Recursive Estimation of a Failure Probability for a Lipschitz Function

Lucie Bernard, Albert Cohen, Arnaud Guyader, Florent Malrieu

Recursive Estimation of a Failure Probability for a Lipschitz Function

Lucie Bernard
IDP, Université de Tours, France
lucie.bernard@live.fr

Albert Cohen
LJLL, Sorbonne Université, France
cohen@ann.jussieu.fr

Arnaud Guyader[1]
LPSM, Sorbonne Université & CERMICS, France
arnaud.guyader@upmc.fr

Florent Malrieu
IDP, Université de Tours, France
florent.malrieu@univ-tours.fr

Abstract

Let $g : \Omega = [0,1]^d \to \mathbb{R}$ denote a Lipschitz function that can be evaluated at each point, but at the price of a heavy computational time. Let X stand for a random variable with values in Ω such that one is able to simulate, at least approximately, according to the restriction of the law of X to any subset of Ω. For example, thanks to Markov chain Monte Carlo techniques, this is always possible when X admits a density that is known up to a normalizing constant. In this context, given a deterministic threshold T such that the failure probability $p := \mathbb{P}(g(X) > T)$ may be very low, our goal is to estimate the latter with a minimal number of calls to g. In this aim, building on Cohen *et al.* [9], we propose a recursive and optimal algorithm that selects on the fly areas of interest and estimate their respective probabilities.

Index Terms: Sequential design, Probability of failure, Sequential Monte Carlo, Tree based algorithms, High dimension.

AMS Subject Classification: 60J20, 65C05, 65C05, 68Q25, 68W20.

[1]Corresponding author.

Contents

1 Introduction

Let $g : \Omega = [0,1]^d \to \mathbb{R}$ denote a function that can be evaluated at any point $x \in \Omega$. Then, considering a random variable X with values in Ω that we can easily simulate, we want to estimate the so-called failure probability

$$p := \mathbb{P}(g(X) > T),$$

where T is a fixed threshold such that p is strictly positive but possibly very low. We are motivated by applications where each evaluation of the function g at a given $x \in \Omega$ is costly. For example, it could be the result of a numerical simulation or of a physical experiment, that has to be repeated for each new value of x. Therefore, one would like to limitate as much as possible the number of queries $x \mapsto g(x)$.

In this framework, a naive Monte Carlo method consists in simulating n independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with the same law as X, and considering the estimator

$$p_n := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{g(X_i)>T}.$$

Since the random variables $\mathbf{1}_{g(X_i)>T}$ are i.i.d. with Bernoulli law $\mathcal{B}(p)$, this estimator is unbiased, strongly consistent, and satisfies the following central limit theorem:

$$\sqrt{n}(p_n - p) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, p(1-p)).$$

However, this is an asymptotic result that is of no practical interest unless n is of order $1/p$. Indeed, if $n \ll 1/p$, as is the case in the situations we have in mind, then most of the time $p_n = 0$ and this estimator is useless.

To circumvent this issue, the purpose of variance reduction techniques is to make the rare event less rare and, in turn, decrease the previous asymptotic variance, that is $\sigma^2 = p(1-p)$. For example, instead of simulating according to the law μ of X, the idea of Importance Sampling is to consider an auxiliary distribution $\tilde{\mu}$ such that, if $\tilde{X} \sim \tilde{\mu}$, the event $\{g(\tilde{X}) > T\}$ is not rare. If this is possible, one then just has to simulate $\tilde{X}_1, \ldots, \tilde{X}_n$ i.i.d. according to $\tilde{\mu}$, and consider the estimator

$$\tilde{p}_n := \frac{1}{n} \sum_{i=1}^{n} \frac{d\mu}{d\tilde{\mu}}(\tilde{X}_i) \mathbf{1}_{g(\tilde{X}_i)>T},$$

where $\frac{d\mu}{d\tilde{\mu}}$ stands for the Radon-Nikodym derivative of μ w.r.t. $\tilde{\mu}$. This technique has been widely applied in practice and may indeed lead to dramatic variance reductions. However, it requires a lot of information about both the failure domain

$$F := \{x \in \Omega \ : \ g(x) > T\}, \tag{1.1}$$

and the law μ in order to find a relevant instrumental distribution $\tilde{\mu}$. There is a huge amount of literature on this topic. Among the first references, we can mention the paper by Kahn and Harris in particle physics [14], while the application to structural safety dates back at least to Harbitz [13]. We refer for example to the monograph [5] for details.

