Research project (PRe)

Speciality : Applied Mathematics
Academic year : 2021/2022

Estimation of the probability of the union of rare events

Author : Mathis Le Bail **Promotion :** 2023

Internship from 16 May to 26 August 2022

Supervisor : Joris Bierkens, Associate Professor at Delft University of Technology
TU Delft, Mekelweg 4, 2628 CD Delft, Netherlands

Referent teacher : Laure Giovangigli , Teacher-researcher
UMA ENSTA Paris, 828 boulevard de Maréchaux, 91120 Palaiseau, France

Confidentiality Notice

This present document is not confidential. It can be communicated outside in paper format or distributed in electronic format

Abstract

We consider the estimation of the probability μ of a union of J rare events H_j defined by a random variable X. To do this, we compare the ALOE estimator and the directional estimator from the papers [AC18] and [AK18] published the same year and which use two different approaches. The ALOE estimator is retained because of its better efficiency and its ability to be generalized to a large number of distributions. The latter, in the paper [AC18], is indeed only implemented for Gaussian distributions and rare events whose boundary is defined by a line. We modify this estimator using an adaptive multilevel splitting method to generalise it to a larger number of distributions and for rare events whose boundaries are defined more generally by the equation $h(x) = \tau$ where $h : \mathbb{R}^d \to \mathbb{R}$ is a general function that we assume we can evaluate at any point $x \in \mathbb{R}^d$. This new estimator is compared to direct estimation of μ using the multilevel splitting method through a theoretical study and two numerical examples.

Keywords

Rare events, Monte Carlo method, Importance sampling, Markov Chain Monte Carlo

Résumé

On veut estimer la probabilité μ de l'union de J évènements rares H_j définis par une variable aléatoire X. Pour ce faire, on compare deux estimateurs proposés par les papiers de recherche [AC18] et [AK18] publiés la même année et qui utilisent deux approches assez différentes. Au vu des résulats obtenus, on conserve seulement l'un des deux, celui qu'on appelle estimateur ALOE. Ce dernier dans le papier [AC18] est seulement implémenté pour des distributions gaussiennes et des évènements rares dont les frontières sont définies par une droite. On modifie cet estimateur en utilisant une méthode multi-niveaux adaptive afin de généraliser son fonctionnement à un plus grand nombre de distributions et pour des évènements rares dont les frontières sont définies plus généralement par l'équation $h(x) = \tau$, où $h : \mathbb{R}^d \to \mathbb{R}$ est une fonction quelconque que l'on suppose pouvoir évaluer en tout point $x \in \mathbb{R}^d$. Ce nouvel estimateur est comparé à une estimation directe de μ utilsiant seulement la méthode multi-niveaux à travers une étude théorique et deux exemples numériques.

Mots clés

Evènements rares, méthode de Monte Carlo, Échantillonnage préférentiel, Chaîne de Markov Monte Carlo

Contents

List of Figures

Chapter 1

Introduction

Let $X \in \mathbb{R}^d$ a random variable representing the state of a system, we note F_X its distribution. We consider the problem of estimating $\pi = \mathbb{P}(X \in A)$ where A is the event of interest. To deal with this kind of calculation we often use the crude Monte Carlo method (CMC) which consists of choosing the estimator $\Pi = \mathbb{1}_{\{X \in A\}}$ that follows a Bernoulli distribution of parameter π and estimate $\pi = \mathbb{P}(X \in A) = \mathbb{E}[\Pi]$ by

$$\pi_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(x_i) \tag{1.1}$$

where the (x_i) are i.i.d realisations of $X \sim F_X$.

The crude Monte Carlo estimator $\hat{\pi_n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(X_i)$ with $X_1, ..., X_n$ i.i.d random variables following the distribution F_X is unbiased and has a variance of $\sigma^2_{\pi_n} = \frac{\pi(1-\pi)}{n}$. This quantity goes to 0 as π tends to 0 so the absolute error $\sigma^2_{\pi_n}$ is small for low probabilities π. But a more relevant quantity to consider is the relative error or coefficient of variation (CV) $\frac{\sigma_{\pi_n}}{\pi}$ which is high for low probability π :

$$\frac{\sigma_{\pi_n}}{\pi} = \frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}\pi} \underset{\pi \to 0}{\sim} \frac{1}{\sqrt{n\pi}} \underset{\pi \to 0}{\to} \infty \tag{1.2}$$

The problem studied in this report is to estimate π when this quantity is very small (order of magnitude 10^{-2} or less), the event is then called a rare event. According to what is written above, we are therefore in the case of a high coefficient of variation. A large relative error means that the width of the confidence interval obtained for π will be in terms of order of magnitude much larger than the estimate $\hat{\pi}$. We can take the example given in [Sør07], if we consider $\hat{\pi_n}$ for a large number N of independent draws, we can approximate its distribution by a Gaussian distribution. We want to determine the sample size needed to acquire a 10% relative precision for the half-width of the 95% confidence interval. We have to resolve the equation

$$\frac{1.96\sqrt{\pi(1-\pi)}}{\pi\sqrt{N}} \approx 0.1 \tag{1.3}$$

which leads to

$$N \approx \frac{100 \times 1.96^2 \pi(1-\pi)}{\pi^2} \sim \frac{100 \times 1.96^2}{\pi} \tag{1.4}$$

The sample size increases as $\frac{1}{\pi}$. For a π of the order $\pi \approx 10^{-8}$, we would have $N \approx 10^{10}$. Such a large number of samples makes it impossible to estimate the probability correctly by this method.

Thus in the case of rare event estimation, estimators have to be designed more thoughtfully. To be effective they often have to be considered specifically for the problem under study. However, in general, there are two types of efficiency criteria for estimators of tail probabilities in the rare-event simulation literature. We repeat here the presentation of these two criteria proposed in the paper [ABL08].

1.1 Efficiency properties in rare-event simulation

Let be a family of rare events $\{R(\tau)\}$, which means that $\pi(\tau) = \mathbb{P}(R(\tau)) \to 0$ when $\tau \to \infty$. For instance, $R(\tau) = (X > \tau)$ where X is a random variable. Suppose that we have an unbiased estimator Π_τ of $\pi(\tau)$ for each τ.

Definition 1. Π_τ *is strongly efficient if*

$$\sup_\tau \frac{Var(\Pi_\tau)}{\pi(\tau)^2} < \infty \tag{1.5}$$

This property would mean that the relative error is bounded as $\tau \to \infty$. The second property of interest for the estimator, although weaker than the first one, is :

Definition 2. Π_τ *is asymptotically optimal or weakly efficient if*

$$\frac{\log(\mathbb{E}(\Pi_\tau^2))}{2\log(\pi(\tau))} \underset{\tau\to\infty}{\to} 1 \tag{1.6}$$

Such a property is also called logarithmic efficiency. The strong efficiency implies the weak efficiency. Indeed

$$\begin{aligned}\frac{\log(\mathbb{E}(\Pi_\tau^2))}{2\log(\pi(\tau))} &= \frac{\log(Var(\Pi_\tau) + \mathbb{E}(\Pi_\tau)^2)}{2\log(\pi(\tau))} \\ &= \frac{2\log(\pi(\tau))}{2\log(\pi(\tau))} + \frac{\log(\frac{Var(\Pi_\tau)}{\pi(\tau)^2} + 1)}{\log(\pi(\tau)^2)} \underset{\tau\to\infty}{\to} 1\end{aligned}$$

because if Π_τ is strongly efficient, $\log(\frac{Var(\Pi_\tau)}{\pi(\tau)^2} + 1) < \infty$ for all τ and $\log(\pi(\tau)^2) \underset{\tau\to\infty}{\to} -\infty$.

These properties are interesting because they allow us to obtain the same accuracy as a CMC method by using less averaging on the estimator. Indeed, we take $\widehat{\Pi_n(\tau)} = \frac{1}{n}\sum_{i=1}^{n}\Pi_\tau^i$ the average of the i.i.d n random variables $\Pi_\tau^1, ..., \Pi_\tau^n$ to estimate $\pi(\tau)$. We want a relative accuracy which verifies :

$$\mathbb{P}\left(\left|\frac{\widehat{\Pi_n(\tau)} - \pi(\tau)}{\pi(\tau)}\right| > \epsilon\right) < \delta \tag{1.7}$$

Chebyshev's inequality gives us

$$\mathbb{P}\left(\left|\frac{\widehat{\Pi_n(\tau)} - \pi(\tau)}{\pi(\tau)}\right| > \epsilon\right) \leq \frac{Var(\frac{\widehat{\Pi_n(\tau)}}{\pi(\tau)})}{\epsilon^2} = \frac{Var(\Pi_\tau)}{n\pi(\tau)^2\epsilon^2} \tag{1.8}$$

Then if Π_τ is strongly efficient, there is a $K \in \mathbb{N}$ such that $\frac{Var(\Pi_\tau)}{\pi(\tau)^2} \leq K$ for all τ. The upper bound on the probability is no longer dependent on τ :

$$\mathbb{P}\left(\left|\frac{\widehat{\Pi_n(\tau)} - \pi(\tau)}{\pi(\tau)}\right| > \epsilon\right) \leq \frac{K}{n\epsilon^2} \tag{1.9}$$

Thus, with ϵ and δ fixed, the required number of replication n is bounded for all τ. The accuracy is guaranteed with $n \geq \frac{K}{\delta\epsilon^2}$. However, if we look at the crude Monte-Carlo estimator Π_τ^{CMC} of variance $Var(\Pi_\tau^{CMC}) = \pi(\tau)(1-\pi(\tau))$ the right-hand side of (1.8) is equal to

$$\frac{Var(\Pi_\tau)}{n\pi(\tau)^2\epsilon^2} = \frac{1-\pi(\tau)}{n\pi(\tau)\epsilon^2} \approx \frac{1}{n\pi(\tau)\epsilon^2} \tag{1.10}$$

To reach the same accuracy, the number of replications needed is $n = O(\frac{1}{\pi(\tau)})$ which soars when $\tau \to \infty$. To be comprehensive, [ABL08] states that if Π_τ is weakly efficient, one needs $n = exp(-o(\log(\pi(\tau))))$, an intermediate number between the two previous results. The paper also points out that this quick calculation does not take into account the overall calculation complexity required to obtain one replication Π_τ which could be longer to compute if this one is strongly efficient compare to a simple Π_τ^{CMC}. However, if the computation times are not disproportionate, one can see the interest of these two properties when looking for estimators of extreme probabilities.

Much work to obtain estimators with these properties focuses on importance sampling as a way to efficiently construct estimators. The principle and reasons for applying this method in the case of estimating probability of a rare event are briefly outlined below. For more details on the different Importance Sampling techniques see [EM21], the explanations given below are inspired by it.

1.2 Importance sampling

As previously written, we can estimate $\pi = \mathbb{P}(X \in A) = \mathbb{E}[\mathbb{1}_{\{X\in A\}}] = \int \mathbb{1}_A(x)\,\mathrm{d}F(x)$ by the estimator $\hat{\Pi_n}^{CMC} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_A(X_i)$ where $X_1, ..., X_n$ are i.i.d random variables

of distribution function F(.). As the event A is rare for the distribution F(.), few X_i will end up in the area of interest A and contribute to the estimator $\hat{\Pi}_n$ which leads to a severe under-estimation of the rare event probability. The basic idea of importance sampling is to change the distribution used to simulate the samples such that under the new proposal probability density function (pdf) the event of interest is less rare. Let $q(\cdot)$ denote the new proposal pdf and $f(\cdot)$ the original one. q must be such that $q(x) > 0$ for all x where $\mathbb{1}_A(x)f(x) \neq 0$. We then have

$$\pi = \int \mathbb{1}_A(x)\, f(x)\mathrm{d}x = \int \mathbb{1}_A(x)\frac{f(x)}{q(x)}q(x)\,\mathrm{d}x = \mathbb{E}_q\left[\mathbb{1}_A(X)\frac{f(X)}{q(X)}\right] \tag{1.11}$$

We introduce the notation $\mathbb{E}_q$ to specify that the random variable X is distributed with the new pdf q. It then follows naturally that a new estimator of π is $\Pi_n^{IS} = \frac{1}{n}\sum_{i=1}^n \frac{f(X_i)}{q(X_i)}\mathbb{1}_A(X_i)$ where this time $X_i \overset{i.i.d}{\sim} q$. The quantities $w_i = \frac{f(x_i)}{q(x_i)}$ where the (x_i) are the realisations of the r.v's (X_i) are called the importance weights. Under q more x_i are in A thus there is a greater contribution to the probability estimator and the importance weights describe how representative the samples simulated from q are when one is interested in computing integrals with respect to f.

$\hat{\Pi_n^{IS}}$ is always unbiased but its variance depends heavily on the proposal pdf q chosen. Its variance when we choose the proposal pdf q is

$$Var(\hat{\Pi_n^{IS}}) = \frac{1}{n}\int \frac{(\mathbb{1}_A(x)\, f(x) - \pi q(x))^2}{q(x)}\,\mathrm{d}x \tag{1.12}$$

The best possible choice of q is then equal to $q^*(x) = \frac{\mathbb{1}_A(x)f(x)}{\pi}$ since it allows us to obtain a zero variance. However, this distribution can never be implemented in practice since it involves the knowledge of the unknown π. Nevertheless, q^* gives a useful intuition to choose the proposal, the efficiency of the estimator is dependent on the mismatch of $\mathbb{1}_A(x)\, f(x)$ and $q(x)$ with this penalization amplified by $q(x)$. We note $\mathbb{P}$ the probability associated to the pdf f and $\mathbb{P}^*$ the probability associated to the pdf q^*, we have $\mathbb{P}^*(.) = \mathbb{P}(.|A)$ and $\mathbb{P}^*(dx) = \frac{\mathbb{1}_A(x)}{\mathbb{P}(A)}\mathbb{P}(dx)$. Thus, we wish to use a sampling distribution close to the conditional distribution of $\mathbb{P}$ given A. This method is often used to obtain very good "rare event" importance sampling algorithms.

