RARE EVENT SIMULATION USING REVERSIBLE SHAKING TRANSFORMATIONS

E. GOBET* AND G. LIU†

Abstract. We introduce random transformations called reversible shaking transformations which we use to design two schemes for estimating rare event probability. One is based on interacting particle systems (IPS) and the other on time-average on a single path (POP) using ergodic theorem. We discuss their convergence rates and provide numerical experiments including continuous stochastic processes and jump processes. Our examples cover rather important situations related to insurance, queueing system and random graph for instance. Both schemes have good performance, with a seemingly better one for POP.

Key words. rare event, Monte Carlo simulations, ergodic properties, interacting particle systems

AMS subject classifications. 65C05, 65C40, 37M25, 60K35, 65C35

This version: August 27, 2014

1. Introduction.

1.1. Context and framework. The analysis of rare events is an important issue in economy, engineering and life sciences among other fields, with significant applications such as actuarial risks [AA10], communication network reliability [Rob03], aircraft safety [PW05], random graphs [Bol01] applied to social networks and analysis of epidemic spreading etc, see [Buc04, RG09] and references therein.

We start by specifying the probabilistic setting, which takes a rather general form to include both finite and infinite dimensional situations, examples are given later. The state space is described by a measurable space $(\mathbb{S}, \mathcal{S})$, where $(\mathbb{S}, d_{\mathbb{S}})$ is a metric space[1] and $\mathcal{S}$ is the Borel sigma-field generated by its open sets. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we consider a random variable (measurable mapping) $X : \Omega \mapsto \mathbb{S}$ and a measurable set $A \subsetneq \mathbb{S}$, then the rare event under investigation is defined by $\{X \in A\}$. In this work, we aim at numerically computing the rare-event statistics

$$\mathbb{P}(X \in A) \qquad \text{and} \qquad \mathbb{E}(\varphi(X)|X \in A) \tag{1.1}$$

for bounded measurable functions $\varphi : \mathbb{S} \mapsto \mathbb{R}$. To avoid uninteresting situations, we assume from now on that $\mathbb{P}(X \in A) > 0$.

If $\mathbb{P}(X \in A)$ is very small (say smaller than 10^{-4}), a plain Monte Carlo method is inefficient because the target rare event is realized in only a very small proportion of simulations. One way to circumvent this problem is to use Important Sampling (IS) techniques [RK08]. In short, IS techniques use information on the specific configuration of the model (given by X and A) to design a new probability measure under which

*Centre de Mathématiques Appliqués, Ecole Polytechnique and CNRS, Route de Saclay, 91128 Palaiseau cedex, France. EMAIL: emmanuel.gobet@polytechnique.edu. The author research is part of the Chair *Financial Risks* of the *Risk Foundation* and the *Finance for Energy Market Research Centre*.

†Centre de Mathématiques Appliqués, Ecole Polytechnique and CNRS, Route de Saclay, 91128 Palaiseau cedex, France. EMAIL: gang.liu1988@gmail.com

[1]For the applications we have in mind, it may be an Euclidean space equipped with the usual distance, or the space of continuous functions $\mathbb{C}([0,1],\mathbb{R})$ equipped with the uniform metric, or the space of càdlàg functions $\mathbb{D}([0,1],\mathbb{R})$ with the Skorohod topology [Bil99].

rare events are more likely to occur. Thus this approach heavily relies on the particularity of the model. More recently, an IPS (Interacting Particle System) method (a.k.a. Genetic Genealogical algorithm) has been designed in [Del04, DG05, CDLL06] to estimate the probability of rare event related to the terminal value of a Markov chain, see [KLL12] for switching diffusions. And in [CDFG12], the IPS algorithm is used to estimate probability of rare event related to a static finite dimensional distribution via a particular Markov kernel, which is called *shaking transformation* in this work. The splitting techniques initiated in [VV91] correspond to another particle-based approach where processes are selected and split randomly as they reach rarer and rarer sets; unlike the usual IPS of [DG05], the number of particles becomes random.

1.2. Contributions. Our methodology is in the vein of methods where sampling is performed under the initial measure (thus different from IS) and where $\{X \in A\}$ is included in a cascade of bigger and bigger events $\{X \in A_k\}_{0 \leqslant k \leqslant n}$ built from nested $\mathbb{S}$-subsets

$$\mathbb{S} := A_0 \supset \cdots \supset A_k \supset \cdots \supset A_n := A, \tag{1.2}$$

similarly to the usual splitting methodology [VV91]. Here the integer number $n \geqslant 2$ of intermediate subsets is fixed. Our main contributions are threefold.

a) In this work, we make use of reversible random transformations (so-called shaking transformations) that leave invariant the distribution of X (see hypothesis **(K)**). Our first contribution (see Section 3) is to give explicit forms of such transformations in many cases (including stochastic processes), which are easier to implement than Metropolis-Hasting or Gibbs type transformations and turn out to be efficient in our subsequent numerical experiments. Their parametrization is rather tractable and is an asset for appropriately tuning the shaking force. In addition, combined with rejection techniques, it allows to define another random transformation leaving unchanged the distribution of X restricted to A_k (see Proposition 2.2).

b) Secondly, we generalize the IPS algorithm in [CDFG12] (where the state space is $\mathbb{R}^d$) to a general metric case $(\mathbb{S}, d_{\mathbb{S}})$, allowing for instance to consider spaces of continuous-time paths. Actually this constitutes a completely new point of view for problems involving stochastic processes, as explained below in Subsection 1.3.

c) Thirdly, for computing (1.1) we propose a new method, called POP (Parallel One-Path ergodic shaking), which numerically evaluates in intermediate steps $\mathbb{P}\left(X \in A_k | X \in A_{k-1}\right)$ for each k. The idea seems similar to the usual splitting methodology [VV91], but the novelty of POP is that each conditional probability is independently estimated as the empirical occupation measure of A_k by a suitable Markov chain of length N. In other words, as a difference with usual methods where the convergence is achieved using a large number of (almost) independent samples, the POP method relies on the ergodicity of Markov chain and converges as $N \to +\infty$. It has the advantage of allowing parallel and separated computations of conditional probabilities, thus reducing the global computational time and the interdependency effects compared to IPS method. This idea of decomposing a rare event into a series of non-rare ones is not new, but to our knowledge it is the first one where those non-rare probability estimations can be made independent. The benefit is visible in our experiments where the standard deviation may be reduced by a factor 4 or more.

Contrary to importance sampling techniques, we don't need any supplementary knowledge about what the typical realizations of rare event are. Thus our methods

apply quite generally without too much caring of specific model configuration. This will be seen in various numerical examples.

The organisation of the paper is as follows: after motivating furthermore our abstract probabilistic model, we describe the algorithms in Section 2. Section 3 gathers various examples of shaking transformations, relevant for the subsequent applications and numerical tests exposed in Section 4. We conclude in Section 5.

1.3. Digression when $\mathbb{S}$ is a space of paths (e.g. $\mathbb{C}([0,1],\mathbb{R})$ or $\mathbb{D}([0,1],\mathbb{R})$). We would like to emphasize the advantage of our abstract setting of metric space $(\mathbb{S}, d_{\mathbb{S}})$, in particular when dealing with continuous-time stochastic processes $(Z_t)_{t\geqslant 0}$. We think that this issue is rather important for applications to stochastic processes. For the exposure, consider for instance that $Z \in \mathbb{C}([0,1],\mathbb{R})$. In our work the methodology for rare-event simulation of Z will not be performed as a scheme creating a time-evolution of the marginals of Z, i.e. $Z_{\tau_0} \to \cdots \to Z_{\tau_i} \to \cdots \to Z_1$ (where $\tau_i = i\delta$ for a time step $\delta > 0$ or τ_i is a specific hitting time): this would be the natural approach for IPS selection-mutation procedure of [DG05, CDLL06][CDG11, Section 6] or splitting methods of [VV91, DD09], where in both cases Z is required to be a Markov process and the skeleton $(Z_{\tau_i})_{i\geqslant 1}$ to be a Markov chain. In our approach, we ignore the dynamics of Z and treat the rare-event issue written on a dynamic model as a *static problem on the space of paths*. Our two schemes based on shaking transformation will produce a Markov chain on the path space $\mathbb{C}([0,1],\mathbb{R})$.

Doing so, we avoid Markovian assumptions on Z and this enables very easily to consider path-dependent stochastic differential equation, like HJM framework for interest rates, see [HJM92], or stochastic evolution equations on Banach spaces, see e.g. [PZ14]. Handling discontinuous paths in $\mathbb{S} = \mathbb{D}([0,1],\mathbb{R})$ includes usual models and important applications for ruin in insurance [AA10], queuing theory in communication networks [Rob03] etc. Our approach can handle problems with infinite-time horizon as well.