Another classical variance reduction technique is Importance Splitting, introduced by Kahn and Harris [14]. The principle is to consider several intermediate levels $-\infty = L_0 < L_1 < \cdots < L_K = L$ such that each conditional probability $p^{(k)} := \mathbb{P}(g(X) > L_k | g(X) > L_{k-1})$ is not small, and to apply the corresponding Bayes formula $p = p^{(1)} \ldots p^{(K)}$. Accordingly, if $\hat{p}_n^{(k)}$ is an estimator of $p^{(k)}$, then a natural estimator for p is simply

$$\hat{p}_n = \hat{p}_n^{(1)} \ldots \hat{p}_n^{(K)}.$$

In our specific context, this is the purpose of Subset Simulation [1, 2] and Adaptive Multilevel Splitting [6, 7, 8]. This is particularly suitable when X

has a density f_X that is known up to a normalizing constant, like for example in Bayesian statistics and statistical physics, for one may then apply Markov Chain Monte Carlo (MCMC) techniques to estimate each intermediate probability p_k. As explained in [12], the best asymptotic variance that one can expect through splitting techniques is $s^2 = p^2 \log(p^{-1})$, which is indeed much lower than $\sigma^2 = p(1-p)$. Nonetheless, if t stands for the number of steps of each Markov chain constructed at each step k, this necessitates about $tn \log(n) \log(p^{-1})$ calls to g, which is much larger than the number n of calls required for a naive Monte Carlo estimator. Therefore, when the simulation budget is severely limited, we can not directly apply these splitting techniques, even if we will recycle some of their ingredients in what follows.

In uncertainty quantification, a standard approach is to make more or less agressive assumptions on the failure domain F and/or the function g. One may trace back this idea to First (respectively Second) Order Reliability Methods, or FORM (respectively SORM) for short. In a nutshell, they assume that one can rewrite the probability of interest as $p = \mathbb{P}(L(Z) < 0)$, where Z stands for a standard Gaussian random vector in dimension d. Denoting $z^\star := \operatorname{argmin}\{\|z\|_2, L(z) = 0\}$ the so-called most probable point, the idea is to approximate p by the probability that Z falls in the neighborhood of $z^\star$. We refer to [10] and references therein for more details.

Alternatively, a widespread Bayesian framework consists in assuming that the function g is the realization of a Gaussian random field, defined as a prior model. Conditionally on observed values of the function, the posterior model is still Gaussian. Its mean function provides a surrogate model used to approximate g while the variance represents the uncertainty of the model (see, e.g., [16]). It is then possible to construct sequential sampling strategies to estimate the probability of failure. It basically consists in determining each new evaluation of g by minimizing a criterion that ensures that the precision of the considered estimator is improved. For instance, one may apply Stepwise Uncertainty Reduction strategies, which are formalized in [3] in this Bayesian framework. Combined with Subset Simulation, this approach can also be found in [4] for the estimation of very small probabilities. Note that this Gaussian process modelling approach corresponds to an assumption on the regularity of g, notably through the choice of the correlation function (see, e.g., [16]).

Let us also finally mention that polynomial chaos expansions represent another set of popular non-intrusive metamodelling techniques. The principle is to approximate the mapping g by a series of multivariate polynomials which are orthogonal with respect to the distributions of the input random variables $X_1, \ldots, X_d$ (see, e.g., [17] and references therein). In particular, it

allows one to compute analytically Sobol' indices, which are a standard tool in uncertainty quantification.

Here we do not adopt a Bayesian/metamodelling approach. Concerning the function g, we suppose that it is L-Lipschitz, with L known, and satisfies a so-called level set condition (see Assumption 3). As for the law of X, we assume that it admits a bounded density f_X that is known up to a normalizing constant, or that we are able to simulate at least approximately according to the restriction of f_X to any subset of Ω. In this framework, building on [9], we show that the failure probability p admits a lower (resp. upper) bound p_n^- (resp. p_n^+) based on n calls to g, and such that the approximation error satisfies, for $d \geqslant 2$,

$$E_n := p_n^+ - p_n^- \leqslant C n^{-\frac{1}{d-1}}. \tag{1.2}$$

Even if this rate of convergence is classic in deterministic numerical integration, one may notice that the quantity of interest

$$p = \int_\Omega \mathbf{1}_{g(x)>T} f_X(x) dx$$

is the integral of a non regular function, which makes the problem non trivial. In fact, we prove in Section 6 that this rate is optimal, meaning that under this set of assumptions, no algorithm based on n calls to g can achieve a better approximation error.

Nevertheless, besides n calls to g, our algorithm requires the sequential evaluation of probabilities of the form $\mathbb{P}(X \in Q)$, where Q stands for a generic dyadic subcube of Ω. It is generally impossible to do this exactly, but in many situations of interest we may apply standard MCMC techniques to estimate these probabilities with an arbitrary small (random) error. More explicitly, we propose to adopt here the same idea as in the abovementioned splitting techniques, by generating for each Q a sample of size N that is approximately i.i.d. according to the restriction of the law of X to Q.

Putting all pieces together, we propose a sequential algorithm with global stochastic error

$$|\hat{E}_n^N| \leqslant C n^{-\frac{1}{d-1}} + O_p(1/\sqrt{N}). \tag{1.3}$$

We point out that, in the latter, since the second term does not require any supplementary evaluation of g, it can easily be made arbitrarily small, so that only the first one matters and, as already explained, this first term is optimal for our set of assumptions.

The article is organized as follows. Section 2 gives in more details the assumptions and the main results of this work. Section 3 explains the deterministic

algorithm that allows us to reach the approximation error E_n in (1.2), while the proof of its optimality is deferred to Section 6. Section 4 makes more explicit the term $O_p(1/\sqrt{N})$ in (1.3) and provides asymptotic confidence intervals for our estimators. All of these results are illustrated on a toy example in Section 5, and the proof of Theorem 4.4 is detailed in Section 7.