Chapter 2

Estimating the probability of the union of rare events

Our objective is a little different from just computing the probability of occurrence of a rare event but to characterise an event that could represent a union of rare events and estimate its probability. The notations for the problem are defined below. We then present two research papers that address this problem, each using a rather different approach. These papers make restrictive assumptions on the boundary characterizing the rare event which is always considered as a line. We will present and compare the theoretical and numerical results of the two estimators proposed by the research papers. In view of the results, we will keep only one of them and from this one we will look for an estimator which still holds for a more general problem.

2.1 Formulation of the problem

Let X a d-dimensional random vector $X = (X_1, ..., X_d)$ and a fixed integer J. For each $j \in \{1, ..., J\}$, we have a fixed unit vector $\omega_j \in \mathbb{R}^d$ ($\omega_j^\top \omega_j = 1$) and $\tau_j \in \mathbb{R}$. Rare events are defined as follows :

$$H_j = \left\{x : x^\top \omega_j \geqslant \tau_j\right\} \tag{2.1}$$

We note the set $H = \cup_{j=1}^J \{x \in \mathbb{R}^d : \omega_j^\top x \geq \tau_j\} = \cup_{j=1}^J H_j$. The probability we want to compute is then $\mu = \mathbb{P}(H)$. We also notice that the $(H_j)_j$ defined are half-spaces because of the linear constraints.

If we want to be more general and compute the expectation of $h(.)$ a real-valued function on H the union of the half-spaces, the quantity of interest would be $\mathbb{E}[h(X)\mathbf{1}_{\{X \in H\}}]$.

The two articles [AC18] and [AK18] are presented in detail below, each proposing a different method for estimating the quantities of interest just presented. Both papers make restrictive assumptions in order to be able to estimate them in practice, in particular on the distribution followed by the vector X. These assumptions will be

discussed and the results of these two estimators will be compared numerically and theoretically.

2.2 Importance sampling the union of rare events using the ALOE algorithm

The paper [AC18] focuses on a mixture importance sampling strategy to estimate the probability that one or more rare events takes place. To do this, globally, the sampler will repeatedly choose a rare event at random, and then samples the system conditionally on that one event taking place. For that reason the algorithm is called in the paper ALOE for "At least one event". To estimate $\mu = \mathbb{P}(H)$, the paper assumes that the random variable X follows a Gaussian distribution : $X \sim \mathcal{N}(\eta, \Sigma)$. To better visualize the $(H_j)_j$ and the set H, the paper provides a graphical example in the case of a standard bivariate Gaussian distribution with $J = 6$.

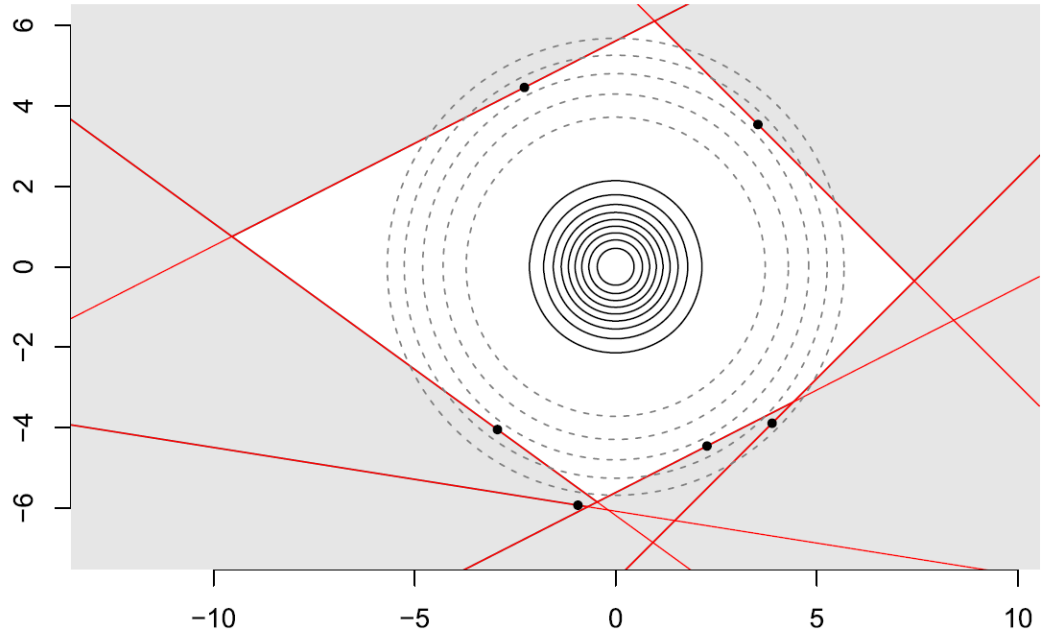

Figure 2.1: Graphical representation of a standard bivariate Gaussian distribution with J = 6 half-spaces which are the rare events

The red lines denote the half-spaces H_j, they represent the boundaries of equations $\omega_j^\top x = \tau_j$. The ωj's define the coefficients of the boundary lines. Hence the unit vectors ωj's are the unit vectors orthogonal to these boundary lines. And τ_j is the distance from the centre to the boundary of H_j in the ω_j direction. The solid circles contain $10\%, 20\%$ up to 90% of the $\mathcal{N}(0, I)$ distribution. The dashed circles contain all but 10^{-k} of the probability for $3 \leq k \leq 7$. The solid points are the corresponding conditional modes i.e. it marks the point of highest probability density knowing that the event H_j occurs.

We can see on this example, and it is always the case, that the boundary lines define a polytope $\mathcal{P}$. It corresponds to the white area in the example.

$$\mathcal{P} = \cap_{j=1}^{J} H_j^c = H^c$$

It is also important to note that the set $\mathcal{P}$ is convex but not always necessarily bounded in every direction.

2.2.1 Standard Multivariate Gaussian Distribution

The first part of the paper focuses on a X which follows a standard Gaussian distribution. For all $j \in \{1, ..., J\}$, $w_j^\top X$ follows a normal distribution since it is a linear combination of components of a gaussian vector.

$$\omega_j^\top X \sim \mathcal{N}(0, \omega_j^\top \mathbf{I} \omega_j) = \mathcal{N}(0, \omega_j^\top \omega_j) = \mathcal{N}(0, 1) \tag{2.2}$$

We can then compute easily the probability that $X \in H_j$:

$$\begin{aligned} P_j &= \mathrm{P}\left(X \in H_j\right) \\ &= P(\omega_j^\top X \geq \tau_j) \\ &= P(\omega_j^\top X \leq -\tau_j) \\ &= \Phi(-\tau_j) \end{aligned} \tag{2.3}$$

where Φ is the cumulative distribution function of the standard gaussian distribution.

One can then define known bounds on μ. It always holds :

$$\max_{1 \leqslant j \leqslant J} P_j =: \underline{\mu} \leqslant \mu \leqslant \bar{\mu} := \sum_{j=1}^{J} P_j \tag{2.4}$$

The right hand side is the union bound which is sometimes very conservative but sometimes quite accurate. We will need these bounds to define new ones on the estimator introduced below.

The paper establishes the following notations. For any $u \subseteq 1 : J \equiv \{1, 2, \ldots, J\}$, let $H_u = \cup_{j \in u} H_j$, so $H_j = H_{\{j\}}$ and by convention $H_\varnothing = \varnothing$. We identify the set H_u with the function $H_u(x) = \mathbb{1}_{H_u}(x)$. Next define $P_u = \mathbb{E}\left(H_u(X)\right)$ for $X \sim \mathcal{N}(0, I)$. We use $-u$ for complementary sets in $1 : J$, and $H_u^c(x)$ for the complementary outcome $1 - H_u(x)$.
Let $S(x) = \sum_{j=1}^{J} H_j(x)$ count the number of rare events that happen. For $s = 0, 1, \ldots, J$, let $T_s = \Pr(S = s)$ give the distribution of S. We use $|u|$ for the cardinality of u.

For the creation of the new estimator of μ, we can see the ALOE as an especially simple mixture sampler where mixture components are conditional distributions. We choose one rare event H_j at random proportionally to P_j, the probability that event j happens and then we sample conditionally on that event taking place. Concretely, the mixture components are the conditional distribution $q_j = \mathcal{L}\left(X \mid \omega_j^\top X \geqslant \tau_j\right)$. And the mixture distribution resulting is :

$$q_\alpha = \sum_{j=1}^{J} \alpha_j q_j \tag{2.5}$$

where α_j is the probability to choose the event H_j. To choose α_j, it seems normal to think that we will sample more conditionally on an event if it is more likely to take place than the others, so we take $\alpha_j = \alpha_j^* \equiv \frac{P_j}{\sum P_j}$. We note

$$q_\alpha^* = \sum_{j=1}^{J} \alpha_j^* q_j \tag{2.6}$$

If we develop the density q_j using the Bayes formula, we obtain :

$$q_j(x) = p(x)H_j(x)/P_j \tag{2.7}$$

where $p(\cdot)$ is the pdf of the initial distribution of X.

Once the distribution q_α^* is defined, we have the following theorem :

Theorem 1. *If* $1 \leq J < \infty$ *and* $\min_j P_j > 0$ *and* $n \geq 1$, *then*

$$\hat{\mu}_{\alpha^*} = \frac{\bar{\mu}}{n} \sum_{i=1}^{n} \frac{1}{S(X_i)} \tag{2.8}$$

where $X_i \sim q_\alpha^*$, *satisfies* $\mathbb{P}(\frac{\bar{\mu}}{J} \leq \hat{\mu}_{\alpha^*} \leq \bar{\mu}) = 1$,

$$\mathbb{E}\left(\hat{\mu}_{\alpha^*}\right) = \mu \tag{2.9}$$

and

$$\mathrm{Var}\left(\hat{\mu}_{\alpha^*}\right) \leq \frac{1}{n}\left(\bar{\mu}\sum_{s=1}^{J}\frac{T_s}{s} - \mu^2\right) \leqslant \frac{\mu(\bar{\mu}-\mu)}{n} \tag{2.10}$$

It is interesting to note that the theorem applies so long as we can sample conditionally on any one event H_j and then determine which other rare events also occur which means it still holds for non-Gaussian distributions.

This estimate $\hat{\mu}_{\alpha^*}$ is a multiplicative adjustment to the union bound. The terms $\frac{1}{S(x_i)}$ range from 1 to $\frac{1}{J}$, it means the estimate will never get larger than the union bound or smaller than the union bound divided by J : $\frac{\bar{\mu}}{J} \leq \hat{\mu_\alpha} \leq \bar{\mu}$. It is convenient since it is also true for the real value μ :

$$\frac{\bar{\mu}}{J} \leq \underline{\mu} \leq \mu \leq \bar{\mu}$$

The left-hand side is obtained as μ is greater than $\underline{\mu} = \max_j P_j$ which is greater than the mean of the P_j's represented by the union bound $\bar{\mu}$ divided by J.

As an another remark, one common way for rare event sampling to be inaccurate is that we might fail to obtain any points where the rare event happens. That leads to a severe under-estimation of the rare event probability. In ALOE, the corresponding problem is the failure to sample any points where two or more of the rare constituent events occur. In that case ALOE will return the union bound as the estimated rare event probability instead of zero. In that case, it is likely that the union bound be a good approximation because the rare events very little overlap and are close to being disjointed.

2.2.2 More generality

Once we have dealt with the case of standard multivariate gaussian distributions, one can always come back to it. Let $Y \sim \mathcal{N}(\eta, \Sigma)$ and the half-spaces are defined by $\gamma_j^\top Y \geq \kappa_j$. We assume that Σ is non-singular and there is a square-root matrix $\Sigma^{\frac{1}{2}}$. We have $X = \Sigma^{-\frac{1}{2}}(Y - \eta) \sim \mathcal{N}(0, \Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}}) = \mathcal{N}(0, I)$ and $Y = \eta + \Sigma^{\frac{1}{2}}X$.

$$\begin{aligned} & \gamma_j^\top y \geq \kappa_j \\ \Leftrightarrow \quad & \gamma_j^\top(\eta + \Sigma^{\frac{1}{2}}x) \geq \kappa_j \\ \Leftrightarrow \quad & \gamma_j^\top \Sigma^{\frac{1}{2}} x \geq \kappa_j - \gamma_j^\top \eta \end{aligned} \tag{2.11}$$

In order to have $\omega_j^\top \omega_j = 1$ and as $(\gamma_j^\top \Sigma^{\frac{1}{2}})(\gamma_j^\top \Sigma^{\frac{1}{2}})^\top = \gamma_j^\top \Sigma \gamma_j$. We take $\omega_j^\top = \frac{\gamma_j^\top \Sigma^{\frac{1}{2}}}{\sqrt{\gamma_j^\top \Sigma \gamma_j}}$ and $\tau_j = \frac{\kappa_j - \gamma_j^\top \eta}{\sqrt{\gamma_j^\top \Sigma \gamma_j}}$. The half-spaces are then given by

$$\omega_j^\top x \geq \tau_j \tag{2.12}$$

For rare events, we have $\gamma_j^\top \mu < \kappa_j$ so τ_j is still strictly positive.