2. Reversible shaking transformation and algorithms.

2.1. Shaking transformation and invariance of conditional distribution. Since $\mathbb{P}(X \in A) > 0$, in view of the inclusion (1.2) we have $\mathbb{P}(X \in A_k) > 0$ for any k, which justifies the decomposition

$$\mathbb{P}(X \in A) = \prod_{k=1}^{n} \mathbb{P}(X \in A_k | X \in A_{k-1}). \tag{2.1}$$

In this section, the standing assumption is

(K) There is a measurable mapping $K : \mathbb{S} \times \mathbb{Y} \mapsto \mathbb{S}$, where $(\mathbb{Y}, \mathcal{Y})$ is a measurable space, and a $\mathbb{Y}$-valued random variable Y independent of X such that the following identity in distribution holds:

$$(X, K(X,Y)) \stackrel{\mathrm{d}}{=} (K(X,Y), X). \tag{2.2}$$

To simplify notation when unambiguous, we simply write $\mathcal{K}(\cdot) := K(\cdot, Y)$. Identity (2.2) implies that X and $\mathcal{K}(X)$ have the same distribution: in our algorithms, $\mathcal{K}$ will serve to build Markov chains with invariant distribution given by that of X and that of X restricted to A_k (see Definitions 2.3 and 2.5). The exact form of K and Y is specific to the model at hand, examples are given in Section 3. We expect the random transformation $X \mapsto \mathcal{K}(X)$ to slightly modify values of X while preserving its distribution: this motivates the label of *shaking transformation*. Based on $\mathcal{K}(\cdot)$, for

each intermediate subset we define a shaking transformation with rejection as follows.

DEFINITION 2.1. *Let $k \in \{0, 1, \cdots, n-1\}$. Under* **(K)**, *define*

$$M_k^{\mathcal{K}} : \begin{cases} \mathbb{S} \times \mathbb{Y} \to \mathbb{S}, \\ (x, y) \mapsto K(x, y)\mathbf{1}_{K(x,y)\in A_k} + x\mathbf{1}_{K(x,y)\notin A_k}. \end{cases} \tag{2.3}$$

We set $\mathcal{M}_k^{\mathcal{K}}(.) := M_k^K(., Y)$ where Y is the generic random variable defined in **(K)**.

In [CDFG12] this kind of transformation is used to design an interacting particle algorithm for rare events related to random variables in $\mathbb{R}^d$. Here we generalize it to the general state space $\mathbb{S}$. Proposition 2.2 and Theorem 2.4 when $\mathbb{S} = \mathbb{R}^d$ have similar counterparts in [CDFG12] which proofs make use of explicit Markov transition kernels. Here in order to generalize, we follow a different presentation, seemingly more adapted to **(K)** and to our general state space setting.

PROPOSITION 2.2. *Let $k \in \{0, 1, \cdots, n-1\}$. The distribution of X conditionally on $\{X \in A_k\}$ is invariant w.r.t. the random transformation $\mathcal{M}_k^{\mathcal{K}}$: i.e. for any bounded measurable $\varphi : \mathbb{S} \to \mathbb{R}$ we have*

$$\mathbb{E}\big(\varphi(\mathcal{M}_k^{\mathcal{K}}(X))|X \in A_k\big) = \mathbb{E}\big(\varphi(X)|X \in A_k\big). \tag{2.4}$$

The above equality still holds if $\varphi(x)$ is replaced by $\varphi(x, U)$ where U is a random variable independent of X and Y (defining $\mathcal{M}_k^{\mathcal{K}}$).

Proof. From Definition 2.1 and **(K)**, we write that $\mathbb{E}\left(\varphi(\mathcal{M}_k^{\mathcal{K}}(X))\mathbf{1}_{X\in A_k}\right)$ equals

$$\begin{aligned} &\mathbb{E}\left(\varphi(\mathcal{K}(X))\mathbf{1}_{X\in A_k}\mathbf{1}_{\mathcal{K}(X)\in A_k}\right) + \mathbb{E}\left(\varphi(X)\mathbf{1}_{X\in A_k}\mathbf{1}_{\mathcal{K}(X)\notin A_k}\right) \\ &= \mathbb{E}\left(\varphi(X)\mathbf{1}_{\mathcal{K}(X)\in A_k}\mathbf{1}_{X\in A_k}\right) + \mathbb{E}\left(\varphi(X)\mathbf{1}_{X\in A_k}\mathbf{1}_{\mathcal{K}(X)\notin A_k}\right) = \mathbb{E}\left(\varphi(X)\mathbf{1}_{X\in A_k}\right). \end{aligned}$$

The equality (2.4) readily follows. The extension to random $\varphi(\cdot, U)$ is similar. □

2.2. Application to IPS algorithm. We are now in a position to put the rare event probability estimation problem in the framework of interacting particles system, which evolves according to the following dynamics.

DEFINITION 2.3. *We define a $\mathbb{S}$-valued Markov chain $(X_i)_{0\leqslant i\leqslant n-1}$, as follows:*

$$X_0 \overset{\mathrm{d}}{=} X, \qquad X_i := \mathcal{M}_i^{\mathcal{K}}(X_{i-1}) = M_i^K(X_{i-1}, Y_{i-1}) \quad \text{for } 1 \leqslant i \leqslant n-1, \tag{2.5}$$

where $(Y_i)_{0\leqslant i\leqslant n-2}$ is a sequence of independent copies of Y (defined in **(K)**) *and independent of X_0.*

The IPS interpretation will follow from the next result.

THEOREM 2.4. *Let $k \in \{1, \cdots, n\}$. We have:*

$$\mathbb{P}\left(X \in A_k\right) = \mathbb{E}\left(\prod_{i=0}^{k-1} \mathbf{1}_{A_{i+1}}(X_i)\right). \tag{2.6}$$

For any bounded measurable function $\varphi : \mathbb{S} \to \mathbb{R}$ we have

$$\mathbb{E}\left(\varphi(X)|X \in A_k\right) = \frac{\mathbb{E}\left(\varphi(X_{k-1})\prod_{i=0}^{k-1} \mathbf{1}_{A_{i+1}}(X_i)\right)}{\mathbb{E}\left(\prod_{i=0}^{k-1} \mathbf{1}_{A_{i+1}}(X_i)\right)}. \tag{2.7}$$

The above formula is still valid if $\varphi(x)$ is replaced by $\varphi(x,U)$ (as in Proposition 2.2) where U is a random variable independent of $(X, X_0, Y_0, \ldots, Y_{k-2})$.

Proof. We first establish (2.7) by induction on k. We start with $k=1$: obviously

$$\mathbb{E}\left(\varphi(X,U)|X\in A_1\right) = \frac{\mathbb{E}\left(\varphi(X,U)\mathbf{1}_{A_1}(X)\right)}{\mathbb{P}\left(X\in A_1\right)} = \frac{\mathbb{E}\left(\varphi(X_0,U)\mathbf{1}_{A_1}(X_0)\right)}{\mathbb{E}\left(\mathbf{1}_{A_1}(X_0)\right)}.$$

Assume now that (2.7) holds for k, any function φ and any random variable U allowed, and let us prove (2.7) for $k+1$. By a slight abuse of notation, we still write $\varphi(x,U) = \varphi(x)$, where U is independent of $(X, X_0, Y_0, \ldots, Y_{k-1})$. We have

$$\begin{aligned}\mathbb{E}\left(\varphi(X_k)\prod_{i=0}^{k}\mathbf{1}_{A_{i+1}}(X_i)\right) &= \mathbb{E}\left(\varphi(\mathcal{M}_k^{\mathcal{K}}(X_{k-1}))\mathbf{1}_{A_{k+1}}(\mathcal{M}_k^{\mathcal{K}}(X_{k-1}))\prod_{i=0}^{k-1}\mathbf{1}_{A_{i+1}}(X_i)\right)\\ &= \mathbb{E}\left(\varphi(\mathcal{M}_k^{\mathcal{K}}(X))\mathbf{1}_{A_{k+1}}(\mathcal{M}_k^{\mathcal{K}}(X))|X\in A_k\right)\mathbb{E}\left(\prod_{i=0}^{k-1}\mathbf{1}_{A_{i+1}}(X_i)\right) \qquad (2.8)\end{aligned}$$

where we have applied the induction hypothesis. Then Proposition 2.2 yields

$$\begin{aligned}\mathbb{E}\left(\varphi(\mathcal{M}_k^{\mathcal{K}}(X))\mathbf{1}_{A_{k+1}}(\mathcal{M}_k^{\mathcal{K}}(X))|X\in A_k\right) &= \mathbb{E}\left(\varphi(X)\mathbf{1}_{A_{k+1}}(X)|X\in A_k\right)\\ &= \mathbb{E}\left(\varphi(X)|X\in A_{k+1}\right)\mathbb{E}\left(\mathbf{1}_{A_{k+1}}(X)|X\in A_k\right). \qquad (2.9)\end{aligned}$$

Another application of Proposition 2.2 and of (2.8) with $\varphi\equiv 1$ shows that

$$\begin{aligned}\mathbb{E}\left(\mathbf{1}_{A_{k+1}}(X)|X\in A_k\right) &= \mathbb{E}\left(\mathbf{1}_{A_{k+1}}(\mathcal{M}_k^{\mathcal{K}}(X))|X\in A_k\right)\\ &= \frac{\mathbb{E}\left(\prod_{i=0}^{k}\mathbf{1}_{A_{i+1}}(X_i)\right)}{\mathbb{E}\left(\prod_{i=0}^{k-1}\mathbf{1}_{A_{i+1}}(X_i)\right)}. \qquad (2.10)\end{aligned}$$

Substituting (2.10) into (2.9) and (2.8) gives the equality (2.7) for $k+1$. Lastly, the proof of (2.6) now follows easily from (2.10) and (2.1). □

By the above theorem, the rare event probability is written in form of an unnormalized Feynman-Kac measure for interacting particle systems (see [Del04, DG05] for detailed discussions). This enables to use numerical algorithms for estimating it. In general, an interacting particle (a.k.a. genetic genealogical) algorithm provides a way to estimate

$$\mathbb{E}\left(f(X_0,\ldots,X_n)\prod_{i=0}^{n-1}G_i(X_i)\right)$$

where f and G_i are bounded and $(X_i)_{0\leqslant i\leqslant n}$ is a Markov chain. In view of (2.6) with $k=n$ the rare event probability corresponds to $f\equiv 1$ and $G_i(\cdot) = \mathbf{1}_{A_{i+1}}(\cdot)$ and the corresponding Markov chain is defined in Definition 2.3.

