stratification method by partitioning the "directional space", in other words the space where $\mathbf{A}$ takes its values, that is to say the $(p-1)$-dimensional unit sphere $S_{p-1} \subset \mathbb{R}^p$, into m strata and we denote $I := \{1, ..., m\}$. The natural strata adapted to directional draws are cones and partitioning S_{p-1} is equivalent to make a partition of the general space into cones. Let us denote by $(q_i)_{i\in I}$ a partition of S_{p-1} into m strata.

Let us denote by n the total number of directional draws, e.g. the number of directions we want to simulate. Let us consider an allocation of the n directional draws in each stratum described by the sequence $\mathbf{w} = (w_i)_{i\in I}$ such that $\sum_{i=1}^m w_i = 1$. The number of draws in the i-th stratum is $n_i = \lfloor nw_i \rfloor$ with $\lfloor . \rfloor$ the floor function (we neglect the rounding errors in the analysis). We can express the expectation P_f as: $P_f = \sum_{i=1}^m \rho_i P_i$ with $P_i = \mathbb{E}\big(\xi(\mathbf{A}^i)\big)$, $\mathbf{A}^i \sim \mathcal{L}(\mathbf{A}|\mathbf{A} \in q_i)$ and $\rho_i = \mathbb{P}(\mathbf{A} \in q_i)$.

Now, we estimate P_i by drawing n_i directions in the i-th stratum. This can be done by using a simple rejection method. We get: $\hat{P}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \xi(\mathbf{A}_j^i)$ with $(\mathbf{A}_j^i)_{j=1,...,n_i}$ a family of independent and identically distributed (i.i.d.) random vectors with distribution $\mathcal{L}(\mathbf{A}|\mathbf{A} \in q_i)$.

We obtain the following unbiased estimator by coupling the Directional Sampling and the Stratification methods: $\hat{P}_f^{DSS} = \sum_{i=1}^m \rho_i \hat{P}_i$ and its variance is given by σ^2_{DSS}/n with $\sigma^2_{DSS} = \sum_{i=1}^m \frac{\rho_i^2 \sigma_i^2}{w_i}$ where $\sigma_i^2 = \mathbb{V}\big(\xi(\mathbf{A}^i)\big)$.

The DSS estimator is consistent and asymptotically normal. Like in the standard stratification method, we can determine the optimal distribution of directional draws in each stratum in order to minimize the variance estimator, which can be useful for the design of an adaptive strategy. The unique optimal solution $\mathbf{w_{opt_1}}$ indicates that we must give large weights to the strata with the largest uncertainties, i.e. variances.

4.3. **ADS-2 method**

4.3.1. *General principle of the ADS-2 method*

Now, we have all the tools to set up our 2-steps adaptive method with an estimator coupling stratification and directional sampling. We will call this estimator the 2-adaptive directional stratification (ADS-2) estimator. It will be denoted by $\hat{P}_f^{ADS-2}$.

Let us explain our strategy. We split the total number n of directional draws into two parts: $\gamma_1(n)n$ and $\gamma_2(n)n$ such that $\sum_{i=1}^2 \gamma_i(n) = 1$ and $(\gamma_i(n))_{i=1,2}$ are functions of variable n. The first part, $\gamma_1(n)n$, will be used to estimate the optimal allocation minimizing the variance (learning step), and the second part, $\gamma_2(n)n$, will be used to estimate the failure probability (estimation step), using the estimation of the optimal allocation.

For instance, we can take $\gamma_1(n) = n^{l-1}$ with $l \in (0, 1)$, or $\gamma_1(n) = \gamma_1$ with $\gamma_1 \in (0, 1)$ and, in an obvious way, we have $\gamma_2(n) = 1 - \gamma_1(n)$. Let us set the strata as the quadrants of the Gaussian space, denoted by $(q_i)_{i\in I}$ with

$m = 2^p$ the number of quadrants. Indeed, without any other information, the most natural choice is quadrants, which are well adapted for directional draws and enable a good survey of the Gaussian space. Also, without prior information, we should choose a prior uniform allocation. We denote it $(w_i)_{i\in I}$, and as we considered quadrants: $w_i = \rho_i = 1/m$ for all $i \in I$. Our method can be readily extended to the case in which a prior information is available on the optimal allocation. From now on, $n_i = \lfloor \gamma_1(n) n w_i \rfloor$ for all $i \in I$ stands for the number of draws achieved in each stratum in the estimation step. Let us suppose that we use the directional and stratification estimator $\hat{P}_f^{DSS}$. The next idea is to replace the initial allocation by an estimation of the optimal allocation, $\mathbf{w_{opt_1}}$, made with the $\gamma_1(n)n$ first draws. But first, we need to define our estimators.