2 Assumptions and main results

Let X be a random variable on $\Omega = [0,1]^d$ with $d \in \mathbb{N}^\star$ and $g : \Omega \to \mathbb{R}$. For a given threshold $T \in \mathbb{R}$, let us denote by F the failure domain and p the failure probability, i.e.,

$$F = \{x \in \Omega \,:\, g(x) > T\} \quad \text{and} \quad p = \mathbb{P}(X \in F) = \mathbb{P}(g(X) > T).$$

We intend to present and analyse an algorithm to estimate this failure probability as precisely as possible for a given total number n of calls to g. In all what follows, the upcoming assumptions will be of constant use.

Assumption 1 (Absolute continuity of the distribution of X)**.** *The distribution of X on Ω admits a bounded density function f_X with respect to the Lebesgue measure λ. In other words*

$$\|f_X\|_{L^\infty} = K < \infty.$$

Assumption 2 (Lipschitz smoothness)**.** *The function g is assumed to be L-Lipschitz with respect to the supremum norm on $\mathbb{R}^d$, i.e.,*

$$|g(x) - g(\tilde{x})| \leqslant L\|x - \tilde{x}\|_\infty, \quad x, \tilde{x} \in \Omega.$$

Equivalently, $\nabla g \in L^\infty(\Omega)$ with $\|\nabla g(x)\|_1 \leqslant L$ almost everywhere in Ω.

Here, we denote by $\|z\|_p$ the ℓ^p norm of a vector $z \in \mathbb{R}^d$. For the Euclidean norm, we sometime simply write $|z| := \|z\|_2$.

Assumption 3 (Level set condition)**.** *There exists a constant $M > 0$ such that*

$$\lambda(\{x \in \Omega \,:\, |g(x) - T| \leqslant \delta\}) \leqslant M\delta, \quad \delta > 0.$$

The constants L and M in Assumptions 2 and 3 are jointly coupled. Indeed, since the failure probability p is such that $0 < p < 1$, there exists x_T such that $g(x_T) = T$, and for all $x \in \Omega$, we have

$$|g(x) - T| \leqslant L\|x - x_T\|_\infty \leqslant L,$$

so that, if Assumption 3 is satisfied,

$$1 = \lambda\left(\{x \in \Omega \,:\, |g(x) - T| \leqslant L\}\right) \leqslant ML,$$

which shows that $ML \geqslant 1$. We introduce the constant

$$C := ML, \tag{2.1}$$

which will appear in the error estimates established for the algorithm presented and analyzed further.

Remark 2.1. The level set Assumption 3 may be thought as reflecting the fact that the function g is not too much flat in the vicinity of the level set $S_T = g^{-1}(\{T\})$. Indeed, when $d = 1$, if x_T is a point such that $g(x_T) = T$ and assuming that g is continuously differentiable, then $g'(x_T) < M^{-1}$ would contradict Assumption 3 for δ small enough. In the case $d \geqslant 2$, assuming that g is continuously differentiable with $\nabla g(x) \neq 0$ for any $x \in S_T$, then S_T is a compact submanifold of dimension $(d-1)$ and the coarea formula (see, e.g., [11], Proposition 3 page 118) says that, for δ small enough,

$$\lambda(\{x \in \Omega \,:\, |g(x) - T| \leqslant \delta\}) = \int_{T-\delta}^{T+\delta} \left(\int_{S_t} \frac{ds}{|\nabla g(s)|} \right) dt, \tag{2.2}$$

where ds stands for the $(d-1)$-dimensional Hausdorff measure on the level set $S_t = g^{-1}(\{t\})$. As a consequence, Assumption 3 is fulfilled with constant M for δ small enough as soon as

$$|\nabla g(x)| > \frac{2H}{M},$$

where H is the $(d-1)$-dimensional Hausdorff measure of S_T, and therefore for all δ up to raising the value of M.

The proof of the following result is housed in Section 3.3 for the first part (definition of the algorithm and error rates), and in Section 6 for the second part (optimality).

Theorem 2.2. *Under Assumptions 1, 2, and 3, there exists an algorithm that, based on n calls to g, constructs two deterministic bounds $p_n^- \leqslant p \leqslant p_n^+$ such that the approximation error $E_n := p_n^+ - p_n^-$ satisfies*

- *If $d = 1$, $E_n \leqslant 2CK\ 2^{-\frac{n}{2C}}$.*
- *If $d \geqslant 2$, $E_n \leqslant 8C^{\frac{d}{d-1}} K\ n^{-\frac{1}{d-1}}$.*

In addition, these rates of convergence are optimal.

Remark 2.3. As it will become clear in Section 3, the algorithm that we propose only requires the knowledge of the Lipschitz constant L (or an upper-bound), while that of K and M is not needed.

The quantities p_n^- and p_n^+ are defined as the measures of certain sets of dyadic cubes that are determined by our algorithm. When $f_X = 1$, that is when X is uniformly distributed, this measure can be computed exactly, otherwise it may need to be estimated. This requires possibly many samples of X, but not any additional call of g.