If we are interested in computing $\mathbb{E}[h(X)\mathbf{1}_{\{X \in H\}}] = \int_H h(x)p(x)\mathrm{d}x := \nu(h)$. The ALOE estimator can be re-written as

$$\hat{\nu}(h) = \frac{\bar{\mu}}{n} \sum_{i=1}^{n} \frac{h(X_i)}{S(X_i)} \tag{2.13}$$

where $X_i \sim q_\alpha^*$. Using the same steps as for the proof of Theorem 1, one can show :

$$\mathbb{E}[\hat{\nu}(h)] = \nu(h) \quad \text{and} \quad Var(\hat{\nu}) = \frac{1}{n}(\bar{\mu} \int_H \frac{h(x)^2 p(x)}{S(x)} \mathrm{d}x - \nu^2) \tag{2.14}$$

2.2.3 Sampling method

To sample $X_i = x_i$ from the q_α^* distribution, we have to be able to sample $X \sim \mathcal{N}(0, I)$ conditionally on $X^\top \omega \geqslant \tau$ for a unit vector ω and scalar τ. For that, we use the following steps :
1) Sample $Z \sim \mathcal{N}(0, I)$.
2) Sample $U \sim \mathbb{U}(0, 1)$.
3) Let $Y = \Phi^{-1}(\Phi(\tau) + U(1 - \Phi(\tau)))$.
4) Deliver $X = \omega Y + \left(I - \omega\omega^\top\right) Z$.

This method is based on the research paper [Dou10]. This paper only focus on the case of a conditional equality gaussian distribution.

For example, if we have $V = \begin{pmatrix} X \\ Y \end{pmatrix}$ with $X \in \mathbb{R}^{n_x}$ and $Y \in \mathbb{R}^{n_y}$ and $V \sim \mathcal{N}(m, \Sigma)$.

$$m = \begin{pmatrix} m_x \\ m_y \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$$

The distribution $X|(Y = y)$ still follows a gaussian distribution which can be described analytically.

$$X|(Y = y) \sim \mathcal{N}(m_{x|y}, \Sigma_{x|y})$$

where $m_{x|y} = m_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - m_y)$ and $\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^\top$

To sample according to the distribution $\mathcal{N}(m_{x|y}, \Sigma_{x|y})$, the paper [Dou10] shows that we can use the estimator

$$\bar{X} = X + \Sigma_{xy}\Sigma_{yy}^{-1}(y - Y)$$

In our case, $X = Z \sim \mathcal{N}(0, I)$, $Y = \omega^\top Z \sim \mathcal{N}(0, 1)$, $\Sigma_{xy} = w$ and $\Sigma_{yy} = 1$. So if we wanted to sample $(Z|\omega^\top Z = \tau)$, we would choose $X = Z + \omega(\tau - \omega^\top Z) = \omega\tau + (I - \omega\omega^\tau)Z$ where $Z \sim \mathcal{N}(0, I)$.

For inequality condition, we take $X = \omega Y + (I - \omega\omega^\tau)Z$ where Y is a truncated standard normal distribution above τ.

Indeed, we can also see the following reasoning for writing X, we consider the vector $z = z_\omega + z_{\omega^\perp} \in \mathbb{R}^n$ where z_ω is the orthogonal projection of z onto the vector space generated by the vector ω. Here,

$$z_\omega = (\omega^\top z)\omega = \omega\omega^\top z \quad \text{and} \quad z_{\omega^\perp} = z - z_\omega = z - \omega\omega^\top z = (I - \omega\omega^\top)z$$

Thus, the random vector $Z \sim \mathcal{N}(0, I)$ can be written as follows :

$$Z = \omega\omega^\top Z + (I - \omega\omega^\top)Z \tag{2.15}$$

$(I - \omega\omega^\top)Z$ is not involved in the scalar product between ω and Z. To meet the inequality condition, we replace $\omega^\top Z$ by Y, the truncated standard normal gaussian above τ.

For a more formal proof, we show that $X = \omega Y + \left(I - \omega\omega^\top\right) Z$ and $Z|\omega^\top Z$ have the same distribution. We can take $\omega = e_1$ without loss of generalities (we can rebuild the space in taking this vector equals to e_1).

$$\begin{aligned} &\mathbb{P}((Z_1, ..., Z_d) \leq (\xi_1, ..., \xi_d) \,|\, \omega^\top Z \geq \tau) \\ =&\mathbb{P}(Z_1 \leq \xi_1, ..., Z_d \leq \xi_d \,|\, {e_1}^\top Z \geq \tau) \\ =&\mathbb{P}(Z_1 \leq \xi_1, ..., Z_d \leq \xi_d \,|\, Z_1 \geq \tau) \end{aligned} \tag{2.16}$$

And $X = e_1 Y + \left(I - e_1 e_1^\top\right) Z = \begin{pmatrix} Y \\ Z_2 \\ \dots \\ Z_d \end{pmatrix}$

Then

$$\begin{aligned} &\mathbb{P}(X \leq (\xi_1, ..., \xi_d)) \\ =&\mathbb{P}(Y \leq \xi_1, ..., Z_d \leq \xi_d) \\ =&\mathbb{P}(Z_1 \leq \xi_1, ..., Z_d \leq \xi_d \,|\, Z_1 \geq \tau) \end{aligned} \tag{2.17}$$

as Y follows a truncated standard gaussian distribution above τ. We have demonstrated what we wanted.

Returning to the practical implementation of this algorithm, step 3 is replaced by $Y = \Phi^{-1}(U\Phi(-\tau))$ and we deliver $X = -X$. This trick is used because we get better numerical stability in sampling $Z \sim \mathcal{N}(0, I)$ conditionally on $X^\top\omega \leq -\tau$. Also X is rewritten as $X = \omega Y + Z - \omega(\omega^\top Z)$ to avoid an expensive multiplication $\left(I - \omega\omega^\top\right) Z$.

2.2.4 Algorithm

To summarize, to approach the expectation of the estimator (2.8), it is done as follows :

For each realisation x_i of $X_i \sim q_\alpha^*$ ($i \in \{1, ..., n\}$):
1. Choose at random one rare event H_j. The probability to choose H_j is equal to $\alpha_j^* = \frac{P_j}{\bar{\mu}}$
2. To sample $X \sim \mathcal{N}(0, 1)$ given that $X^\top\omega_j \geq \tau_j$:
   2.1 Sample $Z \sim \mathcal{N}(0, I)$
   2.2 Sample $U \sim \mathbb{U}(0, 1)$
   2.3 Let $Y = \Phi^{-1}(U\Phi(-\tau_j))$
   2.4 Let $X = \omega_j Y + Z - \omega_j(\omega_j^\top Z)$
   2.5 Deliver $X = -X$

Compute $\frac{\bar{\mu}}{n} \sum_{i=1}^n \frac{1}{S(x_i)}$

For a more general Gaussian distribution than the standard one, it is sufficient to apply the method seen in (2.2.2) before applying the above algorithm.

2.3 Efficient Simulation for Expectations over the Union of Half-Spaces using directional simulation

We present in a similar way the functioning of the algorithm of the paper [AK18], which has the same objective as the previous one, but which uses a method of directional simulation and sampling.

The new approach allows to consider more general distributions than just the Gaussian family, it assumes that the random vector X follows an elliptical distribution. Below we briefly write some results concerning elliptical distributions, they are presented in more detail in [Fra04].

2.3.1 Elliptical distributions

Let X be a d-dimension random vector, X follows a spherical distribution if and only if :

$$X = RS$$

where S a random vector which is uniformly distributed on the unit sphere $S^{d-1} = \{s \in \mathbb{R}^d : s^\top s = 1\}$ and R a nonnegative random variable being independent of S. R is called the generating random variable or generating variate of X.

R has the same distribution as $\|X\|_2$. Indeed, $\|X\|_2^2 = X^\top X = R^2 S^\top S = R^2$. For example, the standard gaussian distribution is a spherical distribution, so if $X \sim \mathcal{N}_d(0, I_d)$ we have $X = RS$ and $R^2 = \|X\|_2^2 = \chi_d^2$. So the generating variate of X follows a $\sqrt{\chi_d^2}$ distribution.

We can transform the spherical distribution to say that X has an elliptical distribution given by:

$$X = \eta + R\Lambda S$$

where S a d-dimensional random vector which is uniformly distributed on the unit sphere S^{d-1}, R a nonnegative random variable being independent of S, $\eta \in \mathbb{R}^d$, $\Lambda \in \mathbb{R}^{d \times d}$ and $\text{rank}(\Lambda) = d$.

The matrix Λ transforms the spherical vector S along an elliptical density surface, as R does not take into account the direction because it is independent of S, the elliptical surface obtained will have the same density at all points. The generating random variable R determines the distribution's shape, in particular the tailedness of the distribution.

Example 1. Let $\eta \in \mathbb{R}^d$ and $\Lambda \in \mathbb{R}^{d\times d}$ such that $\Sigma = \Lambda\Lambda^\top$. $X \sim \mathcal{N}(\eta, \Sigma)$ is an elliptical distribution. X can be written like this :

$$X = \eta + \sqrt{\chi_k^2}\Lambda S$$

We have still $R^2 \sim \chi_d^2$. Indeed if $X \sim \mathcal{N}(\eta, \Sigma)$ then $\Sigma^{-\frac{1}{2}}(X-\eta) \sim \mathcal{N}(0, I_d)$. So $\|\Sigma^{-\frac{1}{2}}(X-\eta)\|^2 \sim \chi_d^2$ and finally

$$\begin{aligned} \|\Sigma^{-\frac{1}{2}}(X-\eta)\|^2 &= (\Sigma^{-\frac{1}{2}}R\Lambda S)(\Sigma^{-\frac{1}{2}}R\Lambda S)^\top \\ &= \Sigma^{-\frac{1}{2}}R\Lambda S S^\top \Lambda^\top R(\Sigma^{-\frac{1}{2}})^\top \\ &= R^2\Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}} \\ &= R^2 \end{aligned} \tag{2.18}$$

△

If X follows an elliptical distribution with the same assumptions as above and that the c.d.f of R is absolutely continuous then it admits a density function f_x :

$$f_X(x) = |\Sigma|^{-\frac{1}{2}} g_R((x-\eta)^\top\Sigma^{-1}(x-\eta)) \tag{2.19}$$

where

$$t \longmapsto g_R(t) = \frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}}\sqrt{t}^{-(d-1)} f_R(\sqrt{t}) \tag{2.20}$$

t > 0, d the dimension of X and f_R is the p.d.f of R.

Example 2. Let X follow a multivariate t-distribution of dimension d with $\nu > 0$ the degrees of freedom and of mean η. Its density function is written as follows :

$$f_X(x) = \frac{\Gamma(\nu+d)}{\Gamma(\frac{\nu}{2})\nu^{\frac{d}{2}}\pi^{\frac{d}{2}}}|\Sigma|^{-\frac{1}{2}}(1+\frac{1}{\nu}(x-\eta)^\top\Sigma^{-1}(x-\eta))^{-\frac{\nu+d}{2}}$$

Σ is assumed to be positive definite

Thus, the density generator R of X is

$$t \longmapsto g_R(t) = \frac{\Gamma(\nu+d)}{\Gamma(\frac{\nu}{2})\nu^{\frac{d}{2}}\pi^{\frac{d}{2}}}(1+\frac{t}{\nu})^{-\frac{\nu+d}{2}}$$

f_R is then defined as follows :

$$
\begin{aligned}
f_R(t) &= \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} t^{d-1} g_R(t^2) \\
&= \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} t^{d-1} \frac{\Gamma(\nu+d)}{\Gamma(\frac{\nu}{2})\nu^{\frac{d}{2}}\pi^{\frac{d}{2}}} (1+\frac{t^2}{\nu})^{-\frac{\nu+d}{2}} \\
&= \frac{\Gamma(\nu+d)}{\Gamma(\frac{\nu}{2})\Gamma(\frac{d}{2})} \frac{2}{\nu^{\frac{d}{2}}} \frac{1}{t} (t^2)^{\frac{d}{2}} d^{\frac{d}{2}} d^{-\frac{d}{2}+1} \frac{1}{d} (1+\frac{t^2}{\nu})^{-\frac{\nu+d}{2}} \\
&= \frac{2t}{d} \frac{\Gamma(\nu+d)}{\Gamma(\frac{\nu}{2})\Gamma(\frac{d}{2})} (\frac{d}{\nu})^{\frac{d}{2}} (\frac{t^2}{d})^{\frac{d}{2}-1} (1+\frac{d}{\nu}\frac{t^2}{d})^{-\frac{\nu+d}{2}} \\
&= \frac{2t}{d} f_F(\frac{t^2}{d})
\end{aligned}
$$

where f_F is the p.d.f function of a F-distribution with degrees of freedom d and ν.

△

These results will be useful later on for the understanding of the paper. We return below to the explanation of the directional sampling method proposed in the paper [AK18].

2.3.2 Method and distribution assumption

[AK18] develops a conditional Monte Carlo method for the random vector X to be sampled from the target region H directly. The aim is again to improve the efficiency of the estimator compared to a crude Monte Carlo method which samples fewer and fewer values in the area of interest as the event H becomes rare.

The method to sample X is based on these assumptions :

The random vector X follows an elliptical distribution, it can be written $X = \eta + R\Lambda\Theta$. As defined previously, $\eta \notin H$ is the fixed mean vector, R is a nonnegative radial random variable, $\Lambda \in \mathbb{R}^{d\times d}$ such that $\Sigma = \Lambda\Lambda^\top$ is positive definite and Θ is uniformly distributed on the unit sphere S^{d-1} and independent of R. We also suppose known the continuous density function f_R, the cumulative distribution function F_R and its inverse for the variable R. The elliptical function can cover several known distributions such as the normal distribution, the t-distribution or the symmetric generalized hyperbolic distribution.