4.3.2. *The non-recycling ADS-2 estimator*

The non-recycling ADS-2 estimator does not use the $\gamma_1(n)n$ first draws for the expectation estimation. It only uses the second part of draws, $\gamma_2(n)n$:

$$\hat{P}_{f,nr}^{ADS-2} = \sum_{i=1}^{m} \rho_i \frac{1}{N_i^{nr}} \sum_{j=1}^{N_i^{nr}} \xi(\mathbf{A}_{n_i+j}^i) \tag{13}$$

where

$$N_i^{nr} := \lfloor (1 - \epsilon(n)) n_i^{nr} + \epsilon(n) \rho_i n \rfloor \tag{14}$$

with $n_i^{nr} = \left\lfloor \gamma_2(n) n \tilde{W}_i^{nr} \right\rfloor$ and $(\tilde{W}_i^{nr})_{i=1,\dots,m}$ given by an empirical estimation of the theoretical optimal allocation $\mathbf{w_{opt_1}}$. Besides, $\epsilon(n) \in (0, 1]$ enables in the estimation step to draw some directions in a stratum even if the estimated allocation returns zero, which can give a correction to the bias brought by a wrong estimation of an allocation. For a limited number of simulations, we should take $\epsilon(n)$ as small as possible: we propose $\epsilon(n) = 2/(n\rho_i) = 2^{p+1}/n$. The non-recycling estimator is consistent, unbiased and we can get an expression of its variance

$$\mathbb{V}\big(\hat{P}_{f,nr}^{ADS-2}\big) = \frac{1}{n} \sum_{i=1}^{m} \rho_i^2 \sigma_i^2 \times \mathbb{E}\left(\left[\big(1 - \epsilon(n)\big) \gamma_2(n) \frac{\rho_i \tilde{\sigma}_i}{\sum_{j=1}^{m} \rho_j \tilde{\sigma}_j} + \epsilon(n) \rho_i \right]^{-1} \right)$$

for all n.

Two central limit theorems can be proved for the non-recycling estimator. Straightforwardly, to get a confidence interval on the estimator, we assume that the error is Gaussian with variance the estimated variance of the estimator.

4.3.3. *Improvement of the non-recycling ADS-2 estimator*

When the physical dimension grows, the number of strata of the ADS-2 method increases exponentially: indeed, in dimension p, the number of quadrants is 2^p. As a minimum of simulations is required to explore each quadrant, the number of directional simulations needed can be too large with respect to the restricted number of simulations we are limited to (C_0).

Now, the idea is to get, with the simulations performed in the first step (learning step), a sort of the random variables in function of their influence on the failure event. Then, we only stratify the most important ones. We use the simulations done in the first step to detect the important variables and to estimate the optimal allocation in the new strata.

In the second step, we get the final estimation with this allocation. To determine if a random variable will be stratified, we propose the following method. We first index the quadrants. The input index $k \in 1, ..., p$ is given the tag i_k which takes its values in $\{-1, 1\}$ and corresponds to the input sign. Thus, each quadrant $i \in \{1, ..., m\}$ is characterized by a p-uple $\mathbf{i} = (i_1, ..., i_p)$. From now on, for the strata indexation, we will use, in an equivalent way, either the indexation $i \in \{1, ..., m\}$ or its associated multi-index $\mathbf{i} = (i_1, ..., i_p) \in \{-1, 1\}^p$.

Then, we define the sequence $\tilde{\mathbf{T}} = (\tilde{T}_k)_{k=1,...,p}$ by:

$$\tilde{T}_k = \sum_{i_l \in \{-1,1\}, l \neq k} \left| \tilde{P}_{(i_1,...,i_{k-1},-1,i_{k+1},...,i_p)} - \tilde{P}_{(i_1,...,i_{k-1},1,i_{k+1},...,i_p)} \right|$$

with $\tilde{P}_{(i_1,...,i_p)}$ the estimation of the expectation in the quadrant $(i_1, ..., i_p)$ obtained during the learning step.

Thus, $\tilde{T}_k$ aggregates the differences of the expectations between the quadrants along dimension k. The larger $\tilde{T}_k$ is, the more influential the k-th input is. Then, we sort sequence $\tilde{\mathbf{T}}$ by decreasing order and we decide to stratify only over the $p' < p$ first dimensions, the other inputs being simulated without stratification.