We begin with an idealized situation. The following result is established in Section 4.1.

Theorem 2.4. *If for each dyadic cube Q of Ω, one is able to simulate an N i.i.d. sample according to the restriction of the law of X to Q, then, without any additional call to g, we can construct two unbiased, strongly consistent and asymptotically Gaussian estimators $p_{n,N}^-$ and $p_{n,N}^+$ of the previous lower and upper bounds, i.e.,*

$$\sqrt{N}\left(p_{n,N}^{\pm} - p_n^{\pm}\right) \xrightarrow[N\to\infty]{d} \mathcal{N}(0, (\sigma_n^{\pm})^2),$$

along with consistent estimators $\sigma_{n,N}^-$ and $\sigma_{n,N}^+$ of the latter asymptotic standard deviations.

Unfortunately, it is usually not possible to simulate an N sample that is exactly i.i.d. according to the restriction of the law of X to Q. However, if the pdf f_X is known up to a normalizing constant (as is the case in many situations of interest), then one can do it at least approximately thanks to a Metropolis-Hastings algorithm. The upcoming proposition gives a flavor of the type of results we obtain in this context.

Proposition 2.5. *If f_X is continuous strictly positive on Ω, and known up to a normalizing constant, then, without any additional call to g, we can construct two estimators $\hat{p}_{n,N}^-$ and $\hat{p}_{n,N}^+$ such that, for all $t \in \mathbb{N}^\star$,*

$$\mathbb{P}\left(\hat{p}_{n,N}^{\pm} = p_{n,N}^{\pm}\right) \geqslant \left(1 - Ar^t\right)^{mN}.$$

for some constants $A > 0$, $0 < r < 1$, and $m \in \mathbb{N}^\star$. The same result holds true for $\sigma_{n,N}^-$ and $\sigma_{n,N}^+$.

The proof of this proposition is detailed in Section 4.2.

3 Approximation error

3.1 Neveu's notation

Let us denote $\mathcal{D}$ the set of all dyadic subcubes of Ω, and $\mathcal{D}_j$ the set of all dyadic cubes with sidelength 2^{-j} for $j \geqslant 1$. Given a dyadic cube Q in $\mathcal{D}$, c_Q stands for the center of Q. Each dyadic cube Q has 2^d children numbered from 1 to 2^d and each $Q \neq \Omega$ has exactly one parent.

In the sequel, we will identify a dyadic cube in $\mathcal{D}$ to a vertex in the infinite 2^d-regular tree $\mathcal{T}$. It will be referred to thanks to Neveu's notation (see [15]): the root of the tree, associated to Ω, is denoted by $\emptyset$ and, for any $k \in \mathbb{N}^\star$ and $1 \leqslant u_1, \ldots, u_k \leqslant 2^d$, the vertex $(u_1, \ldots, u_k)$ is the u_k^{th} child of $(u_1, \ldots, u_{k-1})$. A vertex $u = (u_1, \ldots, u_k)$ in $\mathcal{T}$ is then associated to a cube $Q(u)$ in $\mathcal{D}$. Notice that the sidelength of $Q(u)$ is 2^{-k} where k is the depth of u (distance between the root and u).

If v is a vertex in $\mathcal{T}$, $\bar{v}$ (resp. $\mathcal{C}(v)$) denotes the parent (resp. the set of the 2^d children) of v. The vertex v is said to be an ancestor of u, and we denote $v \leqslant u$, if $Q(u) \subset Q(v)$ or, equivalently, if v is a prefix of u. Notice that $u \leqslant u$. In the sequel, $a(v)$ stands for the set made of the ancestors of v, including v but excluding the root for convenience. Finally, if v and w are two vertices, then $v \wedge w$ stands for the more recent common ancestor of v and w.

We say that Λ is a finite 2^d-regular tree, if it is a finite subset of $\mathcal{T}$ such that $u \in \Lambda$ and $v \leqslant u$ implies that $v \in \Lambda$. For any finite 2^d-regular tree Λ of $\mathcal{T}$, the leaves (resp. internal vertices) of Λ are the vertices in Λ with no child (resp. with 2^d children) in Λ. The depth of Λ is defined as the maximal depth of the vertices in Λ. See Figure 1 for an illustration.

For the purpose of our algorithm, the dyadic cubes (or vertices) are labelled according to the following rule that involves the evaluation of g at their centers.

Definition 3.1 (Label of a cube)**.** *The dyadic cube Q with side length 2^{-j} and center c_Q is labelled*

- $\mathcal{I}$ *(inside) if* $g(c_Q) > T + L2^{-j-1}$,
- $\mathcal{O}$ *(outside) if* $g(c_Q) < T - L2^{-j-1}$,
- $\mathcal{U}$ *(uncertain) otherwise.*