The first approach is to evaluate the expectation by conditioning on Θ. The global idea is that a direction is chosen in choosing at random a unit vector $\theta \in S^{d-1}$ and then average $\mathbb{P}(X \in H)$ along this line. If H^c is bounded in this direction, then the line reaches the set H at distance $\min_j \frac{\tau_j - \omega_j^\top \eta}{\omega_j^\top \Lambda\theta}$ with $\omega_j^\top \Lambda\theta > 0$ in crossing the boundary line of the rare event H_j corresponding. If it is not bounded in this direction, the line never reaches H. In the case where H^c is bounded in the direction

chosen, the average $\mathbb{P}(X \in H)$ along this line depends of the distribution of R which we remind is the same independently of the direction.

If we can easily compute $\mathbb{P}(\eta + R\Lambda\theta \in H)$ when $\Theta = \theta$, then the new estimator does have a reduced variance with respect to the crude Monte-Carlo one. Indeed, the estimator is then $\mathbb{P}(\eta + R\Lambda\Theta \in H|\Theta)$ and we have :

$$\mathbb{P}(X \in H) = \mathbb{E}[\mathbb{1}_{\eta+R\Lambda\Theta\in H}] = \mathbb{E}[\mathbb{E}[\mathbb{1}_{\eta+R\Lambda\Theta\in H}|\Theta]] = \mathbb{E}[P(\eta + R\Lambda\theta \in H|\Theta)] \quad (2.21)$$

$$\begin{aligned} Var(\mathbb{1}_{(X\in H)}) &= Var(\mathbb{E}[\mathbb{1}_{(X\in H)}|\Theta]) + \mathbb{E}[Var(\mathbb{1}_{(X\in H)}|\Theta)] \\ &\geq Var(\mathbb{E}[\mathbb{1}_{(X\in H)}|\Theta]) \\ &= Var(\mathbb{P}(\eta + R\Lambda\Theta \in \mathcal{A}|\Theta)) \end{aligned} \quad (2.22)$$

In practice, the unbiased estimator is :

$$\frac{1}{n}\sum_{i=1}^{n} P(\eta + R\Lambda\theta_i \in H)$$

where $\{\theta_1, ..., \theta_n\}$ is a sample of n random vectors chosen uniformly on the set S^{d-1}.

To compute $\mathbb{P}(\eta+R\Lambda\theta \in H)$, we have to find the region for R such that $\eta+R\Lambda\theta \in H$ for $\Theta = \theta$, say $\mathsf{R}(\theta)$.

As written above, once the direction θ is fixed :

$$\begin{aligned} & X \in H_j \\ \Rightarrow \quad & \omega_j^\top x > \tau_j \\ \Rightarrow \quad & \omega_j^\top (\eta + R\Lambda\theta) > \tau_j \\ \Rightarrow \quad & R\,\omega_j^\top \Lambda\theta > \tau_j - \omega_j^\top \eta \\ \Rightarrow \quad & R > \frac{\tau_j - \omega_j^\top \eta}{\omega_j^\top \Lambda\theta} \end{aligned} \quad (2.23)$$

The third inequality implies that $\omega_j^\top \Lambda\theta > 0$ because $\tau_j > \omega_j^\top \eta$ since $\eta \in H^c$, otherwise H would no longer be a rare event. The relation $\omega_j^\top \Lambda\theta > 0$ requires that the extension of the line in the direction given by $\Lambda\theta$ eventually intersects the half-space H_j. The scalar product between ω_j and $\Lambda\theta$ is positive so $\Lambda\theta$ is in the same half-space than ω_j and as ω_j is orthogonal to the boundary line of H_j then this line intersects at some point the extension of the vector $\Lambda\theta$.

Thus, $X = \eta + R\Lambda\theta \in H_j \Leftrightarrow R > \frac{\tau_j - \omega_j^\top \eta}{\omega_j^\top \Lambda\theta}$ and $\omega_j^\top \Lambda\theta > 0$. If there are several half-spaces H_j intersecting the line of direction $\Lambda\theta$, the point of interest is the smallest distance to the centre for which we enter the set H.

Then one have that $\eta + R\Lambda\theta \in H$ if and only if $R \in]r(\theta), \infty[$, where

$$r(\theta) = \begin{cases} \min\limits_{\{j|\omega_j^\top \Lambda\theta > 0\}} \dfrac{\tau_j - \omega_j^\top \eta}{\omega_j^\top \Lambda\theta} & if \quad \omega_j^\top \Lambda\theta > 0 \quad \text{for some j} \\ \infty \quad \text{otherwise} \end{cases} \tag{2.24}$$

We use the notation $\mathsf{R}(\theta)$ to denote the interval $]r(\theta), \infty[$ for a fixed θ.

Using this interval, the paper provides an estimator to compute the more general quantity $\mathbb{E}[h(X)\mathbb{1}_{(X\in H)}]$ when X follows the assumptions of elliptical distribution written above and $h(.)$ a real-valued function on H.

Theorem 2. *An unbiased estimator for* $\mathbb{E}[h(X)\mathbb{1}_{(X\in H)}]$ *is given by*

$$h(\eta + \tilde{R}\Lambda\Theta)\mathbb{P}(R \in \mathsf{R}(\Theta)|\Theta) \tag{2.25}$$

where $\tilde{R}$ *has the conditional distribution* $(R|R \in \mathsf{R}(\Theta), \Theta)$. *The ratio of the second moments of this estimator and the crude Monte Carlo estimator is bounded above by*

$$\max_{\|\theta\|=1} \mathbb{P}(R \in \mathsf{R}(\theta)) \tag{2.26}$$

The last part of the theorem means that it always holds :

$$\frac{\mathbb{E}[h(\eta + \tilde{R}\Lambda\Theta)^2\mathbb{P}(R \in \mathsf{R}(\Theta)|\Theta)^2]}{\mathbb{E}[h(X)^2\mathbb{1}_{(X\in H)}]} \leq \max_{\|\theta\|=1} \mathbb{P}(R \in \mathsf{R}(\theta)) \tag{2.27}$$

In fact, the bound $\max_{\|\theta\|=1} \mathbb{P}(R \in \mathsf{R}(\theta))$ is conservative. For a fixed $\Theta = \theta$, the gain of the conditional sampling is

$$\frac{\mathbb{E}[h(\eta + \tilde{R}\Lambda\theta)^2\mathbb{P}(R \in \mathsf{R}(\theta))^2|\Theta = \theta]}{\mathbb{E}[h(X)^2\mathbb{1}_{(X\in\mathcal{A})}|\Theta = \theta]} = \mathbb{P}(R \in \mathsf{R}(\theta)) \tag{2.28}$$

The average efficiency gain for every $\Theta = \theta$ is then $\mathbb{E}[\mathbb{P}(R \in \mathsf{R}(\Theta)|\Theta)] = P(X \in H)$.

For $\tilde{R}$, one can sample it from $R|R \in \mathsf{R}(\theta)$ by setting

$$\tilde{R} = F_R^{-1}(F_R(r(\theta)) + U[1 - F_R(r(\theta))]) \tag{2.29}$$

where U is uniformly distributed in [0,1]. This is the truncated distribution of R above $r(\theta)$.

2.3.3 Representation of rare events

For the events $H_j = \{x \in \mathbb{R}^d : \omega_j^\top x > \tau_j\}$, if we replace τ_j by $m\tau_j$, the event becomes rarer as $m \to \infty$. But in the paper, instead of increasing τ_j, they prefer consider the size of X getting smaller. They define the following sequence of decreasing vectors :

$$X^m = \frac{\eta}{m} + \frac{R\Lambda\Theta}{m^a} \quad m \geq 1 \quad a > 0 \tag{2.30}$$

with a parameter a to tune the rate of decay of the $(X^m)_m$.

We quickly present this new way of viewing things.

For η, R, Λ, Σ and Θ, the same definitions are kept. X has the density function $f_X(x) = |\Sigma|^{-\frac{1}{2}} g_R((x-\eta)^\top \Sigma^{-1}(x-\eta))$ where g_R is defined in (2.20).

In the case of X^m, it has the density function :

$$\begin{aligned}
f_{X^m} &= |\Sigma|^{-\frac{1}{2}} \quad g_{\frac{R}{m^a}}((x-\frac{\eta}{m})^\top \Sigma^{-1}(x-\frac{\eta}{m})) \\
&= |\Sigma|^{-\frac{1}{2}} \quad \frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}} \sqrt{(x-\frac{\eta}{m})^\top \Sigma^{-1}(x-\frac{\eta}{m})}^{-(d-1)} f_{\frac{R}{m^a}}(\sqrt{(x-\frac{\eta}{m})^\top \Sigma^{-1}(x-\frac{\eta}{m})}) \\
&= |\Sigma|^{-\frac{1}{2}} \quad \frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}} (m^a)^{d-1} \sqrt{m^{2a}(x-\frac{\eta}{m})^\top \Sigma^{-1}(x-\frac{\eta}{m})}^{-(d-1)} m^a f_R(\sqrt{m^{2a}(x-\frac{\eta}{m})^\top \Sigma^{-1}(x-\frac{\eta}{m})}) \\
&= |\Sigma|^{-\frac{1}{2}} (m^a)^d g_R(m^{2a}(x-\frac{\eta}{m})^\top \Sigma^{-1}(x-\frac{\eta}{m}))
\end{aligned} \tag{2.31}$$

One would like to work with a random vector which is centered so we change the measure so that X^m has mean zero under the new measure.

$$\mathbb{E}[h(X^m)\mathbb{1}_{(X^m \in H)}] = \mathbb{E}^0[h(X^m)\frac{f_m(X^m)}{f_m^0(X^m)}\mathbb{1}_{(X^m \in H)}] \tag{2.32}$$

where $\mathbb{E}^0$ stands for the probability measure $\mathbb{P}^0$ under which X^m is distributed as $m^{-a}R\Lambda\Theta$. Hence, if we come back to the writing of f_{X^m} we take $f^0_{X^m} = |\Sigma|^{-\frac{1}{2}} (m^a)^d g_R(m^{2a} x^\top \Sigma^{-1} x)$

In sampling in the new distribution f_m^0, the unbiased estimator in Theorem 2 can be written as

$$h_m(\tilde{R}\Lambda\Theta)\mathbb{P}(R \in m^a \mathsf{R}(\Theta)|\Theta) \tag{2.33}$$

where $\tilde{R}$ is distributed as R given that $R \in m^a \mathsf{R}(\Theta)$ and the function h_m is

$$h_m(x) = h(m^{-a}x)\frac{g((x-m^{a-1}\eta)^\top \Sigma^{-1}(x-m^{a-1}\eta))}{g(x^\top \Sigma^{-1} x)}$$

2.3.4 Algorithm

We also summarise below the approach in practice to approximate the expectation of the estimator (2.33) :

For each realisation θ_i of Θ_i ($i \in \{1, ..., n\}$) :
1. Sample $\Theta = \theta$ from the uniform distribution on the unit sphere S^{d-1}
2. Find $\mathsf{R}(\theta) = \{r \geq 0 | r\Lambda\theta \in H\}$ (We note that now $r(\theta)$ is computed the same way as above but with $\mu = 0$)
3. If $\mathsf{R}(\theta)$ is nonempty then
   Sample $\tilde{R}$ from $R|R \in \mathsf{R}(\theta)$ by using (2.29)
   Compute $\mathrm{T} = h_m(\tilde{R}\Lambda\Theta)\mathbb{P}(R \in m^a\mathsf{R}(\Theta)|\Theta)$
   Else set $\mathrm{T} = 0$
   End if
4. Deliver T

Compute $\frac{1}{n}\sum_{i=1}^{n} t_i$ where $(t_i)_i$ are the realisations of the $T_i \sim T$

The algorithm is simplified if we are only interested in $\mathbb{P}(X \in H)$ by taking $h \equiv 1$. This is what we will do in practice for comparisons between estimators (2.8) and (2.33) from a theoretical and numerical point of view.

2.4 Comparison between the ALOE sampling and the directional sampling

2.4.1 Distribution assumption

In the ALOE sampling (2.2), the proposed method assumes that the random variable X follows a Gaussian distribution although Theorem 1 holds so long as we can sample conditionally on any one event H_j and then determine which other rare events also occur. This is due to the difficulty of sampling in the conditional distribution $(X|\omega^\top X \geq \tau)$ for a general distribution F_X. If we return to (2.2.3), the necessary condition for returning a realisation of a random variable with the distribution $(X|\omega^\top X \geq \tau)$ is to know the distribution function of $\omega^\top X$ in order to obtain Y as the inverse of this truncated distribution function above τ. However, in the case where X is not a Gaussian vector, it is often complicated to know the distribution followed by $\omega^\top X$. There are though algorithms using Markov Chains Monte Carlo which manage to obtain almost i.i.d samples from distributions of the form $(X|h(X) > \tau)$ where $h(.)$ a real-valued function possible to evaluate and which will be discussed further below in the report.

The directional sampling (2.3) has, at first sight, less restrictive assumptions. It allows to tackle a larger number of distributions than the single Gaussian one (all those characterized as elliptical distribution) but this requires knowledge of the distribution F_R of the generating random variable R. This is for example the case for t-distributions and Gaussian distributions as shown in (2.3.1). We then use the same trick as in (2.2.3) by simulating with (2.29). Another advantage of this algorithm is

that it does not require knowledge of the $(P_j)_j$, which can be difficult to compute when the distribution is not Gaussian. However, unlike Theorem 1, Theorem 2 does not apply to every distribution F_X since it must be elliptical. Thus, if we are able to sample conditionally on any one event H_j for the distribution of interest as discussed in the above paragraph the ALOE estimator can be seen as the most general one.

2.4.2 Efficiency

Both estimators verify properties defined in (1.1). The ALOE estimator (2.2) is strongly efficient while the directional estimator (2.3) is only weakly efficient. We are interested here only in the estimators used to approximate the quantity $\mu = \mathbb{P}(X \in H)$. The results in [AC18] and [AK18] related to these properties are written below.