The detailed description of interacting particle algorithms can be found in [DG05, CDG11] (see also [CDFG12] for $\mathbb{S}=\mathbb{R}^d$). The adaptation to our rare event problem in a general state space $\mathbb{S}$ is made without difficulty. As in [CDG11], we introduce an extra rejection parameter $\alpha\in[0,1]$ which increases the independent resampling effect: in [DG05, CDFG12], $\alpha=1$.

The algorithm below generates at each time $i\in\{0,\ldots,n-1\}$ a sample of M elements in $\mathbb{S}$, whose empirical measure approximates the distribution of X conditionally on $\{X\in A_i\}$. We denote by $(Y_i^{(m)} : 1\leqslant m\leqslant M, 0\leqslant i\leqslant n-2)$

(resp. $(U_i^{(m)} : 1 \leqslant m \leqslant M, 0 \leqslant i \leqslant n-2)$) a sequence of independent copies of Y from Assumption **(K)** (resp. of a uniformly distributed random variable on $[0,1]$).

Initialization:
Draw $(X_0^{(M,m)}, m = 1, \cdots, M)$ which are M independent copies of X
$p_0^{(M)} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}_{A_1}(X_0^{(M,m)})$
for $i = 0$ until $n-2$ **do**
  $I_i = \{m \in \{1, \ldots, M\} \text{ s.t. } X_i^{(M,m)} \in A_{i+1}\}$
  for $m = 1$ until M **do**
    Selection step:
    if $U_i^{(m)} < \alpha$ and $X_i^{(M,m)} \in A_{i+1}$ **then**
      $\hat{X}_i^{(M,m)} = X_i^{(M,m)}$
    else
      $\hat{X}_i^{(M,m)} = X_i^{(M,\hat{m})}$ where $\hat{m}$ is drawn independently of everything else and uniformly in the set I_i
    end
    Mutation step:
    $X_{i+1}^{(M,m)} = M_{i+1}^K(\hat{X}_i^{(M,m)}, Y_i^{(m)})$
  end
  $p_{i+1}^{(M)} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}_{A_{i+2}}(X_{i+1}^{(M,m)})$
end
Result: $p^{(M)} = \prod_{i=0}^{n-1} p_i^{(M)}$

Algorithm 1: Interacting Particle System algorithm

In the case $\alpha = 1$, the above algorithm takes the same form as that in [CDFG12, Section 2] for random variables in $\mathbb{R}^d$. The difference in our work lies in the general state space for which Feynman-Kac formulae (Theorem 2.4) are nevertheless valid due to the assumption **(K)**: once obtained these formulae, deriving the above IPS algorithm follows a standard routine whose details are left to the reader. Applying this algorithm to stochastic processes is exposed in Sections 3 and 4. The convergence properties of this algorithm are postponed to Subsection 2.4.1.

2.3. Application to POP algorithm. In Proposition 2.2, we have seen that the distribution of X conditionally on $\{X \in A_k\}$ is invariant with respect to $\mathcal{M}_k^{\mathcal{K}}$. This property allows us to put the problem of computing $\mathbb{P}(X \in A_{k+1} | X \in A_k)$ or $\mathbb{E}(\varphi(X) | X \in A_k)$ in the ergodic Markov chain setting and therefore to compute $\mathbb{P}(X \in A)$ as a consequence of (2.1). Before entering into details, we recall that provided that $(Z_i)_{i \geqslant 0}$ is a Markov chain on a measurable space with a unique invariant distribution π, the Birkhoff's theorem for ergodic Markov chains gives

$$\frac{1}{N}\sum_{i=0}^{N-1} f(Z_i) \underset{N \to +\infty}{\longrightarrow} \int f \mathrm{d}\pi \qquad a.s. \tag{2.11}$$

for π-a.e. starting point Z_0. Here f is a bounded (or π-integrable) measurable function. See [MT09, Chapter 17] or [DMS14, Chapter 7].

For each k we define a Markov chain as follows.

Definition 2.5. *For each $k = 0, \ldots, n-1$, given a starting point $X_{k,0}$, define*

$$X_{k,i} := \mathcal{M}_k^{\mathcal{K}}(X_{k,i-1}) = M_k^K(X_{k,i-1}, Y_{k,i-1}) \quad \text{for } i \geqslant 1 \tag{2.12}$$

where $(Y_{k,i})_{i\geqslant 0}$ *is a sequence of independent copies of* Y *(defined in* **(K)***) and independent of* $X_{k,0}$.

We assume additionally that the sequences $((Y_{k,i})_{i\geqslant 0} : 0 \leqslant k \leqslant n-1)$ are independent. The above process $X_{k,.}$ is a Markov chain in $\mathbb{S}$ and one invariant measure is the distribution of X conditionally on $\{X \in A_k\}$. Then, provided that this is the unique invariant measure, one can use the approximation (as $N \to +\infty$)

$$\mathbb{E}\left(\varphi(X)|X \in A_k\right) \approx \frac{1}{N}\sum_{i=0}^{N-1} \varphi(X_{k,i}), \tag{2.13}$$

which for $\varphi \equiv \mathbf{1}_{A_{k+1}}$ yields an approximation of $\mathbb{P}\left(X \in A_{k+1}|X \in A_k\right)$ and therefore of the rare event probability $\mathbb{P}\left(X \in A\right)$. Observe that each conditional probability is computed separately, in parallel for each A_k, on a single path: this gives the reason why we call this method POP for Parallel One-Path. Furthermore, these conditional probabilities can be estimated independently, by taking independent initializations (as defined below), i.e. by restarting the initialization from the beginning for each step k with negligible extra time cost since n is usually small. Both the separate and independent evaluations of conditional probabilities are nice properties of POP, and are not shared with other existing algorithms to our knowledge.

The following algorithm evaluating $\mathbb{P}\left(X \in A\right)$ gives a way to automatically initialize each step.[2]

Initialization:
$X_{0,0}$ is a copy of X
for $k = 0$ until $n-1$ **do**
  for $i = 1$ until $N-1$ **do**
    $X_{k,i} = M_k^K(X_{k,i-1}, Y_{k,i-1})$
  end
  $p_k^{(N)} = \frac{1}{N}\sum_{i=0}^{N-1} \mathbf{1}_{A_{k+1}}(X_{k,i})$
  $i_k = \arg\min\{j : X_{k,j} \in A_{k+1}\}$
  $X_{k+1,0} = X_{k,i_k}$
end
Result: $p^{(N)} = \prod_{k=0}^{n-1} p_k^{(N)}$

Algorithm 2: Parallel One-Path algorithm

As previously mentioned, the n steps are almost separated, except for initializations. Thus our POP algorithm can be easily parallelized on different processors: for instance, one can use a preliminary run to get all the initial positions in different subsets/levels. Then all the ergodic time-averages are performed in parallel. We could even use the same copy of Y throughout the different levels to save time used in the generation of random variables Y.

Besides, this algorithm can also serve for estimating $\mathbb{E}(\varphi(X)|X \in A)$ using the Markov chain $(X_{n,.})$ and the approximation (2.13). This should even be less time-consuming than computing $\mathbb{P}(X \in A)$ since we only need to get a starting point satisfying $X \in A$ and then do POP once at $k = n$ to obtain an empirical distribution of $X|X \in A$.

Lastly, observe that increasing the accuracy of POP algorithm is elementary since it suffices to keep on simulating the n Markov chains $((X_{k,.}) : 0 \leqslant k \leqslant n-1)$ until a

[2]Since the initialization is not done with the stationary distribution, in all our numerical examples, we use 1% percent burn-in time to reduce its impact.

larger time horizon N'. This is a significant difference with IPS, for which increasing accuracy implies increasing M and thus resimulating all the M particles system from the beginning (because of interactions).

2.4. Convergence analysis for both algorithms.

2.4.1. Convergence of IPS. Convergence of the IPS algorithm for estimating unnormalized Feynman-Kac measure is well studied in the literature, as the number of particles $M \to +\infty$: under various hypotheses, are proved the law of large number, central limit theorem (at rate $\sqrt{M}$) and non-asymptotic error estimation (fixed M).

Regarding Algorithm 1, it is known that the estimator is unbiased. A non-asymptotic variance control is given in [CDG11]. Since in our algorithm the number of intermediate levels are usually not large, we don't need the assumption $(M)_m$ or $(\hat{M})_m$ in [CDG11] and we can get similar results to Lemma 4.1 and Lemma 4.3 in [CDG11] by only assuming that the following quantity $\hat{\delta}_k$ is finite for each $k = 0, 1, \cdots, n-1$

$$\hat{\delta}_k := \sup_{(x,y)\in A_{k+1}^2} \frac{\mathbb{P}(\mathcal{K}(x) \in A_{k+2}) + \mathbf{1}_{A_{k+2}}(x)\mathbb{P}(\mathcal{K}(x) \notin A_{k+1})}{\mathbb{P}(\mathcal{K}(y) \in A_{k+2}) + \mathbf{1}_{A_{k+2}}(y)\mathbb{P}(\mathcal{K}(y) \notin A_{k+1})} < +\infty \tag{2.14}$$

where by convention $A_{n+1} = A_n$. We adapt [CDG11, Corollary 5.2] to our setting.