A cube with label $\mathcal{I}$ is included in the failure set F. Indeed, for any $Q \in \mathcal{D}_j$ and any $x \in Q$,

$$|g(x) - g(c_Q)| \leqslant L\|x - c_Q\|_\infty \leqslant L2^{-j-1}.$$

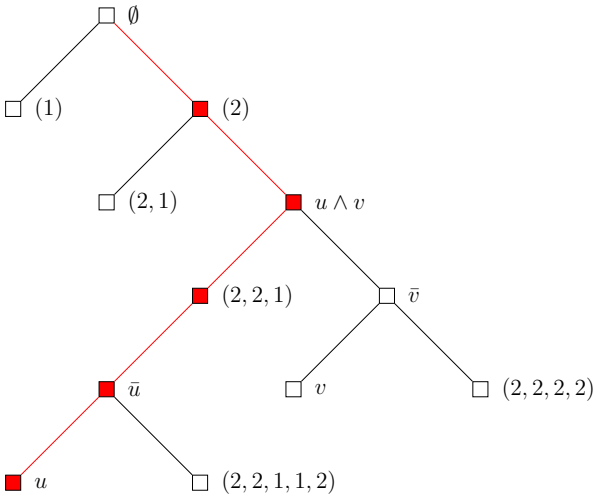

Figure 1: Example of a 2-regular finite tree of depth 5 in dimension 1. The red line represents the set $a(u)$ of the ancestors of the leaf u.

As a consequence, if the label of Q is $\mathcal{I}$, then, for any $x \in Q$, we have

$$g(x) \geqslant g(c_Q) - L2^{-j-1} > T.$$

Likewise, a cube with label $\mathcal{O}$ is included in $F^c := \Omega \setminus F$. Finally, a cube with label $\mathcal{U}$ may intersect F and/or F^c.

3.2 Recursive construction of relevant trees

The algorithm starts with $\Lambda(0) = \{\Omega\}$, where Ω has the label $\mathcal{U}$ and the depth of $\Lambda(0)$ is 0. At a given step $k > 0$, a finite 2^d-regular tree $\Lambda(k)$ of depth k has been constructed with the following features:

(i) internal vertices are all labelled as $\mathcal{U}$,

(ii) leaves of depth lower than k are labelled $\mathcal{I}$ or $\mathcal{O}$,

(iii) leaves of depth k can have any label.

Then, the tree $\Lambda(k+1)$ is obtained by performing a 2^d-split on each leaf with label $\mathcal{U}$ and evaluating g at their center in order to label the new leaves according to Definition 3.1. Clearly the new tree $\Lambda(k+1)$ of depth $(k+1)$ has similar properties (see Figure 2).

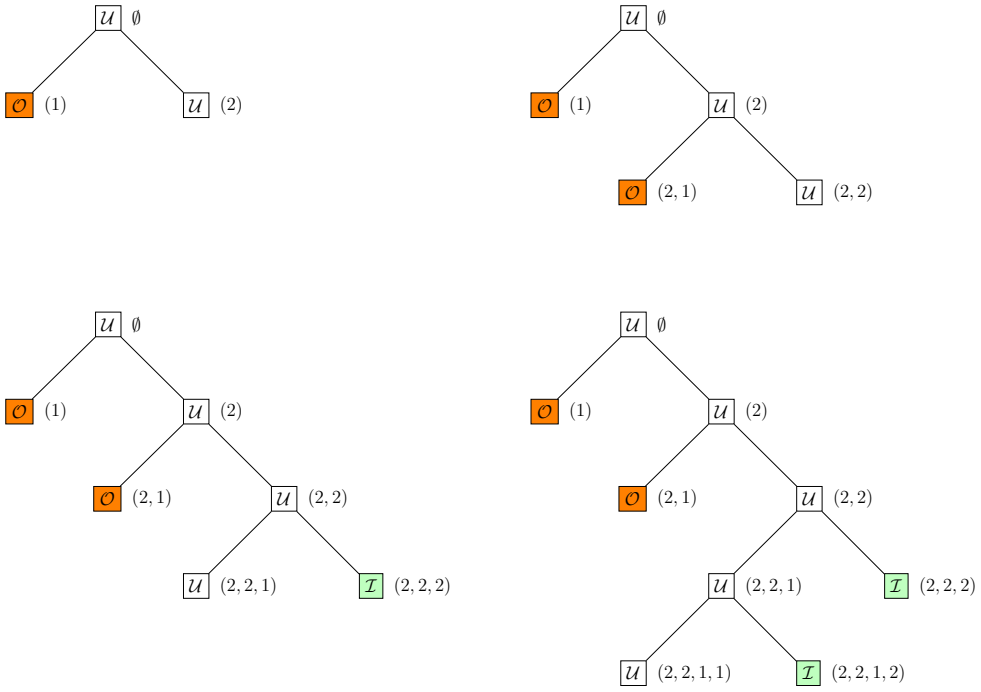

Figure 2: Example of a recursive construction of $\Lambda(1), \ldots, \Lambda(4)$ for $d = 1$.

Denoting $|\Lambda(k)|$ the cardinal of $\Lambda(k)$, the number n_k of evaluations of g that is involved in the construction of $\Lambda(k)$ is therefore given by $n_0 = 0$ and, for all $k \geqslant 1$,

$$n_k = |\Lambda(k)| - 1,$$

since the evaluation at the center of Ω is useless when $p > 0$. The following result gives an upper bound on this number. Recall that C is the constant defined by (2.1).