Next, we estimate the optimal allocation to be achieved in the new $m' = 2^{p'}$ hyper-quadrants. Consequently, the final ADS-2^+ estimator is given by:

$$\hat{P}_{f,nr}^{ADS-2^+}(\tilde{\mu}) = \sum_{i=1}^{m'} \rho'_i \frac{1}{N_i^{nr}(\tilde{\mu})} \sum_{j=1}^{N_i^{nr}(\tilde{\mu})} \xi(\mathbf{A}_j^i) \tag{15}$$

with

$$N_i^{nr}(\tilde{\mu}) := \left\lfloor \big(1 - \epsilon(n)\big)\gamma_2(n) n \tilde{W}_i^{nr}(\tilde{\mu}) + \epsilon(n)\rho'_i n \right\rfloor \tag{16}$$

and $(\mathbf{A}_j^i)_{j=1,...,N_i^{nr}(\tilde{\mu})}$ a family of i.i.d. random vectors with distribution $\mathcal{L}\big(\mathbf{A}|\mathbf{A} \in q_i(\tilde{\mu})\big)$.

4.4. **Comparison of ADS with the directional simulation and the subset simulation methods**

In this section, we apply the ADS methods to the hyperplanes H_1 and H_2 defined by $H_1 : \sum_{i=1}^p x_i = k_1$ and $H_2 : x_p = k_2$. H_2 has one influential variable and H_1 has all its variables influential. We denote by N the total number of runs of the method. We define the estimated percentage of estimations fallen in the estimated two-sided symmetric 95% confidence interval as:

$$\hat{PCI} = \frac{100}{N} \sum_{k=1}^{N} \mathbb{1}_{P_f \in [\hat{P}_k^-; \hat{P}_k^+]} \quad \text{with} \quad \hat{P}_k^{\pm} = \hat{P}_{f,k}^{ADS-2} \pm \alpha_{97,5\%} \frac{\hat{\sigma}_k^{ADS-2}}{\sqrt{n}},$$

where $\alpha_{97,5\%}$ is the $97,5\%$ Gaussian quantile, $\hat{P}_{f,k}^{ADS-2}$ and $\hat{\sigma}_k^{ADS-2}$ the k-th failure probability and standard deviation estimations. We denote $\hat{CV}$ the estimated coefficient of variation.

We compare in tables 4.1 the ADS-2^+ method with the Subset Simulation method (SS), also called multilevel splitting, which is built to overcome the curse of dimensionality and responds to the four constraints presented in section 4.1 (see for example [2]). Also, the results obtained with the ADS-2 method are presented. The SS results have been obtained using the open-source Matlab toolbox: FERUM (Finite Element Reliability Using Matlab) version 4.0.

Method	n	NG	$\hat{CV}$ (%)	$\hat{PCI}$
SS	500	3300	60	77
SS	700	4600	51	88
SS	1000	6700	42	81
DS	600	3100	42	86
DS	900	4501	35	90
ADS-2	1200	3219	55	61
ADS-2	1750	4675	46	61
ADS-2	2550	6760	39	76
ADS-2^+ ($p'=3$)	1200	3139	41	78
ADS-2^+ ($p'=3$)	1800	4799	31	83
ADS-2^+ ($p'=2$)	1500	3962	30	90
ADS-2^+ ($p'=1$)	1500	3957	28	92

(a) Hyperplane H_1. $P_f = 10^{-6}$. $p = 5$. $N = 100$.

Method	n	NG	$\hat{CV}$ (%)	$\hat{PCI}$
SS	300	1989	65	86
SS	500	3330	51	88
SS	700	4630	43	85
SS	1000	6800	36	93
DS	200	3154	62	70
DS	400	6672	50	82
DS	600	10100	42	82
ADS-2	300	2033	16	90
ADS-2	500	3340	13	93
ADS-2	650	4426	11	96
ADS-2^+ ($p'=3$)	700	2487	21	92
ADS-2^+ ($p'=3$)	1200	4333	16	94

(b) Hyperplane H_2. $P_f = 10^{-6}$. $p = 5$. $N = 100$.

Table 4.1 Results on two different hyperplanes.