Corollary 1. *Let* $\hat{\mu}_{\alpha^*} = \frac{\bar{\mu}}{n}\sum_{i=1}^{n}\frac{1}{S(X_i)}$ *where* $X_i \sim q_\alpha^*$ *with* q_α^* *defined in (2.6). Then* $Var(\hat{\mu}_{\alpha^*}) \leq \frac{\bar{\mu}^2}{4n}$. *If* $\underline{\mu} \geq \frac{\bar{\mu}}{2}$ *then also* $Var(\hat{\mu}_{\alpha^*}) \leq \frac{\mu(\bar{\mu}-\mu)}{n}$. *Similarly,*

$$\frac{Var(\hat{\mu}_{\alpha^*})}{\mu^2} \leq \frac{1}{n}\min\left\{\frac{\bar{\mu}}{\underline{\mu}} - 1, J-1\right\} \leq \frac{J-1}{n} \tag{2.34}$$

This corollary gives us the property of strong efficiency for (2.8) since it implies in particular for J fixed :

$$\lim_{\mu\to 0}\frac{Var(\hat{\mu}_{\alpha^*})}{\mu^2} < \infty \tag{2.35}$$

Theorem 3. *For the directional estimator*

$$\hat{\mu}_d = \mathbb{P}(R \in m^a\mathsf{R}(\Theta)|\Theta) \tag{2.36}$$

if R *has a Weibull-like distribution i.e*

$$f_R(r) = \alpha_1 r^{\beta_1} e^{-\alpha_2 r^{\beta_2}} \tag{2.37}$$

for some $\alpha_1, \alpha_2, \beta_2 > 0$ *and* β_1 *then we have*

$$\lim_{m\to\infty}\frac{\log(\mathbb{E}[\mathbb{P}(R \in m^a\mathsf{R}(\Theta)|\Theta)^2])}{\log(\mathbb{E}[\mathbb{P}(R \in m^a\mathsf{R}(\Theta)|\Theta)]^2)} = 1 \tag{2.38}$$

For example, if $X = m^{-a}R\Lambda\Theta \sim \mathcal{N}(0,\Sigma)$ then $R \sim \sqrt{\chi_d^2}$ which gives

$$f_R(r) = \frac{1}{2^{\frac{d}{2}-1}\Gamma(\frac{d}{2})} r^{d-1} e^{-\frac{r^2}{2}} \tag{2.39}$$

R does follow a Weibull-like distribution. So in the Gaussian case the provided estimator is asymptotically optimal or weakly efficient.

To visualise these two results numerically, we construct an example. We choose the dimension $d = 20$ and set the number of rare events $J = 2000$. We take $X \sim \mathcal{N}(0, I_d)$. To define the $(H_j)_j$, the unit vectors $(\omega_j)_j$ are chosen at random and the $(\tau_j)_j$ are all set to a common value τ. Since $X \sim \mathcal{N}(0, I)$, we have from (2.3) $P_j = \Phi(-\tau)$ for all j and the union bound $\bar{\mu} = \sum P_j = J\Phi(-\tau)$. To represent the evolution of quantities (2.35) and (2.38) in our numerical example, we build K similar problems by keeping the same constants but simply changing τ and as $\tau = -\Phi^{-1}(\frac{\bar{\mu}}{J})$ we rather vary $\log_{10}$ of the union bound $\bar{\mu}$ to control the order of magnitude of μ. Thus we construct K identical problems by just varying $\log_{10}(\bar{\mu})$ from -1 to $-K$ in steps of -1.

We try first the directional estimator from (2.3) and we compute the ratio of logarithms appearing in the quantity (2.38) for each problem. We note that, in order not to make the task more complex, we have set the parameter $a = 1$ so it becomes equivalent to describe this ratio as a function of τ instead of m. We denote $l(\tau)$ this ratio.

$$l(\tau) = \frac{\log(\mathbb{E}[\mathbb{P}(R \in \mathsf{R}(\Theta)|\Theta)^2])}{\log(\mathbb{E}[\mathbb{P}(R \in \mathsf{R}(\Theta)|\Theta)]^2)} \tag{2.40}$$

with $\mathsf{R}(\Theta) =]r(\Theta), \infty[$ and $r(\Theta)$ depends on τ according to its definition (2.24).

Theorem 3 gives us $\lim_{\tau \to \infty} l(\tau) = 1$. For $K = 130$, Figure 2.2 shows the estimated values of this ratio as a function of τ. It seems to well describe the expected behaviour. To better visualise the orders of magnitude of μ associated with such choices of τ, Figure 2.3 shows the same points, this time with $x = \log_{10}(\bar{\mu})$.

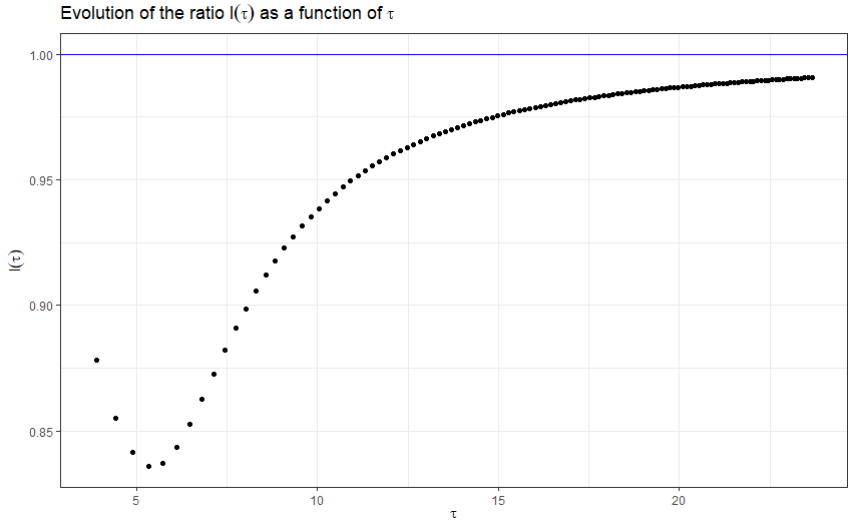

Figure 2.2: Results of the computation of the ratios $l(\tau)$ where the first and the second moments have been estimated from 1000 samples in a log-log plot

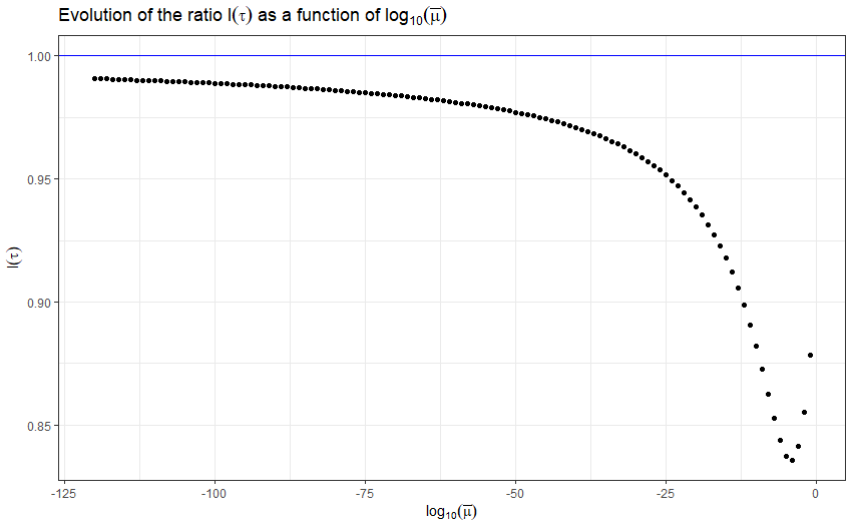

Figure 2.3: Results of the computation of the ratios $l(\tau)$ where the first and the second moments have been estimated from 1000 samples in a log-log plot

Similarly, we want to observe the strong efficiency of the ALOE estimator (2.2). We also compute the ratio $r(\tau) = \frac{Var(\hat{\mu}_{\alpha*})}{\mu^2}$ which is a function of τ for each problem. This time we take $K = 60$ so that we can always obtain values that are numerically calculable. Corollary 1 states this quantity remains bounded for J fixed even when $\tau \to \infty$ which means $\mu \to 0$. Figure 2.4 shows the estimated values of this ratio as a function of τ or $\log_{10}(\bar{\mu})$. $r(\tau)$ does not tend towards a very large value when the order of magnitude of μ becomes very small (always less than or equal to that of $\bar{\mu}$), in particular here it tends to 0 when the events $(H_j)_j$ become particularly rare. As τ increases, the events $(H_j)_j$ overlap less and become almost disjoint. In this configuration, μ is really close to $\bar{\mu}$ with almost zero variance.

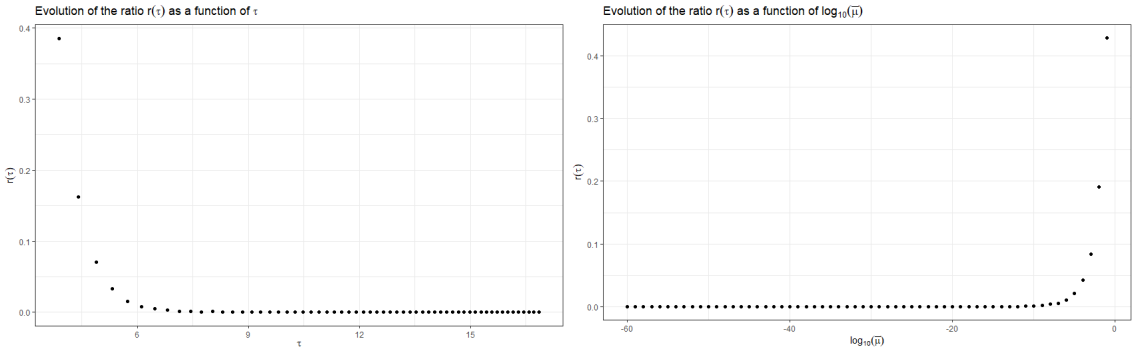

Figure 2.4: Results of the computation of the ratios $r(\tau)$ where the first and the second moments have been estimated from 1000 samples in a log-log plot

The paper [AK18] does not specify whether or not the directional estimator $\hat{\mu}_d$ provided is strongly efficient. We compute in our example $r_d(\tau) = \frac{Var(\hat{\mu}_d)}{\mathbb{E}[\hat{\mu}_d]^2}$ as we did for the ALOE estimator. Figure 2.5 shows that this quantity does not seem to be bounded independently of μ. The largest τ considered is associated with a union

bound $\bar{\mu}$ of order 10^{-60} which is already quite small. In view of the results, one could then assume that the directional estimator $\hat{\mu}_d$ is probably not strongly efficient at least for the standard Gaussian distribution.

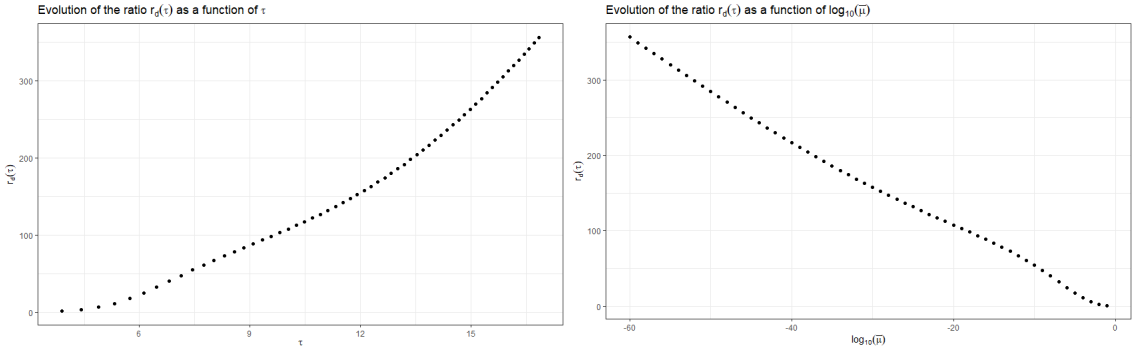

Figure 2.5: Results of the computation of the ratios $r_d(\tau)$ where the first and the second moments have been estimated from 1000 samples in a log-log plot

Chapter 3

Union of rare events where boundary lines are defined by the general equation $h(x) = \tau$

One possibility to generalise the problem (2.1), is to define the events $(H_j)_j$ not only through a boundary represented by a line of equation $\omega^\top x = \tau$ but with a boundary of equation $h(x) = \tau$ where $h(\cdot)$ a general function. Furthermore, it would be interesting to be able to estimate the probability of the union of these events for a maximum of different distributions and not only Gaussian ones. To do this, we take one of the two estimators we studied earlier and modify the procedure associated with it to address this more general problem. As discussed above, if we can overcome the difficulty of simulating samples (X_i) in the distribution $(X|h(X) \geq \tau_j)$ and computing the $(P_j)_j$, the ALOE estimator can be generalized to any distribution unlike the directional estimator. Moreover, it has a stronger property since it is strongly efficient while the directional estimator is a priori only weakly efficient at least for the Gaussian distribution. For these reasons, we will work with the ALOE estimator and will no longer study the directional estimator.

The new rare events $(H_j)_j$ are redefined as follows

$$H_j = \{x : h_j(x) \geqslant \tau_j\} \tag{3.1}$$

where $h_j : \mathbb{R}^d \to \mathbb{R}$ is a general function that we assume we can evaluate at any point $x \in \mathbb{R}^d$

To adapt the ALOE estimator, we need to be able to estimate all the probabilities (P_j) associated with these events and simulate a number of points in the distribution $(X|h(X) \geq \tau_j)$ for each H_j. In order to do this, we will put into practice an adaptive multilevel splitting method presented in the paper [GHM11].