THEOREM 2.6. *Under the assumption that all $\hat{\delta}_k$'s are finite, we have the following non-asymptotic control*

$$\mathbb{E}\left(\left|\frac{p^{(M)}}{p} - 1\right|^2\right) \leqslant \frac{4}{M}\left(\sum_{s=0}^{n-1} \frac{\Delta_s}{\mathbb{P}\left(X \in A_{s+1} | X \in A_s\right)} + 1\right) \tag{2.15}$$

where $\Delta_s = \prod_{k=s}^{n-1} \hat{\delta}_k$.

Proof. We follow very closely the proof of [CDG11]. Actually in our setting, we avoid their assumptions $(M)_m$ or $(\hat{M})_m$ which role is partly to get better estimates as n is large: we thus just emphasize how to get rid of these assumptions in our work. Firstly, we easily check that $\hat{\delta}_k$ defined in their assumption $(\hat{H})_m$ is the one given in our theorem. Secondly with this estimate at hand, we can prove that (using notation of their Equation (4.3))

$$\sup_{x,y\in A_{k+1}^2} \frac{\hat{Q}_{k,n}(1)(x)}{\hat{Q}_{k,n}(1)(y)} \leqslant \Delta_k.$$

Therefore, the upper bound on the r.h.s. of (4.5) in their Lemma 4.3 becomes Δ_k (by noticing that their $\tilde{\delta}_k = 1$). Lastly, the rest of the proof is similar in that the above estimate propagates to their Corollary 5.2 in the form of our inequality (2.15). □

The upper bound of Theorem 2.6 is useful to appropriately choose the shaking transformation in order to make the error smaller. First, obviously we have $\hat{\delta}_k \leqslant \sup_{y\in A_{k+1}} \frac{1}{\mathbb{P}(\mathcal{K}(y)\in A_{k+2})}$. In case of slight shaking, $\mathcal{K}(y)$ will differ little from y, so the probability of going from A_{k+1} to A_{k+2} is small and $\hat{\delta}_k$ is large. Conversely, in case of strong shaking and since A_{k+2} is expected to be small, $\mathcal{K}(y)$ will be very likely to exit A_{k+2}, resulting in a large $\hat{\delta}_k$. Hence choosing an intermediate shaking force is presumably the best choice, see later numerical experiments.

Finally, the upper bound (2.15) is rather qualitative and can not be seemingly quantitatively computed except in very special cases (like Gaussian distribution). More general cases are left to further research.

2.4.2. Convergence of POP. (2.11) with its assumptions gives

$$\frac{1}{N}\sum_{i=0}^{N-1} \mathbf{1}_{A_{k+1}}(X_{k,i}) \underset{N\to+\infty}{\longrightarrow} \mathbb{P}(X \in A_{k+1}|X \in A_k) \qquad a.s. \tag{2.16}$$

for a.e. starting point $X_{k,0}$.

The convergence of ergodic theorem has been much studied in the literature, with results like almost sure convergence, asymptotic and non-asymptotic fluctuations, see for instance [MT09, DMS14]. Here we apply the recent work [LMN13, Theorem 3.1] in our rare event setting.

THEOREM 2.7. *For each k in $\{0, , \cdots, n-1\}$, assume that $(A_k, d_{\mathbb{S}})$ is a Polish space and that the Markov chain $(X_{k,i})_{i\geqslant 0}$ is π_k-irreducible and Harris recurrent, where π_k is the distribution of X conditionally on $\{X \in A_k\}$. If in addition the "small set" condition holds: "there exists a Borel set $F_k \subset A_k$ of positive π_k measure, a positive number $\beta_k > 0$ and a probability measure ν_k such that $P_k(x,\cdot) \geqslant \beta_k \nu_k(\cdot), \forall x \in F_k$" where $P_k(.,.)$ is the transition kernel of $X_{k,.}$, then there exists a constant C_k depending on the model such that*

$$\mathbb{E}\left(\left(p_k^{(N)} - \mathbb{P}(X \in A_{k+1}|X \in A_k)\right)^2\right) \leqslant \frac{C_k}{N}.$$

For application in our rare event examples, the "Polish assumption" is usually satisfied when we consider the space of continuous functions $\mathbb{C}([0,T],\mathbb{R}^d)$ $(T > 0)$ (example of Subsection 4.1), or the space of càdlàg functions $\mathbb{D}([0,T],\mathbb{R}^d)$ (jump processes in insurance, queueing system and Hawkes process in Subsections 4.2-4.3-4.5) or $\mathbb{R}^{\mathbb{N}}$, see [Bil99] for details. The random graph example in Subsection 4.4 is associated to a finite space and the "Polish assumption" is thus trivial.

But verification of the small set condition is more difficult. In the random graph example, this condition is satisfied obviously since the state space is finite. In general, extra work is still needed to verify the small set condition in each particular example.

Finally recall that our estimated rare event probability $p^{(N)}$ is the product of all $p_k^{(N)}$'s. Since we have already error control for each $p_k^{(N)}$ and since these quantities are bounded, by easy computations we can establish that the convergence rate is also $\sqrt{N}$ for the estimation of rare event probability.

3. Construction of shaking transformation. In order to make previous algorithms applicable, we now provide reversible shaking transformations in various situations (for random variables and random processes). Of course, Metropolis-Hastings and Gibbs type transformations are natural candidates but our aim is to provide other transformations that are presumably more suitable to the model under consideration. Recall that one has to exhibit a shaking map $K(\cdot,\cdot)$ and a random variable Y such that $(X, K(X,Y)) \overset{\mathrm{d}}{=} (K(X,Y), X)$.

3.1. Gaussian variable/process and stochastic differential equation (SDE in short) driven by Brownian motion. For a standard Gaussian variable $X := G$ in $\mathbb{R}^d$, a simple shaking transformation is

$$K(G, G') = \rho G + \sqrt{1-\rho^2} G'$$

with $\rho \in (-1, 1)$ and $Y := G'$ is a independent copy of G. Figure 1 gathers two graphs of 100000 independent simulations of $(G, \mathcal{K}(G))$ with their respective marginal

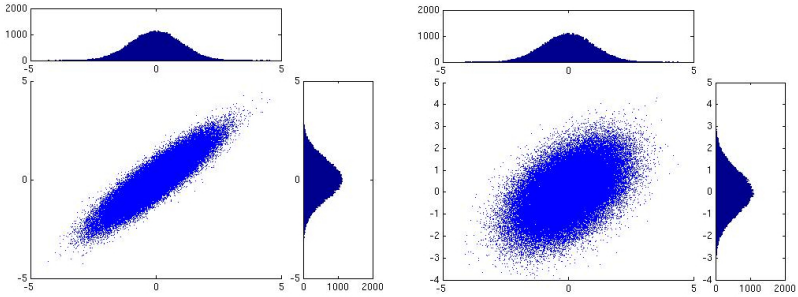

FIG. 1. *Shaking Gaussian variables in dimension 1, with* $\rho = 0.9$ *(left) and* $\rho = 0.5$ *(right)*

histograms (of course close to the Gaussian distribution). The larger the value of ρ, the slighter the shaking, the closer the points to the diagonal.

The same linear transformation works for Gaussian processes $X := (G_t)_{0 \leqslant t \leqslant T}$ ($T > 0$ fixed), with zero mean and any co-variance function. In the case where X is a d-dimensional Brownian motion, one can take slightly more general transformation based on Wiener integrals: namely, for the i-th component of $\mathcal{K}(G)$, take a measurable function $\rho_i : [0,T] \mapsto [-1,1]^d$ with $|\rho_{i,t}| \leqslant 1$ and set

$$\mathcal{K}_i(G) = \left(\int_0^t \rho_{i,s} \cdot \mathrm{d}G_s + \int_0^t \sqrt{1 - |\rho_{i,s}|^2} \mathrm{d}G'_{i,s} \right)_{0 \leqslant t \leqslant T} \tag{3.1}$$

where $G' = (G'_1, \ldots, G'_d)$ is another independent Brownian motion in $\mathbb{R}^d$. Provided that the matrix $\rho_t = (\rho_{1,t}, \ldots, \rho_{d,t})$ is symmetric and $\rho_i \cdot \rho_j \equiv 0$ for all $i \neq j$, this transformation satisfies **(K)**.

With this tool at hand, it is then straightforward to define reversible shaking transformations of solution to a stochastic differential equation of the form

$$\mathrm{d}Z_t = b(t, Z_t)\mathrm{d}t + \sigma(t, Z_t)\mathrm{d}G_t, \quad Z_0 \text{ independent of } G, \tag{3.2}$$

where coefficients b and σ fulfill appropriate smoothness and growth conditions in order to have a unique strong solution [RY99]. Setting $X = (Z_t)_{0 \leqslant t \leqslant T}$ it suffices to define $\mathcal{K}(X)$ as the (strong) solution of $\mathrm{d}Z'_t = b(t, Z'_t)\mathrm{d}t + \sigma(t, Z'_t)\mathrm{d}K(G, G')_t, Z'_0 = Z_0$ where the shaken Brownian motion $K(G, G') = \mathcal{K}(G)$ is defined in (3.1): this procedure satisfies **(K)**. This will be applied to the example of Ornstein-Uhlenbeck process in Subsection 4.1. Observe that this method can be directly extended to non-Markovian equations driven by Brownian motion.