Table 4.1 confirms the efficiency of the ADS-2^+ method in comparison with the ADS-2 method. Moreover, in comparison with the SS method and for approximately the same $\hat{CV}$ and $\hat{PCI}$, the ADS-2^+ method ($p' = 3$) enables to reduce the number of calls to the failure function by approximately a factor 2. Also, for a better choice of p' (equal to 1), we can reduce the $\hat{CV}$ of 25% for approximately the same number of calls to the failure function. Finally, the same study has been performed on H_1 and H_2 for p going from 5 to 8 and $P_f = 10^{-6}$ and 10^{-8}. As predictable, the SS method is completely robust with respect to the increase of the dimension and gives almost exactly the same results. On H_1, the ADS-2 method always outperforms the SS method: for instance in dimension 7 for a failure probability of 10^{-8}, we divide by a factor 2 the $\hat{CV}$ with the half number of calls to the failure function used for SS (8000) while keeping a good $\hat{PCI}$. In the worst case, i.e. when considering H_2, $P_f = 10^{-8}$ and $p = 7$, the results between SS and ADS-2^+ are equivalent. Finally, we can see in tables 4.1 that the ADS methods outperform the simple directional simulation method, slightly on H_2 and completely on H_1.

4.5. **Discussion and conclusion**

The ADS-2 strategy concentrates the runs into the most important parts of the sample space, which results in an asymptotic optimal error variance. For a limited number of simulations, theoretical and numerical results show that a relevant solution to get a confidence interval on the estimation is to consider the error Gaussian and to consider the non-asymptotic estimation of the variance of the estimator.

Furthermore, when the number of simulations is limited, the most natural assumption will be to choose $\gamma_1(n) = \gamma_1$, in order to have a sufficient number of draws in both learning and estimation steps of the ADS-2 method. A numerical study over this parameter has been performed in [34]. Also in order to avoid a constant bias on the estimation, it is important to choose a non zero value for the parameter $\epsilon(n)$ which enables, during the estimation step, to perform some simulations in the quadrants which are not detected as important ones in the learning step.

Finally, we compare the ADS-2 and ADS-2^+ methods to the subset simulation method which is one of the most relevant methods to use in the context described in section 4.1. The results show that in the considered configurations the ADS-2^+ method outperforms the subset simulation method. Hence, the ADS methods are very efficient when the following conditions are met: a number of calls to the failure function of a few thousand, an order of magnitude of the failure probability less than 10^{-4}, a dimension less than 7, no regularity assumption on the failure function, and a need for explicit and trustworthy error controls.

References

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

[2] S. K. Au and J. L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263-277, 2001.

[3] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, New-York, 1978.

[4] J. H. Blanchet, P. W. Glynn, and J. C. Liu. Fluid heuristics, Lyapunov bounds, and efficient importance sampling for a heavy-tailed $G/G/1$ queue. *Queueing Systems*, 57:99–113, 2007.

[5] J. H. Blanchet, K. Leder, and P. W. Glynn. Efficient simulation of light-tailed sums: an old-folk song sung to a faster new tune... In P. L'Ecuyer and A. B. Owen, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2008*, pages 227–248, Berlin, 2009. Springer-Verlag.

[6] Z. I. Botev and D. P. Kroese. An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting. *Methodology and Computing in Applied Probability*, 10(4):471-505, 2008.

[7] M. Broniatowski and V. Caron. Long runs under a conditional limit distribution. *arXiv:1010.3616*, 2012.

[8] M. Broniatowski and V. Caron. Small variance estimators for rare event probabilities. *ACM Transactions on Modeling and Computer Simulation,* 23(1):Article 6, 2013.

[9] J. A. Bucklew. *Introduction to rare event simulation*. Springer Series in Statistics. Springer-Verlag, New York, 2004.

[10] J. A. Bucklew, P. Ney and J. S. Sadowsky. Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *Journal of Applied Probability*, 27(11):49-61, 1990.

[11] H. Cancela, P. L'Ecuyer, G. Rubino, and B. Tuffin. Combination of conditional Monte Carlo and approximate zero-variance importance sampling for network reliability estimation. In B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, editors, *Proceedings of the 2010 Winter Simulation Conference*, pages 1263–1274, 2010.

[12] V. Caron. A conditional limit theorem. Applications to conditional inference and Importance Sampling methods. *HAL : tel-00763369*, 2012.

[13] F. Cérou, P. Del Moral, T. Furon, and A. Guyader. Sequential Monte Carlo for Rare Event Estimation. *Statistics and Computing*, 22(3), pp.795-808, 2012.

[14] F. Cérou, A. Guyader, T. Lelièvre, and D. Pommier. A multiple replica approach to simulate reactive trajectories. *The Journal of Chemical Physics*, 134(5):054108, 2011.