Proposition 3.1. *Let $k \geqslant 0$. If $d = 1$, the number n_k of evaluations of g satisfies*

$$n_k \leqslant 2Ck.$$

If $d \geqslant 2$, then we have

$$n_k \leqslant 4C\ 2^{(d-1)k}.$$

Proof. Since $n_0 = 0$, the result is clear for $k = 0$. Therefore, let us consider the case where $k \geqslant 1$. For any $0 \leqslant j \leqslant k-1$, let us denote by $\mathcal{U}(j)$ the set of leaves of $\Lambda(j)$ with label $\mathcal{U}$ (see Figure 2). Recall from Definition 3.1 that this set is made of dyadic cubes with side length 2^{-j} such that, for any $Q \in \mathcal{U}(j)$,

$$|g(c_Q) - T| \leqslant \frac{L}{2^{j+1}}.$$

As a consequence, for any $x \in Q \in \mathcal{U}(j)$, Assumption 2 gives

$$|g(x) - T| \leqslant |g(x) - g(c_Q)| + |g(c_Q) - T| \leqslant \frac{L}{2^j}.$$

This ensures that

$$\bigcup_{Q \in \mathcal{U}(j)} Q \subset \left\{ x \in \Omega \ : \ |g(x) - T| \leqslant \frac{L}{2^j} \right\}.$$

Since the volume of each cube in $\mathcal{U}(j)$ is 2^{-jd}, this yields

$$\frac{|\mathcal{U}(j)|}{2^{jd}} \leqslant \lambda\left(\left\{ x \in \Omega \ : \ |g(x) - T| \leqslant \frac{L}{2^j} \right\} \right),$$

with the understanding that $|\mathcal{U}(j)|$ is the cardinal of $\mathcal{U}(j)$. Thanks to Assumption 3, we get that

$$|\mathcal{U}(j)| \leqslant \left(\frac{C}{L} \frac{L}{2^j} \right) 2^{jd} = C 2^{j(d-1)} =: \mu_j. \tag{3.1}$$

Thus, the construction of $\Lambda(j+1)$ requires at most $2^d \mu_j$ evaluations of g. As a consequence, we can bound the total number n_k of calls to g to construct $\Lambda(k)$ as follows:

$$n_k \leqslant \sum_{j=0}^{k-1} 2^d \mu_j.$$

If $d \geqslant 2$, we are led to

$$n_k \leqslant C 2^d \frac{2^{(d-1)k} - 1}{2^{d-1} - 1} \leqslant 4C \ 2^{(d-1)k}.$$

In the case $d = 1$, we obtain $n_k \leqslant 2Ck$. □

3.3 Control of the error

For any $k \geqslant 0$, we denote by $\mathcal{I}(k)$ (resp. $\mathcal{U}(k)$) the leaves of $\Lambda(k)$ with label $\mathcal{I}$ (resp. $\mathcal{U}$). We can readily estimate the failure probability p thanks to the tree $\Lambda(k)$ as follows:

$$p^-(k) \leqslant p \leqslant p^+(k),$$

where

$$p^-(k) := \sum_{Q \in \mathcal{I}(k)} \mathbb{P}(X \in Q) \quad \text{and} \quad p^+(k) := p^-(k) + \sum_{Q \in \mathcal{U}(k)} \mathbb{P}(X \in Q). \tag{3.2}$$

Lemma 3.2 (Control of the error)**.** *For any $k \geqslant 0$, the estimations $p^-(k)$ and $p^+(k)$ of the failure probability p given by the tree $\Lambda(k)$ are such that*

$$0 \leqslant p^+(k) - p^-(k) \leqslant CK2^{-k}.$$

Proof. Under Assumption 1, for any $k \in \mathbb{N}$ and $Q \in \mathcal{U}(k)$, we have

$$\lambda(Q) = 2^{-dk} \quad \text{and} \quad \mathbb{P}(X \in Q) \leqslant 2^{-dk}K.$$

As a consequence, the definition of $p^-(k)$ and $p^+(k)$ together with Equation (3.1) ensure that

$$p^+(k) - p^-(k) = \sum_{Q \in \mathcal{U}(k)} \mathbb{P}(X \in Q) \leqslant \frac{|\mathcal{U}(k)|K}{2^{dk}} \leqslant CK\frac{2^{k(d-1)}}{2^{dk}}.$$

This concludes the proof. □

Before going further, let us notice that, for any $n \geqslant 1$, there exists $k \geqslant 0$ such that $n_k \leqslant n < n_{k+1}$, with the convention $n_0 = 0$. We can apply the same algorithm as before, with the understanding that all the leaves of the tree $\Lambda(k)$ are explored while this is the case only for $(n - n_k)$ leaves with depth $(k + 1)$ of the tree $\Lambda(k + 1)$. This defines a subtree Λ_n of $\Lambda(k + 1)$. With obvious notation, the leaves of Λ_n can be partitioned as $\mathcal{I}_n \cup \mathcal{U}_n \cup \mathcal{O}_n$. In this respect, we deduce upper and lower bounds p_n^- and p_n^+ for p as follows:

$$p_n^- := \sum_{Q \in \mathcal{I}_n} \mathbb{P}(X \in Q) \quad \text{and} \quad p_n^+ := p_n^- + \sum_{Q \in \mathcal{U}_n} \mathbb{P}(X \in Q). \tag{3.3}$$