3.2. Poisson variable and compound Poisson process. For a Poisson variable $X := P \overset{\mathrm{d}}{=} \mathtt{Poisson}(\lambda)$ with parameter $\lambda > 0$, a possible transformation is

$$K(P, [\mathtt{Bin}(P, 1-p), \mathtt{Poisson}(p\lambda)]) = \mathtt{Bin}(P, 1-p) + \mathtt{Poisson}(p\lambda)$$

where $p \in (0,1)$, using extra independent Binomial and Poisson random variables, see [Kin93, Chapter 5].

Invoking the same reference, the above decomposition holds also for compound Poisson process (CPP in short) with parameter (λ, μ), i.e. $X := (P_t)_{0 \leqslant t \leqslant T}$ where $P_t = \sum_{k=1}^{N_t} J_k$ where N is a standard Poisson process with intensity λ and $(J_k)_k$ are

i.i.d with distribution μ. Let $p \in (0,1)$: by coloring at random the jumps of N in `red` with probability $1-p$ and in `green` with probability p, we can write $N_t = N_t^{\mathtt{r}} + N_t^{\mathtt{g}}$ and $X_t = X_t^{\mathtt{r}} + X_t^{\mathtt{g}}$, using obvious notations. Then $X^{\mathtt{r}}$ and $X^{\mathtt{g}}$ are two independent CPP with parameters $((1-p)\lambda, \mu)$ and $(p\lambda, \mu)$. Using an extra independent CPP Y distributed as $X^{\mathtt{g}}$, it is easy to check that the following transformation satisfies **(K)**:

$$K(X,Y) = (X_t^{\mathtt{r}} + Y_t)_{0 \leqslant t \leqslant T}. \tag{3.3}$$

In Subsection 4.3, we will use this shaking transformation for the example of queueing system with exponential inter-arrival time .

3.3. Other random variables and processes.

▷ *Gamma, Exponential, χ^2 distributions.* For a random variable $X = \Gamma$ with Gamma distribution $\mathtt{Gamma}(a,b)$ $(a>0, b>0)$ defined by $\mathbb{P}(\Gamma \in \mathrm{d}x) = c_a b^a x^{a-1} e^{-bx} \mathbf{1}_{x>0} \mathrm{d}x$ for a normalizing constant c_a, we can provide a simple transformation based on the beta-gamma algebra (see [CY12, Chapter 4]). Let $p \in (0,1)$: with the notation of **(K)**, take $Y = (\mathtt{Beta}(a(1-p), ap), \mathtt{Gamma}(ap, b))$ with two extra independent Beta and Gamma distributed random variables, and set

$$\mathcal{K}(\Gamma) = \Gamma\ \mathtt{Beta}(a(1-p), ap) + \mathtt{Gamma}(ap, b). \tag{3.4}$$

It satisfies **(K)**. In particular, in the case $a = 1$ we recover the case of exponential distribution $\mathtt{Exp}(b)$. Note also that shaking $\chi^2(k)$ distribution directly follows from the above since this distribution coincides with that of $2\mathtt{Gamma}(\frac{k}{2}, 1)$. Figure 2 represents 100000 independent simulations of $(\Gamma, \mathcal{K}(\Gamma))$ with their respective marginal histograms. The smaller the value of p, the slighter the shaking. On the plots, observe that Γ and $\mathcal{K}(\Gamma)$ have the same distribution (coherently with **(K)**).

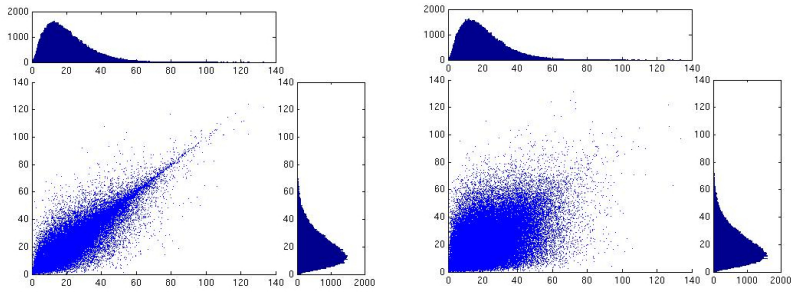

FIG. 2. *Shaking* $\mathtt{Gamma}(2.5, 0.12)$ *variables with* $p = 0.1$ *(left) and* $p = 0.5$ *(right)*

▷ *(s)-distribution.* For $X := T$ with the (s)-distribution given by $\mathbb{P}(T \in \mathrm{d}t) = \frac{1}{\sqrt{2\pi t^3}} \exp(-\frac{1}{2t}) \mathbf{1}_{t>0} \mathrm{d}t$ which represents the hitting time of level 1 by a standard Brownian motion, we have (at least) two shaking transformations. First, we can shake Brownian motion as previously exposed. Alternatively, we can use the well-known identity $T \overset{\mathrm{d}}{=} G^{-2}$ where G is a standard Gaussian random variable and apply the Gaussian shaking transformation. The underlying change of variable principle can be stated "informally" as follows: assume that $X \overset{\mathrm{d}}{=} f(Z)$ for some invertible function f and for a random variable Z having a shaking transformation $\mathcal{K}_Z$, then $\mathcal{K}_X(\cdot) = f(\mathcal{K}_Z(f^{-1}(\cdot)))$ defines a natural shaking transformation for X.

▷ *Uniform variable on* $[0,1]$. We can rely on the relation with exponential distribution to write $X := U \overset{\mathrm{d}}{=} \exp(-\texttt{Exp}(1))$. Let $p \in (0,1)$: in view of (3.4), the following transformation satisfies **(K)**,

$$\mathcal{K}(U) = U^{\texttt{Beta}(1-p,p)} \exp(-\texttt{Gamma}(p,1))$$

with extra independent Beta and Gamma random variables.

Now that we are in a position to shake uniform distribution, it is easy to shake any distribution on $\mathbb{R}$ having a continuous CDF function F since $F(X) \overset{\mathrm{d}}{=} \texttt{Unif}$. This is useful in the case F and F^{-1} are easily tractable. We do not list all the possibilities.

▷ *Other shakings for random variables.* In cases where explicit transformation is not available, we can use implicit transformation. Namely, assume for instance that $X := f(Z_1, \cdots, Z_n)$ with independent $(Z_i)_i$, which serves to simulate X through the simulation of $(Z_i)_{1 \leqslant i \leqslant n}$, and suppose that each Z_i has an explicit shaking transformation. Then the implicit shaking transformation for X is

$$\mathcal{K}(X) = f(\mathcal{K}_1(Z_1), \cdots, \mathcal{K}_n(Z_n))$$

where the exact expression of $\mathcal{K}_i$ may be different according to the type of random variables Z_i and each shaking is made independently of the others. For example, this can be applied to X having Beta distribution because of the identity $\texttt{Beta}(a,b) \overset{\mathrm{d}}{=} \frac{\texttt{Gamma}(a,1)}{\texttt{Gamma}(a,1)+\texttt{Gamma}(b,1)}$ with independent Gamma distributions.

3.4. Other variations on the shaking.

▷ *Randomized shaking.* Actually, in the previous examples $K(\cdot,\cdot)$ is often written as $K_\theta(\cdot,\cdot)$ for a parameter θ serving to tune the shaking force. A first remark is that instead of fixing the parameter value of θ, one can also randomize it, which gives rise to another reversible shaking transformation.

LEMMA 3.1. *Assume that $K_\theta(\cdot, Y)$ satisfies* **(K)** *for any θ in a measurable space Θ and that $K_\cdot(\cdot,\cdot)$ defines a measurable function from $\Theta \times \mathbb{S} \times \mathbb{Y}$ into $\mathbb{S}$. Let T be any Θ-valued random variable independent of Y, then $K_T(\cdot, Y)$ satisfies* **(K)**.

The proof is easy and left to the reader. As a consequence, all the shaking transformations presented before can be generalized with a random parameter. This randomization technique will be seen useful in the example of Section 4.6.

▷ *Partial shaking.* When the random variable X is built on several independent random variables, it may be relevant to shake only some of them. For instance, consider a general pure jump process (including CPP or renewal process), where $A = (A_n)_{n \geqslant 1}$ represents the inter-arrival times and $B = (B_n)_{n \geqslant 1}$ represents the jump sizes, A and B being independent: the shaking transformation may concern both A and B, or only A (the jump times), or only B (the jump sizes). These alternatives are tested in the subsequent examples on insurance and queueing system. Similarly, for a SDE driven by both Brownian motion and another independent Levy process, we can shake the first driving process or the second, or both.

Another strategy is to apply *randomized partial shaking*. For a model of the form $X := f(Z_i, 1 \leqslant i \leqslant n)$ with independent $(Z_i)_i$, when n large or $n = +\infty$ we can reduce the computational cost by picking at random a subset of coordinates and only shake independently the corresponding random variables. The property of reversible shaking transformation is preserved owing to Lemma 3.1. This method will be used in the random graph example of Subsection 4.4.

4. Numerical examples. We now aim at comparing the numerical performances of IPS and POP algorithms, using shaking transformations presented in Section 3. The examples below are chosen according to their importance in applications and also because they are numerically challenging, moreover for some of them we can compute benchmark values using importance sampling techniques.

We have not optimized the choice of intermediate levels in these examples: only very rough preliminary runs are done to make all conditional probabilities of the same magnitude. The design and analysis of adaptive algorithms in the vein of [CDFG12] is left to future research. However, observe that adding extra intermediate levels in POP can be done directly, without changing the estimation for other levels, thus preliminary runs are not necessary for POP; this is a significant advantage compared to IPS where one would to resimulate the whole particle system.

Another important remark concerns the memory. While for IPS one has to store all the particles (due to interactions), for POP only one particle per level needs to be stored which constitutes a large memory save.