[15] N. Chopin and C. P. Robert. Properties of nested sampling. *Biometrika*, 97(3):741-755, 2010.

[16] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B*, 68(3):411-436, 2006.

[17] P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. *Stochastics and Stochastics Reports*, 76:481–508, 2004.

[18] P. Dupuis and H. Wang. Dynamic importance sampling for uniformly recurrent Markov chains. *Annals of Applied Probability*, 15:1–38, 2005.

[19] K. T. Fang, S. Kotz, and K. Ng. *Symmetric Multivariate and Related Distributions*. Monographs on statistics and applied probability, Chapman and Hall, London, New York, 1990.

[20] P. Glasserman and J. Li. Importance sampling for portfolio credit risk. *Management Science*, 51:1643-1656, 2005.

[21] P. Glasserman and Y. Wang. Counterexamples in importance sampling for large deviations probabilities. *Annals of Applied Probability,* 7(3):731-746, 1997.

[22] A. Guyader, N. Hengartner, and E. Matzner-Løber. Simulation and estimation of extreme quantiles and extreme probabilities. *Applied Mathematics and Optimization*, 64:171-196, 2011.

[23] J. M. Hammersley and D. C. Handscomb. *Monte Carlo methods*. Methuen & Co. Ltd., London, 1965.

[24] H. Kahn and T. E. Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards Appl. Math. Series*, 12:27-30, 1951.

[25] 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.

[26] P. L'Ecuyer, F. Le Gland, P. Lezaud, and B. Tuffin. Splitting techniques. In G. Rubino and B. Tuffin, editors, *Rare Event Simulation Using Monte Carlo Methods*, pages 39–61. Wiley, 2009. Chapter 3.

[27] 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.

[28] P. L'Ecuyer, G. Rubino, S. Saggadi, and B. Tuffin. Approximate zero-variance importance sampling for static network reliability estimation. *IEEE Transactions on Reliability*, 8(4):590–604, 2011.

[29] P. L'Ecuyer, S. Saggadi, and B. Tuffin. Graph reductions to speed up importance sampling-based static reliability estimation. In *Proceedings of the 2011 Winter Simulation Conference*, 2011.

[30] P. L'Ecuyer and B. Tuffin. Approximating zero-variance importance sampling in a reliability setting. *Annals of Operations Research*, 189:277–297, 2011.

[31] P. Liu and A. D. Kiureghian. Structural reliability under incomplete probability information. *Journal of Engineering Mechanics*, 112(1):85–104, 1986.

[32] H. Madsen and O. Ditlevsen. Structural reliability methods. *Probabilistic Engineering Mechanics*, Wiley, 1996.

[33] M. Munoz Zuniga, J. Garnier, E. Remy, and E. Rocquigny. Analysis of adaptive directional stratification for the controlled estimation of rare event probabilities. *Statistics and Computing*, 22(3):809–821, 2012.

[34] M. Munoz Zuniga. Méthodes stochastiques pour l'estimation contrôlée de faibles probabilités sur des modèles physiques complexes. Application au domaine nucléaire. *Ph.D. thesis, University of Paris VII*, 2011.
[35] M. K. Nakayama. General conditions for bounded relative error in simulations of highly reliable Markovian systems. *Advances in Applied Probability*, 28:687–727, 1996.
[36] P. Ney. Dominating points and the asymptotics of large deviations for random walk on $\mathbb{R}^d$. *Annals of Probability*, 11(1):158-167, 1983.
[37] S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queue. *IEEE Transactions on Automatic Control*, 34:158-167, 1990.
[38] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics, Springer-Verlag, New York, second edition, 2004.
[39] G. Rubino and B. Tuffin, editors. *Rare Event Simulation using Monte Carlo Methods*. Wiley, 2009.
[40] R. Rubinstein. The Gibbs cloner for combinatorial optimization, counting and sampling. *Methodology and Computing in Applied Probability*, 11(4):491-549, 2009.
[41] J. S. Sadowsky. On Monte Carlo estimation of large deviations probabilities. *Annals of Applied Probability,* 9(2):493-503, 1996.
[42] M. J. Schervish. *Theory of statistics*. Springer Series in Statistics, Springer-Verlag, New York, 1995.
[43] P. Shahabuddin. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science*, 40(3):333–352, 1994.
[44] J. Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833-859, 2006.
[45] B. Tuffin. *La simulation de Monte Carlo*. Hermès, 2010.