3.1 Estimation of the probability of the rare event $h(X) \geq \tau$ - Method

We would like to estimate $p = \mathbb{P}(h(X) \geq \tau)$ knowing that this probability is very low. Thus, for the reasons given in the introduction, we cannot use the crude CMC approach. A possible method explained in detail in the paper [GHM11] is the adaptive multilevel splitting. We fix a specific number of increasing levels $-\infty = L_0 < L_1 < ... < L_m = \tau$ and we rewrite p by making all these levels appear thanks to the conditional probabilities :

$$p = \mathbb{P}(h(X) \geq \tau) = \prod_{k=1}^{m} \mathbb{P}(h(X) > L_k | h(X) > L_{k-1}) \tag{3.2}$$

Then the probabilities $(p_k)_k = (\mathbb{P}(h(X) > L_k | h(X) > L_{k-1}))_k$ are estimated iteratively.

To do this, we start with an i.i.d. sample $(X_1, ..., X_N)$ from the distribution π. π is the probability distribution of X on the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. It is assumed we can draw samples from this distribution. These samples are called particles in the paper [GHM11]. The idea is to "evolve" all these "particles" until they all land in the area of interest $\{x : h(x) \geq \tau\}$ and ideally follow the distribution $(X | h(X) \geq \tau_j)$.

We initialize $L_0 = -\infty$ and $X_1^1 = X_1, ..., X_N^1 = X_N$. For $k = 1, ..., m$, we set

$$L_k = \min(h(X_1^k), ..., h(X_N^k)) \tag{3.3}$$

Thus the levels are defined adaptively rather than fixed beforehand. And for all $i = 1, ..., N$, we set

$$X_i^{k+1} = \begin{cases} X_i^k & \text{if } h(X_i^k) > L_k \\ X^* \sim (X | h(X) > L_k) & \text{if } h(X_i^k) = L_k \end{cases} \tag{3.4}$$

where X^* is ideally independent from of $(X_1^k, ..., X_N^k)$.

At each step, if we try to estimate $p_k = \mathbb{P}(h(X) > L_k | h(X) > L_{k-1})$ from the simulated N particles, L_k is chosen such that we always have the estimation $p_k \approx 1 - \frac{1}{N}$ since there is only one particle that does not satisfy the condition $h(X_i^k) > L_k$. Thus, we continue to iterate the above steps until $k = m = \max\{k : L_k \leq \tau\}$ and define for the estimator of p

$$\hat{p} = \left(1 - \frac{1}{N}\right)^M \tag{3.5}$$

where M is a random variable and m its realisation

One of the difficulties of this algorithm is to simulate according to the distribution $(X|h(X) > L_k)$ independently of the samples $(X_1^k, ..., X_N^k)$. To do so ideally is in fact impossible in general. [GHM11] proposes a method using Markov Chain Monte Carlo techniques to do so approximately. We first need a π-symmetric and one-step π-irreducible kernel K. For that we choose to use a Metropolis-Hasting kernel K based on a one-step π-irreducible proposal distribution $Q(.|.)$. In practice we take the Q associated with the Gaussian random walk which is in general a good choice if π is a Gaussian distribution for instance.

Then, once we have the kernel K, considering we know $L_1 = l_1$,$L_2 = l_2, ...$ and the sets $A_k = \{x : h(x) > l_k\}$, we define π_k the normalised restriction of π on A_k :

$$\pi_k(dx) = \frac{1}{\pi(A_k)} \mathbb{1}_{A_k}(x)\pi(dx) \tag{3.6}$$

and also the transition kernel K_k :

$$K_k(x, dx') = \mathbb{1}_{A_k^c}(x)\delta_x(dx') + \mathbb{1}_{A_k}(x)(K(x, dx')\mathbb{1}_{A_k}(x') + K(x, A_k^c)\delta_x(dx')) \tag{3.7}$$

In practice, what the kernel K_k does is equivalent to this : K proposes a transition from x to x' and if $h(x') > l_k$ then it is accepted otherwise x stays at the same place. But by writing the definition of K_k formally, we can see that π_k is invariant by K_k. In addition, as explained in [GHM11] the results in [Tie94] and [Mey93] ensure that for any initial distribution ν such that $\nu(A_k) = 1$

$$\|\nu K_k^n - \pi_k\| \underset{n\to\infty}{\longrightarrow} 0 \tag{3.8}$$

where $\|.\|$ is the total variation norm

For clarity, if $X \sim \pi$ then the distribution π_k corresponds to the distribution $(X|h(X) > L_k)$. Thus, this is the distribution we want to aim for to distribute our new value X^*. We unroll the first step of our algorithm to better understand the step (3.4) in practice.

We start with an i.i.d sample $(X_1, ..., X_N)$ from π. We initialize $X_1^1 = X_1, ..., X_N^1 = X_N$. We assume without loss of generalities that $h(X_1^1) < ... < h(X_N^1)$. Then $L_1 = h(X_1^1)$. We iterate $X_2^2 = X_2^1, ..., X_N^2 = X_N^1$.

At $L_1 = l_1$ fixed and known, $(X_2^2, ..., X_N^2)$ is i.i.d with distribution π_1. We want our new value X_1^2 to be added to this sample to also follow this distribution while remaining independent of it. To do this, we choose at random one sample X_i^2 and set $X_0^* = X_i^2$. To achieve independence, we apply iteratively the kernel K_1. As $X_0^* = X_i^2 = x_i^2$, the initial measure is $\delta_{x_i^2}$ which verifies $\delta_{x_i^2}(A_1) = 1$. Thus, (3.8) ensures that

$$\left\| \int \delta_{x_i^2} K_1^n - \pi_1 \right\| \underset{n\to\infty}{\longrightarrow} 0 \tag{3.9}$$

So after several iterations by the kernel K_1, we can consider our new resulting "particle" X^* as following the distribution π_1 while being "almost" independent from X_i^2. If we note $X_1^2 = X^*$, the new random variables constructed $(X_1^2, ...X_N^2)$ are ideally i.i.d of common distribution π_1. We then apply this principle iteratively until we obtain a sample of i.i.d. random variables of distribution π_m.

As for the number of iterations needed to consider X^* "almost" independent of its initial value, it is rather difficult to give a general guideline as this is often dependent on the given problem configuration. Later, in the practical part, we will test several number of iterations (between 5 and 40) for our problem with a fixed number of particles N to observe its influence on the relative error of the estimation.

However, for the theoretical study of the algorithm, we will consider that at each iteration, $(X_1^k, ...X_N^k)$ is a sample of random variables perfectly i.i.d of common distribution π_k. This is never true in practice but it would increase the difficulty of the analysis to take into account the convergence properties of the MCMC method used. The algorithm studied is then called the idealized algorithm.

3.2 Properties of the estimator of $\mathbb{P}(h(X) > \tau)$ for the idealized algorithm

The properties of the estimator $\hat{p}$ (3.5) in the case of the idealized algorithm are briefly presented below, they are presented in more detail with their proof in [GHM11].

Proposition 1. *The estimator $\hat{p}$ (3.5) is a discrete random variable taking values in*

$$\left\{1, \left(1 - \frac{1}{N}\right), \left(1 - \frac{1}{N}\right)^2, ...\right\}$$

with probability

$$\mathbb{P}\left(\hat{p} = \left(1 - \frac{1}{N}\right)^k\right) = \frac{p^N(-N\log(p))^k}{k!}, \quad k = 0, 1, 2, ... \tag{3.10}$$

It follows that

$$\mathbb{E}[\hat{p}] = p \quad and \quad Var(\hat{p}) = p^2(p^{-\frac{1}{N}} - 1) \tag{3.11}$$

As mentioned in (1.2), the unbiased crude CMC estimator constructed with N replications has a coefficient of variation

$$CV_{CMC}^2 = \frac{1}{Np} \tag{3.12}$$

In comparison, the new unbiased estimator $\hat{p}$ shows an improvement with

$$CV^2 = \frac{Var(\hat{p})}{p^2} = (p^{-\frac{1}{N}} - 1) \approx \frac{-\log(p)}{N} \tag{3.13}$$

Where the crude CMC method required a sample size of the order of magnitude of $\frac{1}{p}$ to have a reasonable accuracy, this can be achieved with only a sample size of the order of magnitude $-\log(p)$, a reduction of a factor $-p\log(p)$. [GHM11] qualifies this result by adding that the computational complexity of the algorithm is $\mathcal{O}(-N\log(N)\log(p))$ versus only $\mathcal{O}(N)$ for the crude CMC method but for a probability p small enough, the reduction in the variance outweights the increased computational costs.

[GHM11] derives from the above results confidence intervals for p. We fix α a number between 0 and 1 and we note $Z_{1-\frac{\alpha}{2}}$ the quantile of order $1-\frac{\alpha}{2}$ of the standard Gaussian distribution.

Proposition 2. *$I_{1-\alpha}(p) = [\hat{p}_-, \hat{p}_+]$ is a $100(1-\alpha)\%$ asymptotic confidence interval for p where*

$$\hat{p}_\pm = \hat{p}\exp\left(\pm\frac{Z_{1-\frac{\alpha}{2}}}{\sqrt{N}}\sqrt{-\log(\hat{p}) + \frac{Z^2_{1-\frac{\alpha}{2}}}{4N}} - \frac{Z^2_{1-\frac{\alpha}{2}}}{2N}\right) \tag{3.14}$$

We can simplify the interval by neglecting the terms in $\frac{1}{N}$ for N large enough

$$\mathbb{P}\left(\hat{p}\exp\left(-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{-\log(\hat{p})}{N}}\right) \leq p \leq \hat{p}\exp\left(+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{-\log(\hat{p})}{N}}\right)\right) \approx 1-\alpha \tag{3.15}$$

The main argument used for Proposition 2 is that the distribution of $\hat{l} = \log(\hat{p})$ can be approximated by a Gaussian $\mathcal{N}(l, -\frac{l}{N})$ for N large where $l = \log(p)$.

3.3 New estimator for the probability of union of rare events for general boundary lines

The adaptive multilevel splitting method discussed in the two previous sections allows us to obtain for each rare event H_j an estimate of its probability P_j as well as N realizations of more or less independent random variables following the distribution $(X|h_j(X) > \tau_j)$ where X is a random variable following the initial distribution π. We can modify the ALOE estimator in replacing the probabilities P_j by their estimators $\hat{P}_j$. Following the assumption made in [GHM11], we suppose for the theoretical study that the sample $(X_1^j, ..., X_N^j)$ simulated at the end of the algorithm is ideally i.i.d with distribution $(X|h_j(X) > \tau)$. As we previously stated, it is impossible in general to simulate exactly according to $(X|h(X) > L_m)$ and we can just do so approximately using the Markov Chain Monte Carlo technique. It would be an interesting project, to look at the convergence properties of the MCMC algorithm in order to better characterise the real density followed by the samples. We therefore consider that we also have the ability to draw independent $(X_i^j)_i$ following all the conditional distribution $q_j(x) = \frac{p(x)H_j(x)}{P_j}$ for each j. $p(\cdot)$ is the p.d.f of the

distribution π, it is suppose known as we need it to apply the multilevel splitting method.

The same notations are used as in section (2.2), the new importance sampling distribution conditionally on $\hat{P}_1, ..., \hat{P}_J$ is then

$$\hat{q}_{\alpha^*|\hat{P}_1,...,\hat{P}_J}(x) \approx \sum_{j=1}^{J} \frac{\hat{P}_j}{\sum\limits_k \hat{P}_k} \frac{p(x)H_j(x)}{P_j} \tag{3.16}$$

We use the sign $\approx$ instead of the equality since as mentioned above, the new density is not really this one. The samples $(X_i^j)_i$ drawn with probability $\frac{\hat{P}_j}{\sum_k \hat{P}_k}$ do not really follow the distribution q_j.

The new mixture importance sampling estimate of μ based on n draws where x_i are i.i.d realisations of $X \sim \hat{q_{\alpha^*}}$ is

$$\hat{\mu}_{\alpha^*} = \frac{1}{n} \sum_{i=1}^{n} \frac{p(x_i)H_{1:J}(x_i)}{\sum_{j=1}^{J} \frac{\hat{P}_j}{\sum\limits_k \hat{P}_k} \frac{p(x_i)H_j(x_i)}{P_j}} = \frac{\sum\limits_k \hat{P}_k}{n} \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{J} \frac{\hat{P}_j}{P_j} H_j(x_i)} \tag{3.17}$$

We can call this new estimator ALOE-MLS, as it is a mixture between the ALOE and the adaptive multilevel splitting method.

Proposition 3. *Let $\hat{\mu}_{\alpha^*}$ given by (3.17). Then*

$$\mathbb{E}[\hat{\mu}_{\alpha^*}] \approx \mu \tag{3.18}$$

and

$$Var(\hat{\mu}_{\alpha^*}) \approx \frac{1}{n} \mathbb{E}\left[\sum_k \hat{P}_k \int_H \frac{1}{\sum_{j=1}^{J} \frac{\hat{P}_j}{P_j} H_j(x)} p(x)\, \mathrm{d}x \right] - \frac{\mu^2}{n} \tag{3.19}$$

Proof : The proposition is proved in the appendix.

For the variance, the expectation $\mathbb{E}$ relates to the random variables $(\hat{P}_j)_j$. Once again, the sign $\approx$ is to be understood in the sense that we are studying the ideal algorithm where the samples $(X_i^j)_i$ are i.i.d of common distribution q_j, which is not the case in practice. Thus, in reality it is possible that we have a little bias on the expectation as the number of chain iterations for each step in the multilevel splitting is finite.