Lastly we report the means and standard deviations of the algorithms outputs which are evaluated empirically by several runs (say 50 or 100).

Usually in POP method, one could use some burn-in time to reduce influence of the initial position of Markov chain. In all examples, we use the first 1 percent transitions as the burn-in time, except for the first level where no burn-in time is needed.

4.1. One dimensional Ornstein-Uhlenbeck (OU in short) process driven by Brownian motion. The OU process we consider is given by

$$\mathrm{d}Z_t = -Z_t \mathrm{d}t + \mathrm{d}G_t, \quad Z_0 = 0, \tag{4.1}$$

where G is a standard Brownian motion: it is in the form (3.2) and in the sequel, we apply the Brownian motion shaking (3.1) with constant ρ.

Actually, the following rare events are described in terms of the path of $(Z_t)_{0\leqslant t\leqslant T}$ with $T = 1$: instead of an exact simulation, we simply use an Euler scheme $\tilde{Z}$ with time step $h := T/m$ for $m = 100$ and piecewise constant path approximation between the times $t_l := lh$. This discretization scheme does not alter significantly the performance of IPS and POP algorithms.

4.1.1. Maximum of OU process. Here the rare event is given by $\{\max_{0\leqslant l\leqslant m} \tilde{Z}_{t_l} > L\}$ with $L = 3.6$. Because of the mean reverting effect, the related probability is rather small. By 10^7 direct Monte Carlo simulations with importance sampling technique under the new probability $\mathrm{d}\mathbb{Q} = \exp(aG_T - \frac{1}{2}a^2T)\mathrm{d}\mathbb{P}$ where $a = 5$, we derive a 99% confidence interval of the requested probability $[0.977, 1.004] \times 10^{-7}$.

In (1.2) we take $n = 5$ intermediate sets associated to the levels $L_k = L\sqrt{\frac{k}{n}}, k = 1, \cdots, n$. In the experiments we report, we change the values of ρ, α N and M.

Results. For the IPS and POP algorithms, we take respectively $M = 100000$ and $N = 100000$ so that the computational effort is similar. The following tables show results for different values of (α, ρ) for IPS and of ρ for POP. Output statistics (mean, standard deviation) are computed with 50 algorithm runs.

IPS	$\alpha = 1$			$\alpha = 0.5$			$\alpha = 0$		
	mean	std	std/mean	mean	std	std/mean	mean	std	std/mean
$\rho = 0.9$	1.06e-07	5.12e-08	0.48	1.01e-07	3.67e-08	0.36	1.01e-07	3.94e-08	0.39
$\rho = 0.75$	9.51e-08	2.15e-08	0.22	9.81e-08	1.76e-08	0.18	9.98e-08	2.46e-08	0.25
$\rho = 0.5$	9.32e-08	9.42e-08	1.01	7.32e-08	9.18e-08	1.25	8.27e-08	1.18e-07	1.42

POP	mean	std	std/mean
$\rho = 0.9$	9.80e-08	6.74e-09	0.07
$\rho = 0.75$	1.00e-08	9.52e-09	0.10
$\rho = 0.5$	1.05e-07	2.78e-08	0.27

We first notice that the probability is estimated coherently regarding the benchmark value (obtained by importance sampling). We note that POP has a better performance compared to IPS (see the column std/mean), whatever the value of ρ is. Regarding the variance, we observe that the value of α (used for extra resampling) has no significant impact on IPS algorithm, while the value of ρ is important for both IPS and POP with respectively $\rho = 0.75$ and $\rho = 0.9$ as optimal values. The above standard deviations are comparable to the one using importance sampling, but our approaches have the advantage to work in a rather general setting.

4.1.2. Oscillation of OU process. Now the rare event is associated to a large oscillation of the OU process, i.e. we compute $\mathbb{P}\left(\max_{0\leqslant l\leqslant m} \tilde{Z}_{t_l} > L \text{ and } \min_{0\leqslant l\leqslant m} \tilde{Z}_{t_l} < -L\right)$ with $L = 1.6$. Remark that in this situation *standard* important sampling techniques with shifted Brownian motion do not work any more. By a crude Monte Carlo algorithm with 7×10^9 simulations, we obtain a 99% confidence interval equal to $[3.97, 4.37] \times 10^{-7}$.

In our IPS and POP approaches, we simply take $L_k = L\sqrt{\frac{k}{5}}$ for $k = 1, \ldots, 5$ and define intermediate events as $\{\max_{0\leqslant l\leqslant m} \tilde{Z}_{t_l} > L_k \text{ and } \min_{0\leqslant l\leqslant m} \tilde{Z}_{t_l} < -L_k\}$.

Results. In the following tables the empirical results of IPS and POP algorithms are computed over 100 experiments, respectively with $M = 100000$ and $N = 100000$.

IPS	$\alpha = 1$			$\alpha = 0.5$			$\alpha = 0$		
	mean	std	std/mean	mean	std	std/mean	mean	std	std/mean
$\rho = 0.9$	4.01e-07	1.23e-07	0.31	3.94e-07	1.08e-07	0.27	4.18e-07	1.08e-07	0.26
$\rho = 0.75$	4.10e-07	1.67e-07	0.41	4.12e-07	1.89e-07	0.46	4.20e-07	2.02e-07	0.48
$\rho = 0.5$	2.44e-07	4.76e-07	1.95	3.41e-07	9.89e-07	2.90	2.66e-07	4.61e-07	1.73

POP	mean	std	std/mean
$\rho = 0.9$	4.14e-07	2.68e-08	0.06
$\rho = 0.75$	4.18e-07	4.60e-08	0.11
$\rho = 0.5$	4.29e-07	1.26e-07	0.29

As before, both algorithms seemingly converge, with better results for POP (the std for POP is about 4-5 times smaller than for IPS). Here again, the value of α is not so crucial while the value of ρ has important impact on the variance. In all the following IPS algorithms, we fix α equal to 1 (i.e. we skip the resampling step).

In Figure 3, we show empirical variances of 100 experiments results for M and N respectively equal to 100000, 10000, 5000, 3000 and 2000. These variances are not perfectly estimated since we use only 100 runs, nevertheless we approximately obtain a linear convergence with respect to $1/M$ and $1/N$, as expected from theoretical results (see Theorems 2.6 and 2.7).

4.2. Insurance. The capital reserve of an insurance company is modeled by

$$R_t = x + ct - \sum_{k=1}^{N_t} Z_k$$

where x is the initial reserve, c is the premium rate, N is a Poisson process with intensity λ and $(Z_k)_k$ are amounts of claims in case of accident or natural disaster [AA10]. In the following example, we take $c = 1, \lambda = 0.005, x = 100, T = 1$ and

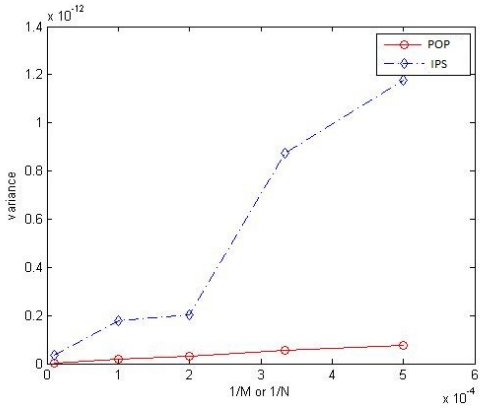

FIG. 3. *Variance for IPS and POP methods as a function of* $1/M$ *and* $1/N$ *respectively*

suppose $(Z_k)_k$ are Gamma variables with parameters $(a, b) = (2.5, 0.12)$. We aim at computing the probability of bankruptcy before T, i.e. $\mathbb{P}\left(\min_{0\leqslant t\leqslant T} R_t < 0\right)$.

Using Esscher transformation, we get the 99% confidence interval for this probability: $[1.042, 1.188] \times 10^{-6}$ through 10^5 Monte Carlo simulations under the new probability $\mathrm{d}\mathbb{Q} = \exp(\sum_{k=1}^{N_T} f(Z_k) - \int_{\mathbb{R}}(e^{f(y)} - 1)\lambda T\nu(\mathrm{d}y))\mathrm{d}\mathbb{P}$ where $f(y) = 0.09y$ and $\nu(\mathrm{d}y)$ is the probability measure of $\texttt{Gamma}(a, b)$. We can easily check that the distribution of Z_k is still of Gamma type under this new probability.

We take $n = 5$ intermediate levels, defined by $L_k = x(1 - (\frac{k}{5})^2)$ for $k = 1, \ldots, 5$.

IPS algorithm. We apply the partial shaking to the jump sizes (and not to the jump times), i.e. we shake all $(Z_k)_k$ with the shaking transformations for Gamma variables (with parameter p), then we get the following results ($M = 10000$, over 100 times experiments).

	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$
mean	1.25e-006	1.11e-006	1.01e-006	1.02e-006	1.15e-006	1.09e-006
std	2.82e-006	1.30e-006	6.46e-007	8.39e-007	5.15e-007	4.11e-007
std/mean	2.26	1.17	0.64	0.82	0.45	0.38

With the Poisson process decomposition shaking with parameter p (M =10000, over 100 times experiments), results become as follows.