Clearly, we have

$$p^-(k) \leqslant p_n^- \leqslant p \leqslant p_n^+ \leqslant p^+(k),$$

so that the approximation error $E_n := p_n^+ - p_n^-$ satisfies $E_n \leqslant p^+(k) - p^-(k)$. With this in mind, we can now complete the proof of Theorem 2.2. When $d \geqslant 2$, according to Proposition 3.1, we may write

$$n < n_{k+1} \leqslant 4C\ 2^{(d-1)(k+1)} \quad \text{or, equivalently,} \quad 2^{-k} \leqslant 2\Big(\frac{n}{4C}\Big)^{-\frac{1}{d-1}}.$$

Hence, Lemma 3.2 yields

$$E_n \leqslant p^+(k) - p^-(k) \leqslant CK2^{-k} \leqslant 8C^{\frac{d}{d-1}}Kn^{-\frac{1}{d-1}}.$$

When $d = 1$, the same reasoning gives

$$n < n_{k+1} \leqslant 2C(k + 1) \quad \text{or, equivalently,} \quad 2^{-k} \leqslant 2^{1-\frac{n}{2C}},$$

so that

$$E_n \leqslant p^+(k) - p^-(k) \leqslant 2CK\ 2^{-\frac{n}{2C}}.$$

This terminates the proof of the first part of Theorem 2.2. The fact that this error is optimal is shown in Section 6.

4 Estimation error

We return to the notation of Section 3.3 and recall that, for any $n \geqslant 2$, we denote by $\mathcal{I}_n$ (resp. $\mathcal{U}_n$) the leaves of Λ_n with label $\mathcal{I}$ (resp. $\mathcal{U}$), so that

$$p_n^- \leqslant p \leqslant p_n^+,$$

where

$$p_n^- := \sum_{Q \in \mathcal{I}_n} \mathbb{P}(X \in Q) \quad \text{and} \quad p_n^+ := p_n^- + \sum_{Q \in \mathcal{U}_n} \mathbb{P}(X \in Q). \tag{4.1}$$

Our goal in this section is to estimate p_n^- and p_n^+ with no additional call to g. We first do it by assuming that, for each vertex $u \in \Lambda_n$, we can simulate an N i.i.d. sample distributed according to the law of X given that it belongs to $Q(u)$. This allows us to propose in Section 4.1 two idealized estimators $p_{n,N}^-$ and $p_{n,N}^+$ along with their asymptotic variances. In Section 4.2, thanks to MCMC techniques, we construct two estimators $\hat{p}_{n,N}^-$ and $\hat{p}_{n,N}^+$ of the latters provided that the density f_X is known up to a normalizing constant.

4.1 Estimation error in an idealized case

From a given tree Λ_n, one can estimate the failure probability p thanks to p_n^- and p_n^+ defined in Equation (4.1). To that end, one has to compute (or estimate) the probability

$$p(u) := \mathbb{P}(X \in Q(u)),$$

for each leaf u of Λ_k. If u is far from the root, then $p(u)$ should be very small and difficult to estimate directly through a naive Monte Carlo method, as explained in Section 1. Therefore, we propose to apply a splitting strategy inspired by rare event estimation.

For a given leaf $u \in \Lambda_n$, recall that $a(u)$ stands for the set of the ancestors of u, including u but excluding the root for convenience. Since $\mathbb{P}(X \in \Omega) = 1$, Bayes formula ensures that

$$\mathbb{P}(X \in Q(u)) = \prod_{v \in a(u)} \mathbb{P}(X \in Q(v) | X \in Q(\bar{v})),$$

which can be reformulated as follows

$$p(u) = \prod_{v \in a(u)} q(v) \quad \text{where} \quad q(v) := \mathbb{P}(X \in Q(v) | X \in Q(\bar{v})).$$

Assumption 4 (Perfect samplings)**.** *Recall that $\mathcal{T}$ stands for the infinite 2^d-regular tree. For any $v \in \mathcal{T}$, consider a sequence $(X_i^v)_{i\geqslant 1}$ of i.i.d. random variables with distribution $\mathcal{L}(X|X \in Q(v))$ and assume that the sequences $(X^v)_{v\in\mathcal{T}}$ are independent.*

Definition 4.1 (Ideal estimators)**.** *For $N \geqslant 1$ and $u, v \in \mathcal{T}$, we define*

$$C_N^v := \sum_{i=1}^{N} \mathbf{1}_{\left\{X_i^{\bar{v}} \in Q(v)\right\}}, \quad q_N(v) := \frac{C_N^v}{N} \quad \textit{and} \quad p_N(u) := \prod_{v\in a(u)} q_N(v). \quad (4.2)$$

Remark 4.2 (Multinomial distribution and unbiasedness)**.** The random variable C_N^v is the number of random variables $(X_i^{\bar{v}})_{1\leqslant i\leqslant N}$ which are in fact in the cube $Q(v)$. Let w be a fixed vertex. The distribution of the random vector $(C_N^v)_{v\in c(w)}$ is the multinomial distribution with parameters N and $(q(v))_{v\in c(w)}$. As a consequence, $q_N(v)$ is a strongly consistent and unbiased estimator of $q(v)$. In addition, for k different vertices $w_1, w_2, \ldots, w_k$ in $\mathcal{T}$, the vectors

$$(C_N^v)_{v\in c(w_1)}, (C_N^v)_{v\in c(w_2)}, \ldots, (C_N^v)_{v\in c(w_k)}$$

are independent. From this we deduce that $p_N(u)$ is also unbiased.