We are interested in finding a more explicit upper bound for the variance of the estimator in the case of the idealised algorithm. To do this, we would like to use the same calculation trick that appears in Theorem 1, but in order to apply it we must first find a common lower bound to the quantities $\frac{\hat{P}_j}{P_j}$ independent of j. Indeed, if there is a ϵ such that $\frac{\hat{P}_j}{P_j} \geq \epsilon$ for all j, we would have

$$\begin{aligned}
\mathbb{E}\left[(\sum_k \hat{P}_k)\int_H \frac{1}{\sum_{j=1}^J \frac{\hat{P}_j}{P_j}H_j(x)}p(x)\,\mathrm{d}x\right] &\leq \frac{1}{\epsilon}\sum_k \mathbb{E}\left[\int_H \frac{\hat{P}_k}{\sum_{j=1}^J H_j(x)}p(x)\,\mathrm{d}x\right] \\
&= \frac{1}{\epsilon}\sum_k \int_H \frac{\mathbb{E}\left[\hat{P}_k\right]}{\sum_{j=1}^J H_j(x)}p(x)\,\mathrm{d}x \\
&= \frac{1}{\epsilon}\sum_k P_k \int_H \frac{1}{\sum_{j=1}^J H_j(x)}p(x)\,\mathrm{d}x \\
&= \frac{\bar{\mu}}{\epsilon}\sum_{s=1}^J \frac{T_s}{s}
\end{aligned} \tag{3.20}$$

Here, the assumptions to interchange the integral and the expectation which is an infinite sum are verified. The upper bound of (3.19) would then be

$$\frac{1}{n}\frac{\bar{\mu}}{\epsilon}\sum_{s=1}^J \frac{T_s}{s} - \frac{\mu^2}{n} \leq \frac{\bar{\mu}}{n\epsilon}(1-T_0) - \frac{\mu^2}{n} = \frac{\mu}{n}\left(\frac{\bar{\mu}}{\epsilon} - \mu\right) \tag{3.21}$$

This is in fact the same as that established in Theorem 1 for the original ALOE estimator except the appearance of ϵ which can deteriorate the bound.

However, there is not a $\epsilon > 0$ which is almost surely a lower bound of all $(\frac{\hat{P}_j}{P_j})_j$ unless we change the implementation of the multilevel splitting algorithm computing the $(\hat{P}_j)_j$. We add a stopping condition on the number of iterations k and we define

$$\hat{P}_j = (1 - \frac{1}{N})^{M_j} \tag{3.22}$$

where $M_j = \min(\{k : L_k \leq \tau_j\}, I^j_{\max})$ and $I^j_{\max}$ the maximum number of iterations permitted. If $L_{M_j+1} < \tau_j$ then the estimator is biased but if $I^j_{\max}$ is large, the bias can be assumed to be quite small. This additional condition ensures

$$\frac{\hat{P}_j}{P_j} \geq \frac{(1-\frac{1}{N})^{I^j_{max}}}{P_j} \geq \min_j \frac{(1-\frac{1}{N})^{I^j_{max}}}{P_j} = \epsilon \tag{3.23}$$

ϵ is smaller the larger the $(I^j_{\max})$ are. In practice, the number of iterations k needed to have $L_k \geq \tau$ is quite high for very small probability. In our simulations, for a probability of the order of magnitude 10^{-8} and an initial Gaussian distribution, the number of iterations required is of the order of 100,000. Thus, to reduce the bias, $I_{\max}$ should be chosen at least in this range. This implies that ϵ is usually a very conservative lower bound and severely deteriorate the upper bound in (3.21).

For better lower bounds ϵ, it is necessary to accept that inequality can only occur with a certain probability. As stated in (3.2), the distribution of $\hat{l}_j = \log(\hat{P}_j)$ can be approximated by a Gaussian $\mathcal{N}(l_j, -\frac{l_j}{N})$ for N large where $l_j = \log(P_j)$. Then we have

$$\mathbb{P}\left(\frac{\hat{l}_j - l_j}{\sqrt{-\frac{l_j}{N}}} \geq -Z_{1-\alpha}\right) \underset{N\to\infty}{\to} 1-\alpha \tag{3.24}$$

We remind that $Z_{1-\alpha}$ is the quantile of order $1-\alpha$ of the standard Gaussian distribution. This implies that

$$\mathbb{P}\left(\frac{\hat{P}_j}{P_j} \geq e^{-Z_{1-\alpha}\sqrt{\frac{-\log(P_j)}{N}}}\right) \underset{N\to\infty}{\to} 1-\alpha \tag{3.25}$$

and in particular

$$1-\alpha \leq \lim_{N\to\infty} \mathbb{P}\left(\frac{\hat{P}_j}{P_j} \geq e^{-Z_{1-\alpha}\sqrt{\frac{-\log(\min_j P_j)}{N}}}\right) \tag{3.26}$$

For $\alpha = 0.05$, $\epsilon = e^{-1.96\sqrt{\frac{-\log(\min_j P_j)}{N}}}$ is much more closer to 1 for $(P_j)_j$ of order of magnitude 10^{-8} for instance and it is a lower bound with probability of at least 0.95. It is not almost surely so it can not be used for the upper bound of the variance of the estimator but it can give a more precise indication on the real variance.

3.4 Comparison of the new ALOE-MLS estimator with the direct method MLS

We compare the estimator created in section (3.3) to which we will refer as the new ALOE-MLS estimator with the direct estimation of μ using only the adaptive multilevel splitting algorithm. In this method, we directly use the function

$$\Psi(x) = \max_j \frac{h_j(x)}{\tau_j} \tag{3.27}$$

And it would then estimate the probability $\mathbb{P}(\Psi(x) \geq 1) = \mu$. According to the previous results, the estimator obtained by this method is unbiased and its variance is equal to $\mu^2(\mu^{-\frac{1}{kN}} - 1)$. We introduce $k \geq 1$ here to differentiate the total number of particles chosen kN when computing μ directly and N taken as a reference for the number of particles used to calculate a single probability P_j. For example, if we want the same number of particles used for both methods, we would choose $k = J$. If we compute the ratio of the upper bound of $Var(\hat{\mu}_{\alpha^*})$ (3.21) over this variance, we have

$$\frac{\frac{\mu}{n}\left(\frac{\bar{\mu}}{\epsilon} - \mu\right)}{\mu^2(\mu^{-\frac{1}{kN}} - 1)} = \frac{1}{n}\frac{\frac{1}{\epsilon}\frac{\bar{\mu}}{\mu} - 1}{\mu^{-\frac{1}{kN}} - 1} \underset{N\gg 1}{\approx} \frac{1}{n}\frac{\frac{1}{\epsilon}\frac{\bar{\mu}}{\mu} - 1}{\frac{-\log(\mu)}{kN}} = \frac{kN}{n}\frac{\frac{1}{\epsilon}\frac{\bar{\mu}}{\mu} - 1}{-\log(\mu)} \tag{3.28}$$

The term $-\log(\mu)$ in the denominator tends to ∞ when μ tends 0. However, it is more difficult to properly study the asymptotic behaviour of $\frac{1}{\epsilon}\frac{\bar{\mu}}{\mu} - 1$ when μ tends to 0. This is due to ϵ which depends on μ and if we take its definition (3.23), its

asymptotic behaviour is not clear. But if we were able to show that $\frac{1}{\epsilon}\frac{\bar{\mu}}{\mu} - 1 \underset{\mu \to 0}{\to} c \neq \infty$ then it would state that the variance of the new ALOE estimator is smaller for very small probabilities μ.

Another very interesting feature to take into account is the computational complexity. As mentioned above, the expected complexity of the adaptive multilevel splitting algorithm used to estimate a probability p is $\mathcal{O}(-N\log(N)\log(p))$. So if we assume that all the $(P_j)_j$ are in the same order of magnitude i.e $P_j \approx P$, then the resulting complexity of the algorithm used to compute the ALOE-MLS estimator is $\mathcal{O}(-N\log(N)\log(P)J)$ as we iterate the method for each P_j and we take each time the same number of particles N. The direct estimation that we can call the MLS method also has a linear complexity in J since if we assume that the evaluation of the functions h_i in x is $\mathcal{O}(1)$, this is no longer the case for Ψ. Each evaluation of Ψ is linear in J as we have to find the maximum over all the $h_j(x)$. Thus, if the total number of particles used is kN, the computational complexity is $\mathcal{O}(-kN\log(kN)log(\mu)J)$. In order to compare, we can make the approximation that $\mu \approx JP$ in order of magnitude. We can then write the following ratio :

$$\frac{-N\log(N)\log(P)J}{-kN\log(kN)log(\mu)J} = \frac{\log(N)\log(P)}{k\log(kN)\log(\mu)} = \frac{\log(N)\log(P)}{k\log(kN)\log(JP)} = \frac{1}{k}\frac{1}{1+\frac{\log(k)}{\log(N)}}\frac{1}{1+\frac{\log(J)}{\log(P)}} \tag{3.29}$$

The quantities $\frac{1}{1+\frac{\log(k)}{\log(N)}}$ and $\frac{1}{1+\frac{\log(J)}{\log(P)}}$ both tend towards 1 respectively when $N \to \infty$ and when $P \to 0$. If we take only $N = 1000$ and $P = 10^{-6}$ with $J = 10$ and $k = 5$ for instance, we would already have $\frac{1}{1+\frac{\log(k)}{\log(N)}}\frac{1}{1+\frac{\log(J)}{\log(P)}} \approx 0.97$. So we consider

$$\frac{-N\log(N)\log(P)J}{-kN\log(kN)\log(\mu)J} = \frac{\log(N)\log(P)}{k\log(kN)\log(\mu)} \approx \frac{1}{k} \tag{3.30}$$

for N quite large and μ quite small.

If $k \gg 1$, it means that the computational complexity will be better for the ALOE-MLS estimator. For instance, if we decide to run the two algorithms using the same number of particles in total, we will choose $k = J$. For a problem where the number of rare events is quite high, this can give a rather low ratio (3.30). However, this is just an indication and does not mean that this algorithm is necessarily better since possibly an equally good variance can potentially be achieved by the MLS method with a smaller number of particles.

In fact, to take into account both computational complexity and variance it is often common to look at the product of it which defines the efficiency of a Monte Carlo process according to [Ham65]. We could then compare this product for the two algorithms and have a more objectively criterion. However, we return to the same difficulty of comparing the explicit bounds of the variances of the estimators which is mainly due to the complex writing of ϵ. We did not study this criterion further in a theoretical way and we just looked at its behaviour in the case of specific examples.

3.5 Numerical results for the new ALOE-MLS estimator and the direct estimation of μ

3.5.1 First example : Circumscribed polygon

For the first numerical example, we choose the circumscribed polygon with a standard Gaussian distribution that is presented in the paper [AC18]. The rare events are still defined by lines and it is not a very original distribution but it ensures that the new algorithm still works for the more "basic" cases. Moreover, we know the exact value of μ in this case, which will allow us to compute the relative error.

In our example, we have $\mathcal{P}(J, \tau) \subset \mathbb{R}^2$ a regular polygon of $J \geq 3$ sides circumscribed around the circle of radius $\tau > 0$. We have $H = \mathcal{P}^c$ which means it is the intersection of the $(H_j^c)_j$ where $H_j = \{x : \omega_j^\top x \geq \tau\}$ with $\omega_j^\top = (\sin(\frac{2\pi j}{J}), \cos(\frac{2\pi j}{J}))$ for $j = 1, ..., J$. [AC18] explains that we have the following result :

$$1 \geq \frac{\mathbb{P}(X \in H)}{\exp(\frac{-\tau^2}{2})} \geq 1 - \frac{\pi^2 \tau^2}{6J^2} \tag{3.31}$$

for large J

It also outlines that for $J = 360$ and $\tau = 6$, the lower bound is about 0.9995 times the upper bound, so it treats the upper bound as exact. We will make here the same assumption, we take the same J and τ for our numeric example and consider that $\mu = \exp(\frac{-6^2}{2}) = 1.52 \times 10^{-8}$.

We run the new ALOE-MLS algorithm to compute the relative error on its estimate $\hat{\mu}$. Figure 3.1 illustrates the paragraph at the end of section (3.1) in showing the evolution in a logarithm scale of the relative error as a function of the number of iterations of the Markov chain. The relative error decreases as the number of iterations of the chain increases, since this improves the estimations of the probabilities P_j but above all improves the quality of the samples $(X_i^j)_i$ for each j. Similarly, the evolution in logarithmic scale of the relative error as a function of the number of particles used for each probability P_j is summarized in Figure 3.2. These representations make it possible to estimate the slope of the decay. Figure 3.3 summarizes the comparison between the ALOE-MLS and MLS algorithms in this numerical example. The slope of decay seems to be to the advantage of the new ALOE-MLS method, however it should be noted that if one accepts a slightly large relative error (which is still of the order of 10^{-2}), the simple MLS method seems to obtain equivalent results in less computation time. The unit of the computation time is the second.