	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$
mean	3.10e-006	2.02e-006	2.93e-006	8.63e-008	9.32e-007	2.08e-006
std	1.76e-005	1.39e-005	1.65e-005	1.32e-007	8.68e-006	1.42e-005
std/mean	5.68	6.85	5.62	1.53	9.32	6.81

We observe that Poisson shaking can not even produce a good mean value and that partial shaking on Gamma variables is much better than Poisson decomposition shaking. This can be explained as follows: in this particular insurance reserve example where there are very few jumps with important jump sizes, the Poisson shaking gives large perturbation of the system (opposite to the spirit of slight shaking), since by removing a jump time (and therefore the claim amount at this instant), this may completely change the situation of the company, from being close to bankruptcy to running with good profit.

Obviously, partial shaking involving Gamma variables doesn't cause this kind of problem since we keep every jump time and only modify claim amount. In this sense,

the Gamma shaking is more continuous and suits better this example.

Shaking all the inter-arrival and jump variables yields the following results (over 100 experiments with $M = 10000$), which gives larger variance than for Gamma shaking only, as expected.

	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$
mean	9.75e-07	9.35e-07	1.22e-06	2.97e-07	9.75e-07	1.10e-06
std	4.73e-06	3.63e-06	7.14e-06	5.90e-07	8.15e-06	9.80e-06
std/mean	4.85	3.89	5.87	1.99	8.36	8.94

POP algorithm. When using Poisson shaking or Gamma shaking for our POP algorithm, we have observed that both of them fail. The reason for Poisson shaking is similar to IPS case. As for Gamma shaking, the difference between IPS and POP is that, in IPS we sample M trajectories and we pick those with jumps, while in POP algorithm we have only one trajectory and (in this insurance example) the initial configuration may have no jump with a large probability, yielding that the output of POP algorithm is doomed to be 0.

To retrieve good convergence properties, we simply apply the shaking for inter-arrival and jump variables and we get the following results (over 100 experiments with $M = 10000$), which are slightly more accurate than IPS.

	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$
mean	1.14e-06	1.11e-06	1.12e-06	1.05e-06	1.12e-06	9.29e-07
std	5.08e-07	4.44e-07	4.80e-07	6.74e-07	8.24e-07	9.52e-07
std/mean	0.45	0.40	0.43	0.64	0.74	1.02

4.3. Queueing system. Suppose we have a 2-nodes Jackson network [Rob03, Chapter 4]. All the costumers arrive at node 1 and when they are served they go to node 2. The customers' arrival times are jump times of a Poisson process with intensity λ. The serving time at node 1 and at node 2 are respectively exponential variables with parameters μ_1 and μ_2. Our purpose is to compute the probability that at some time before T, the number of customers in the system reaches a fixed level K, i.e. $\mathbb{P}(\max_{0\leqslant t\leqslant T} M_t > K)$ where M_t denotes the number of customers in the system at time t.

Given the Poisson process N representing customers' arrival time, we define two compound Poisson processes

$$Z_t^A = \sum_{k=1}^{N_t} A_k, \qquad Z_t^B = \sum_{k=1}^{N_t} B_k,$$

to which we are to apply shaking transformations. Here A_k and B_k are respectively the serving times of k-th customer at node 1 and at node 2. We now claim that $\max_{0\leqslant t\leqslant T} M_t = \Phi((Z_t^A)_{0\leqslant t\leqslant T}, (Z_t^B)_{0\leqslant t\leqslant T})$ for a functional Φ, this representation will be the basis for our algorithms. To justify this, denote by a_k the arrival time of the k-th customer (i.e. the k-th jump of N). Then if we note by e_k^1 the instant when the service for k-th customer at node 1 is finished, we can find the following recursive relation: $e_{k+1}^1 = \max(a_{k+1}, e_k^1) + A_{k+1}$ with the initial condition $e_1^1 = a_1 + A_1$. Remark that the service finishing time at node 1 is the customer arrival time at node 2, we have the same recursive relation for e_k^2, the instants when service for k-th customer at node 2 is finished: $e_{k+1}^2 = \max(e_{k+1}^1, e_k^2) + B_{k+1}$ with the initial condition $e_1^2 = e_1^1 + B_1$. Then when the k-th customer enters the system, the number of customers in the system is $k - \#\{e_j^2 : e_j^2 < a_k\}$; since the maximal number of customers in the system is possibly reached only when a new customer enters the

system we have $\max_{0\leqslant t\leqslant T} M_t = \max_{0\leqslant k\leqslant N_T}(k - \#\{e_j^2 : e_j^2 < a_k\})$ which leads to our claim that the maximal number of customers in the system before T is determined by the two CPP's $(Z_t^A)_{0\leqslant t\leqslant T}$ and $(Z_t^B)_{0\leqslant t\leqslant T}$.

We take $\lambda = 0.5, \mu_1 = 1, \mu_2 = 1, T = 10$ and $n = 10$ intermediate levels defined as $L_k = K\sqrt{\frac{k}{n}}, k = 1, \cdots, n$. We set $K = 20$. For the benchmark value, we use an importance sampling method (with 10^7 simulations) based on Esscher transformation using the new probability $\mathrm{d}\mathbb{Q} = \exp(cN_T - (e^c - 1)\lambda T)\mathrm{d}\mathbb{P}$ where $c = 1.5$: the resulting 99% confidence interval for $\mathbb{P}(\max_{0\leqslant t\leqslant T} M_t > K)$ is $[4.6380, 5.1210] \times 10^{-10}$. The shaking transformation we use here is defined in (3.3), with different values of p.

Results. The following IPS and POP results are computed with $M = 10000$ and $N = 10000$ respectively, over 100 times experiments with each parameter.

IPS	$p = 0.1$	$p = 0.3$	$p = 0.5$
mean	4.35e-10	5.04e-10	5.58e-10
std	5.33e-10	4.26e-10	2.00e-09
std/mean	1.23	0.84	3.58

POP	$p = 0.1$	$p = 0.3$	$p = 0.5$
mean	4.93e-010	5.24e-010	5.27e-010
std	1.33e-010	2.09e-010	4.62e-010
std/mean	0.27	0.40	0.88

The POP algorithm provides more accurate results than IPS, and seemingly more stable as p is modified. If we only shake service times A and B instead of the Poisson process Z^A and Z^B (as in (3.3)), both algorithms fail to work, almost systematically the output of algorithm is 0. This is not surprising since by shaking the service time, we will never increase the number of clients in the system.

The POP method has been tested in the case of renewal process where inter-arrival and service times are uniformly distributed. The performances are good too.

4.4. Random graph. An Erdös-Rényi random graph [Bol01] is a graph with V vertices where every pair of vertices are connected with probability q, independently of the others. It constitutes a toy model for the study of social networks, epidemic... The graph is presented by the upper triangular matrix $X := (X_{ij})_{1\leqslant i<j\leqslant V}$ where $X_{ij} = 1$ if vertices i and j are connected, and $X_{ij} = 0$ otherwise. If vertices i, j and k are all connected to each other, they form a triangle. Thus the number of triangles in the graph is given by

$$T(X) := \sum_{1\leqslant i<j<k\leqslant V} X_{ij}X_{jk}X_{ik}.$$

We easily check that $\mathbb{E}(T(X)) = \frac{V(V-1)(V-2)}{6}q^3$ and as a rare event, we consider the deviation event $\{T(X) > \frac{V(V-1)(V-2)}{6}t^3\}$ for $t > q$. This problem has deserved recent interest in [CV11] with theoretical results and in [BHLN13] with numerical computations based on importance sampling techniques.

The total number of possible connections is $\frac{V(V-1)}{2}$ and may be rather large even for small graphs. In our case we take $V = 64$, $q = 0.35$ and $t = 0.4$: the corresponding estimation given in [BHLN13] is about $2.19e-06$. To reduce the complexity of IPS and POP algorithms, we use the technique of partial shaking, by picking randomly a proportion c of X_{ij} and shake them independently. Regarding the reversible shaking transformation of each Bernoulli random variable X_{ij}, the only possibility is described by a transition matrix $P(x, y)$ $(x, y \in \{0, 1\}^2)$ which satisfies the following condition

$$qP(1, 0) = (1 - q)P(0, 1),$$

i.e. $P(0, 1) = \frac{q}{1-q}P(1, 0)$. Since in this example $\frac{q}{1-q} < 1$, $P(1, 0)$ can be any value in $[0, 1]$ and it is parametrizing the force of shaking. The larger the value of $P(1, 0)$, the more important the change in the graph configuration.

Numerical results are performed with $n = 5$ intermediate levels given by $L_k = \frac{V(V-1)(V-2)}{6} t^3 (\frac{k}{5})^{\frac{1}{5}}$ with $k = 1, \dots, n$.

Results. First, we take $c = 10\%$ and statistics are computed over 50 algorithm experiments. For IPS and POP algorithms, we take respectively $M = 10000$ and $N = 10000$ and we obtain the following results.

IPS - $P(1,0)$	0.25	0.5	0.75	1
mean	1.79e-06	1.83e-06	1.92e-06	2.10e-06
std	2.29e-06	1.30e-06	1.04e-06	8.79e-07
std/mean	1.28	0.71	0.54	0.42

POP - $P(1,0)$	0.25	0.5	0.75	1
mean	2.15e-06	2.05e-06	2.06e-06	2.13e-06
std	5.76e-07	4.52e-07	3.23e-07	3.35e-07
std/mean	0.27	0.22	0.16	0.16

The performance of POP appears rather stable w.r.t. $P(1,0)$ and systematically better than the IPS method.

Second we can modify the value of c by keeping the product $Mc = Nc$ constant (the computational effort remains the same). Taking $c = 1\%$ yields less accurate results we do not report. In the opposite direction, taking $c = 100\%$ fails to work. The question of the best choice of c and $P(0,1)$ according to t, q, V is open.

4.5. Hawkes process. The Hawkes process [Haw71] is a self-exciting counting process $(N_t)_{t \geqslant 0}$ which intensity evolves as

$$\mathrm{d}\lambda_t = \theta(\mu - \lambda_t)\mathrm{d}t + \mathrm{d}N_t.$$

In the last years, it has become rather popular to model earthquakes activity, high-frequency financial data, information flow on internet (Twitter...) etc. We guess that this is challenging model for rare event simulation because of its self-exciting property. Here we set $\theta = 2$, $\mu = 1$, the terminal time $T = 24$ and $\lambda_0 = 1$. We denote all the jump instants before T by $(\tau_j)_{j \geqslant 1}$ and define $H = \max\{\tau_j - \tau_i : \tau_{k+1} - \tau_k < 0.5, i \leqslant k < j - 1\}$, which is the longest period between jump instants during which all jump inter-arrivals are less than 0.5. Our aim is estimate $\mathbb{P}(H > 11)$: using 3×10^8 crude Monte Carlo simulations gives a 99% confidence interval $[3.2469, 3.8064] \times 10^{-6}$.

According to [Oga81, Algorithm 2], Hawkes process (and thus H) can be seen as a functional of countable number of uniform variables in $[0,1]$, which fits our general setting.[3]. Thus we can use the shaking transformation for uniform variables in our algorithms. We define $n = 5$ intermediate sets as $\{H > L_k\}$ where $(L_k)_{k=1,\cdots,5} = [3.5, 5.5, 7.5, 9.5, 11]$. Results over 50 experiments for different shaking coefficients are listed in the following (with $M = N = 10^4$).

IPS	$p = 0.1$	$p = 0.3$	$p = 0.5$
mean	3.30e-06	5.19e-06	3.88e-06
std	2.84e-06	1.37e-05	1.60e-05
std/mean	0.86	2.64	4.12

POP	$p = 0.1$	$p = 0.3$	$p = 0.5$
mean	3.33e-06	3.51e-06	2.69e-06
std	1.25e-06	2.92e-06	3.71e-06
std/mean	0.37	0.83	1.38

We observe good performance of POP (about three times more accurate than IPS). Both algorithms are much more accurate than the crude Monte Carlo method, as expected.

4.6. An example of randomized shaking transformation. We conclude this presentation of numerical experiments by illustrating the benefit of randomization of shaking parameter as exposed in Proposition 2.2 of Subsection 3.4.

[3]During implementation, we only need to keep record of uniform variables that have been used.

We consider the simple problem of estimating $\mathbb{P}(G > 6 \text{ or } G < -5)$, where $X := G$ is a standard Gaussian variable. Of course, one could compute respectively $\mathbb{P}(G > 6)$ and $\mathbb{P}(G < -5)$ then add them up. But this solution requires extra knowledge about the problem that we could not afford in general; hence for the sake of exposure, we do not use this decomposition.

If we use the POP method on the initial problem with intermediate levels defined by $\{G > \sqrt{\frac{k}{5}} \times 6 \text{ or } G < -\sqrt{\frac{k}{5}} \times 5\}, k = 1, \cdots, 5$, the results are rather unstable. Over 100 experiments with the shaking $G = \rho G + \sqrt{1-\rho^2} G'$ where $\rho = 0.75$, 23 outputs are of order 10^{-9} and the others are of order 10^{-7}. This is due to the fact that in POP method we average only one path. When shaking level after level, this path tends gradually either towards $\{G > 6\}$ or towards $\{G < -5\}$ and it becomes practically impossible to realize the jump from $\{G > \sqrt{\frac{k}{5}} \times 6\}$ to $\{G < -\sqrt{\frac{k}{5}} \times 5\}$. As a consequence, only one part of the distribution is selected and estimated[4]. The IPS approach is less sensitive to this problem since it is based a large sample of paths.

To circumvent this problem for POP, we can take a random ρ such that $\rho = 0.75$ with probability 0.8 and $\rho = -0.75$ with probability 0.2: this enables the path to sometimes jump from $\{G > \sqrt{\frac{k}{5}} \times 6\}$ to $\{G < -\sqrt{\frac{k}{5}} \times 5\}$, thus to yield a better performance. Indeed over 100 experiments, with fixed ρ we get mean $2.84e-07$ and standard deviation $1.70e-07$, while with the random ρ we get mean $2.81e-07$ and standard deviation $6.73e-08$. We recall that $\mathbb{P}(G > 6 \text{ or } G < -5) = 2.8764e-07$.

In more general situations, randomization is certainly beneficial to explore disjoint configurations. The right tuning is a delicate question since too much randomization may alter the benefit of POP method. This issue is left to future investigation.

5. Conclusion. We have designed two methods to tackle the problem of rare event estimation, by building suitable Markov chains valued in the state space. This approach has the advantage to suit well to finite and infinite-dimensional situations, such as stochastic processes (possibly without Markovian assumptions). The IPS method is inspired by the well-known Interacting Particle System approach. The POP method is new and relies on ergodic properties to compute in parallel non-rare conditional probabilities. These methods make use of reversible shaking transformations and we exhibit important examples of such transformations, that are relevant for applications. Our numerical experiments show that both algorithms converge, with globally a better performance (accuracy and memory and other implementation issues) of the POP algorithm compared to the IPS one. Some theoretical estimates support the convergence rates.

For future research, it will be worth investigating the choice of optimal shaking parameters together with explicit non asymptotic error estimates. Adaptive choice of intermediate subsets $(A_k)_k$ is another important concern.

REFERENCES

[AA10] S. Asmussen and H. Albrecher. *Ruin probabilities*. Advanced Series on Statistical Science & Applied Probability, 14. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, second edition, 2010.

[BHLN13] S. Bhamidi, J. Hannig, C.Y. Lee, and J. Nolen. The importance sampling technique for understanding rare event in Erdos-Renyi random graphs. *arXiv preprint arXiv:1302.6551*, 2013.

[4]The same phenomenon occurs using importance sampling techniques and splitting methods.

[Bil99] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[Bol01] B. Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001.

[Buc04] J.A. Bucklew. *Introduction to Rare Event Simulation*. Springer, 2004.

[CDFG12] F. Cérou, P. Del Moral, T. Furon, and A. Guyader. Sequential Monte Carlo for rare event estimation. *Stat. Comput.*, 22(3):795–808, 2012.

[CDG11] F. Cérou, P. Del Moral, and A. Guyader. A nonasymptotic theorem for unnormalized Feynman-Kac particle models. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(3):629–649, 2011.

[CDLL06] F. Cérou, P. Del Moral, F. Le Gland, and P. Lezaud. Genetic genealogical models in rare event analysis. *ALEA Lat. Am. J. Probab. Math. Stat.*, 1:181–203, 2006.

[CV11] S. Chatterjee and S.R.S. Varadhan. The large deviation principle for the Erdős-Rényi random graph. *Eur. J. Comb.*, 32(7):1000–1017, 2011.

[CY12] L. Chaumont and M. Yor. *Exercises in probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, second edition, 2012. A guided tour from measure theory to random processes, via conditioning.

[DD09] T. Dean and P. Dupuis. Splitting for rare event simulation: a large deviation approach to design and analysis. *Stochastic Process. Appl.*, 119(2):562–587, 2009.

[Del04] P. Del Moral. *Feynman-Kac formulae: Genealogical and Interacting Particle Systems with applications*. Springer, New-York, 2004.

[DG05] P. Del Moral and J. Garnier. Genealogical particle analysis of rare events. *Annals of Applied Probability*, 15:2496–2534, 2005.

[DMS14] R. Douc, E. Moulines, and D. Stoffer. *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. CRC Press, 2014.

[Haw71] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58:83–95, 1971.

[HJM92] D. Heath, R. Jarrow, and A. Morton. Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation. *Econometrica*, 60, 1992.

[Kin93] J.F.C. Kingman. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1993. Oxford Science Publications.

[KLL12] J. Krystul, F. Le Gland, and P. Lezaud. Sampling per mode for rare event simulation in switching diffusions. *Stochastic Process. Appl.*, 122(7):2639–2667, 2012.

[ŁMN13] K. Łatuszyński, B. Miasojedow, and W. Niemiro. Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli*, 19(5A):2033–2066, 2013.

[MT09] S. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009.

[Oga81] Y. Ogata. On Lewis' simulation method for point processes. *Information Theory, IEEE Transactions on*, 27(1):23–31, 1981.

[PW05] M. Prandini and O.J. Watkins. Probabilistic aircraft conflict detection. *HYBRIDGE WP3: Reachability analysis for probabilistic hybrid systems*, 2005.

[PZ14] G. Da Prato and J. Zabczyk. *Stochastic equations in infinite dimensions*, volume 152. Cambridge university press, 2014.

[RG09] G. Rubino and B. Gerardo, editors. *Rare event simulation using Monte Carlo methods*. Wiley, Chichester, 2009.

[RK08] R.Y. Rubinstein and D.P. Kroese. *Simulation and the Monte Carlo method*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2008.

[Rob03] P. Robert. *Stochastic networks and queues*, volume 52 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, french edition, 2003. Stochastic Modelling and Applied Probability.

[RY99] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*. Comprehensive Studies in Mathematics. Berlin: Springer, third edition, 1999.

[VV91] M. Villén-Altamirano and J. Villén-Altamirano. RESTART: a method for accelerating rare event simulations. In *Proceedings of the 13th International Teletraffic Congress, Copenhagen*, pages 71–76, 1991.