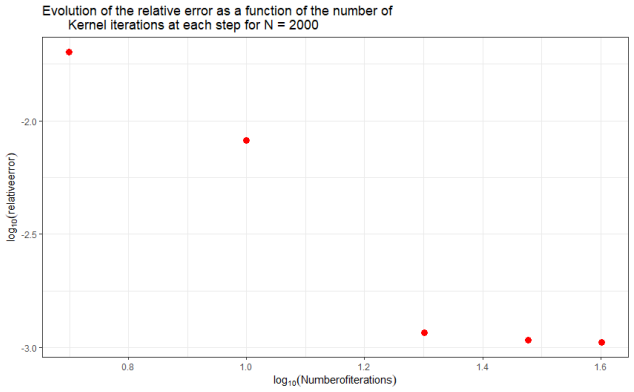

Figure 3.1: Relative error on the estimate $\hat{\mu}$ of the ALOE-MLS algorithm in logarithmic scale for different numbers of kernel iterations at each step in the adaptive multilevel splitting part of the algorithm

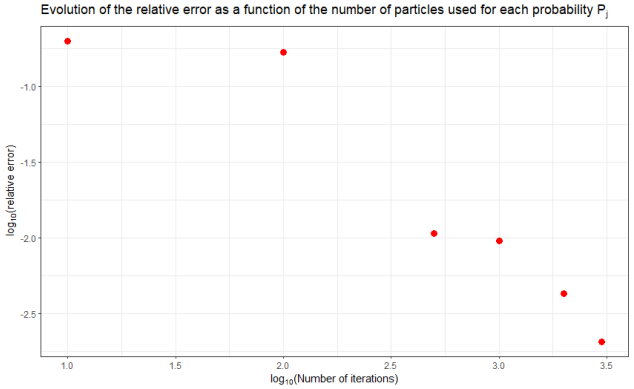

Figure 3.2: Relative error on the estimate $\hat{\mu}$ of the ALOE-MLS algorithm in logarithmic scale for different numbers N of particles

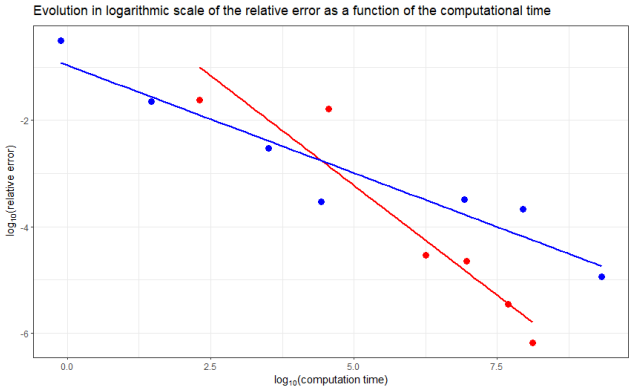

Figure 3.3: Relative error as a function of the computational time. The red dots are obtained from the ALOE-MLS estimator and the blue ones from the direct MLS estimator. The lines are the linear regressions performed on these sets of points. The calculation time was increased by improving the number of particles used in the multilevel splitting step in both cases.

3.5.2 Second example : Boundaries defined by quadratic curves

The second numerical example is constructed with rare events H_j whose boundaries are not defined by a line but by a quadratic curve. We are still in $\mathbb{R}^2$.

$$H_j = \{(x, y) : c(x\sin(\theta_j) + y\cos(\theta_j))^2 + (x\cos(\theta_j) - y\sin(\theta_j)) \geq \tau_j\} \tag{3.32}$$

where c a constant and $0 \leq \theta_j \leq 2\pi$

Figure 3.4 gives a graphical example for $J = 4$. For our simulation, we choose to define the events by evolving θ from 0 to 2π by steps of $\frac{\pi}{8}$, we have then $J = 16$. We keep the standard Gaussian distribution. Figure 3.5 shows the results obtained for the standard deviation of the estimation of μ and the computation time for the two algorithms when we take $\tau_j = \tau = 4$. For $N = 3000$, the new ALOE estimator gives $\hat{\mu} = 2.854815 \times 10^{-7}$. The new ALOE algorithm provides a lower standard deviation than its counterpart for equivalent computing times. In this example, it seems to perform better if we assume that neither algorithm is biased.

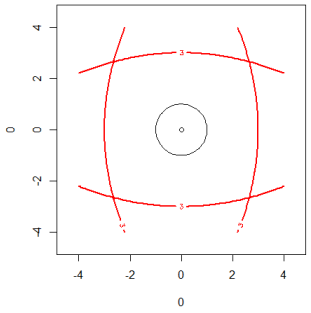

Figure 3.4: Graphical example of our configuration where there are 4 rare events whose boundaries have been defined by the angles $\theta = \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ and $\tau = 3$. The union of the rare events is the complementary of the closed area defined by the intersection of the 4 red lines

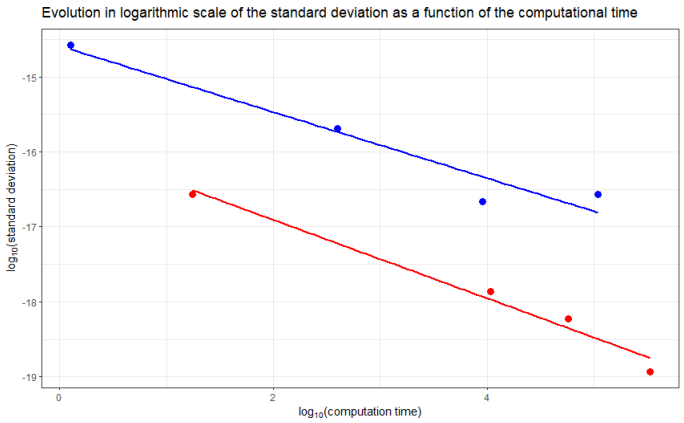

Figure 3.5: Standard deviation of $\hat{\mu}$ as a function of the computation time. The red dots are obtained from the ALOE-MLS estimator and the blue ones from the direct MLS estimator . The lines are the linear regressions performed on these sets of points. The calculation time was increased by improving the number of particles used in the multilevel splitting step in both cases.

Chapter 4

Conclusion

We have introduced a new estimator to compute the probability of the union of rare events which is the mixture of two already existing methods. In theory, this new ALOE estimator can be used for any distribution whose density is known and from which it is possible to draw samples, as well as for configurations where the rare events are described more generally by an equation of the form $h(x) \geq \tau$. It was sought to show that this algorithm has advantages over a known method that uses only adaptive multilevel splitting, which was simply called the MLS method. As possible research directions, we suggest pursuing the theoretical study on the comparison of variances and computational complexity in (3.4). The two numerical examples are encouraging and suggest that, at least in some situations, the newly highlighted estimator ALOE-MLS performs better than the simple MLS. If we wish to continue to study this estimator, it could be interesting to push the theoretical study further. In particular, finding an upper bound on its variance that is not too conservative. More ideally, it would be interesting to study in theory the real behaviour of the new ALOE algorithm taking into account the convergence properties of the Markov chains used in the multilevel splitting parts.

Appendix

Proof of Proposition 3 :

We show that this new estimator is unbiased :

$$\begin{aligned}
\mathbb{E}[\hat{\mu}_{\alpha^*}] &= \mathbb{E}\left[\frac{\sum_k \hat{P}_k}{n}\sum_{i=1}^{n}\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[(\sum_k \hat{P}_k)\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}|\hat{P}_1,...,\hat{P}_J\right]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\sum_k \hat{P}_k)\int_H \frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(x)}\hat{q}_{\alpha^*|\hat{P}_1,...,\hat{P}_J}(x)\,\mathrm{d}x\right] \\
&\approx \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\sum_k \hat{P}_k)\int_H \frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(x)}\sum_{j=1}^{J}\frac{\hat{P}_j}{\sum_k \hat{P}_k}\frac{p(x)H_j(x)}{P_j}\,\mathrm{d}x\right] \\
&\approx \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\int_H \frac{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(x)}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(x)}p(x)\,\mathrm{d}x\right] \\
&\approx \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\int \mathbb{1}_{\{x\in H\}}\,p(x)\,\mathrm{d}x\right] \\
&\approx \mu
\end{aligned} \tag{4.1}$$

where $X_1,...,X_n$ are i.i.d random variables following the distribution $\hat{q}_{\alpha^*|\hat{P}_1,...,\hat{P}_J}$

For the variance,

$$\begin{aligned}
Var(\hat{\mu}_{\alpha^*}) &= Var\left(\frac{\sum_k \hat{P}_k}{n}\sum_{i=1}^{n}\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}\right) \\
&= \frac{1}{n^2}\mathbb{E}\left[Var\left((\sum_k \hat{P}_k)\sum_{i=1}^{n}\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}|\hat{P}_1,...,\hat{P}_J\right)\right] \\
&+ \frac{1}{n}Var\left(\mathbb{E}\left[(\sum_k \hat{P}_k)\sum_{i=1}^{n}\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}|\hat{P}_1,...,\hat{P}_J\right]\right)
\end{aligned} \tag{4.2}$$

According to (4.1),

$$\mathbb{E}\left[(\sum_k \hat{P}_k)\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}|\hat{P}_1,...,\hat{P}_J\right]\approx\mu$$

then

$$Var\left(\mathbb{E}\left[(\sum_k \hat{P}_k)\sum_{i=1}^{n}\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}|\hat{P}_1,...,\hat{P}_J\right]\right)\approx 0$$

And

$$\begin{aligned}
&\mathbb{E}\left[Var\left((\sum_k \hat{P}_k)\sum_{i=1}^{n}\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}|\hat{P}_1,...,\hat{P}_J\right)\right]\\
=&\mathbb{E}\left[(\sum_k \hat{P}_k)^2\,Var\left(\sum_{i=1}^{n}\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}|\hat{P}_1,...,\hat{P}_J\right)\right]\\
=&\mathbb{E}\left[(\sum_k \hat{P}_k)^2\,nVar\left(\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}|\hat{P}_1,...,\hat{P}_J\right)\right]\\
=&\mathbb{E}\left[(\sum_k \hat{P}_k)^2\,n\left(\mathbb{E}\left(\left(\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}\right)^2|\hat{P}_1,...,\hat{P}_J\right)-\mathbb{E}\left(\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}|\hat{P}_1,...,\hat{P}_J\right)^2\right)\right]
\end{aligned}\tag{4.3}$$

We have

$$\mathbb{E}\left[\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}|\hat{P}_1,...,\hat{P}_J\right]^2\approx\frac{\mu^2}{(\sum_k \hat{P}_k)^2}\tag{4.4}$$

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(X_i)}\right)^2|\hat{P}_1,...,\hat{P}_J\right]&\approx\int_H\left(\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(x)}\right)^2\sum_{j=1}^{J}\frac{\hat{P}_j}{\sum_k \hat{P}_k}\frac{p(x)H_j(x)}{P_j}\,\mathrm{d}x\\
&\approx\frac{1}{\sum_k \hat{P}_k}\int_H\frac{1}{\sum_{j=1}^{J}\frac{\hat{P}_j}{P_j}H_j(x)}p(x)\,\mathrm{d}x
\end{aligned}\tag{4.5}$$

Finally,

$$
\begin{aligned}
Var(\hat{\mu}_{\alpha^*}) &\approx \frac{1}{n}\mathbb{E}\left(\left(\sum_k \hat{P}_k\right)^2\left(\frac{1}{\sum_k \hat{P}_k}\int_H \frac{1}{\sum_{j=1}^J \frac{\hat{P}_j}{P_j}H_j(x)}p(x)\,\mathrm{d}x - \frac{\mu^2}{(\sum_k \hat{P}_k)^2}\right)\right) \\
&\approx \frac{1}{n}\mathbb{E}\left((\sum_k \hat{P}_k)\int_H \frac{1}{\sum_{j=1}^J \frac{\hat{P}_j}{P_j}H_j(x)}p(x)\,\mathrm{d}x - \mu^2\right) \qquad (4.6) \\
&\approx \frac{1}{n}\mathbb{E}\left((\sum_k \hat{P}_k)\int_H \frac{1}{\sum_{j=1}^J \frac{\hat{P}_j}{P_j}H_j(x)}p(x)\,\mathrm{d}x - \mu^2\right)
\end{aligned}
$$

Bibliography

[Ham65] Handscomb D. Hammersley J. *Monte Carlo Methods*. London, Methuen, 1965.

[Mey93] Tweedie R. Meyn S. *Markov Chains and Stochastic Stability*. London: Springer, 1993.

[Tie94] L. Tierney. "Markov Chains for Exploring Posterior Distributions". In: *The Annals of Statistics* 22 (1994), pp. 1701–1762. URL: http://dx.doi.org/10.1214/aos/1176325750.

[Fra04] Gabriel Frahm. "Generalized Elliptical Distributions: Theory and Applications". PhD thesis. Universität zu Köln, 2004.

[Sør07] Peter W. Glynn Søren Asmussen. Springer, 2007.

[ABL08] Robert J. Adler, Jose Blanchet, and Jingchen Liu. "Efficient simulation for tail probabilities of Gaussian random fields". In: *2008 Winter Simulation Conference*. 2008, pp. 328–336. DOI: 10.1109/WSC.2008.4736085.

[Dou10] A. Doucet. *A note on efficient conditional simulation of Gaussian distributions*. University of British Columbia, Technical report. 2010.

[GHM11] Arnaud Guyader, Nicolas Hengartner, and Eric Matzner-Løber. "Simulation and Estimation of Extreme Quantiles and Extreme Probabilities". In: *Applied Mathematics & Optimization* 64.2 (Oct. 2011), pp. 171–196. ISSN: 1432-0606. DOI: 10.1007/s00245-011-9135-z. URL: https://doi.org/10.1007/s00245-011-9135-z.

[AK18] Dohyun Ahn and Kyoung Kuk Kim. "Efficient simulation for expectations over the union of half-spaces". In: *ACM Transactions on Modeling and Computer Simulation* 28.23 (2018), pp. 1–20.

[AC18] Yury Maximov Art B. Owen and Michael Chertkov. "Importance sampling the union of rare events with an application to power systems analysis". In: *Electronic Journal of Statistics* 13.1 (2018), pp. 231–254.

[EM21] Víctor Elvira and Luca Martino. *Advances in Importance Sampling*. 2021. DOI: 10.48550/ARXIV.2102.05407. URL: https://arxiv.org/abs/2102.05407.