Remark 4.3. The random variables $(q_N(v))_{v\in a(u)}$ are independent. Nevertheless, for two different leaves u and u', $p_N(u)$ and $p_N(u')$ are not independent.

Our next result, whose proof is deferred to Section 7, provides asymptotic properties (namely, consistency and asymptotic normality) for the estimator $p_N(\mathcal{S})$ of the probability $p(\mathcal{S})$ associated to any set of leaves $\mathcal{S}$. One may keep in mind that, for our problem, we will apply this result with $\mathcal{S} = \mathcal{I}_n$ and $\mathcal{S} = \mathcal{I}_n \cup \mathcal{U}_n$, in which case $p(\mathcal{S})$ (respectively $p_N(\mathcal{S})$) corresponds to p_n^- and p_n^+ (respectively $p_{n,N}^-$ and $p_{n,N}^+$).

Theorem 4.4. *For any set $\mathcal{S}$ of leaves of a tree Λ, one can estimate*

$$p(\mathcal{S}) := \sum_{u\in\mathcal{S}} p(u) \quad \textit{by} \quad p_N(\mathcal{S}) := \sum_{u\in\mathcal{S}} p_N(u),$$

where $p_N(u)$ is defined in (4.2). The estimator $p_N(\mathcal{S})$ is unbiased and strongly consistent:

$$p_N(\mathcal{S}) \xrightarrow[N\to\infty]{a.s.} p(\mathcal{S}).$$

Moreover, it is asymptotically normal, namely

$$\sqrt{N}(p_N(\mathcal{S}) - p(\mathcal{S})) \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

where

$$\sigma^2 = \sum_{u\in\mathcal{S}} p(u)^2 \sum_{v\in a(u)} \frac{1-q(v)}{q(v)} + \sum_{\substack{u,u'\in\mathcal{S}\\ u\neq u'}} p(u)p(u') \left[\sum_{v\in a(u\wedge u')} \frac{1-q(v)}{q(v)} - 1 \right].$$

Remark that if $q(v) = 0$ then $p(u) = 0$ whenever $v \leqslant u$, so that one can cancel u from the set of leaves $\mathcal{S}$ and the expression of σ^2 is always well-defined.

Remark 4.5 (Variance estimation)**.** Recall that each $q(v)$ is strictly positive and consistently estimated on the fly by $q_N(v)$, so that σ^2 is readily estimated by

$$\sigma_N^2 = \sum_{u\in\mathcal{S}} p_N(u)^2 \sum_{v\in a(u)} \frac{1-q_N(v)}{q_N(v)} + \sum_{\substack{u,u'\in\mathcal{S}\\ u\neq u'}} p_N(u)p_N(u') \left[\sum_{v\in a(u\wedge u')} \frac{1-q_N(v)}{q_N(v)} - 1 \right]$$

and σ_N^2 goes almost surely to σ^2 when N goes to infinity. Hence, Slutsky's lemma ensures that

$$\sqrt{N}\, \frac{p_N(\mathcal{S}) - p(\mathcal{S})}{\sigma_N} \xrightarrow[N\to\infty]{d} \mathcal{N}(0,1).$$

In particular, the latter provides asymptotic confidence intervals for $p(\mathcal{S})$.

For our concern, recall that $p_n^- \leqslant p \leqslant p_n^+$ where

$$p_n^- = p(\mathcal{I}_n) = \sum_{u\in\mathcal{I}_n} p(u) \quad \text{and} \quad p_n^+ = p(\mathcal{I}_n \cup \mathcal{U}_n).$$

Hence, the sets of leaves of interest are $\mathcal{S} = \mathcal{I}_n$ and $\mathcal{S} = \mathcal{I}_n \cup \mathcal{U}_n$. Indeed, the previous results establish that

$$p_{n,N}^- = p_N(\mathcal{I}_n) \xrightarrow[N\to\infty]{a.s.} p_n^- \leqslant p \leqslant p_n^+ \xleftarrow[N\to\infty]{a.s.} p_N(\mathcal{I}_n \cup \mathcal{U}_n) = p_{n,N}^+,$$

as well as

$$\sqrt{N}\left(p_{n,N}^\pm - p_n^\pm\right) \xrightarrow[N\to\infty]{d} \mathcal{N}(0, (\sigma_n^\pm)^2).$$

In addition, by Remark 4.5, we can construct on the fly consistent estimators $\sigma_{n,N}^-$ and $\sigma_{n,N}^+$ of the latter asymptotic standard deviations. This closes the proof of Theorem 2.4.

Remark 4.6 (Asymptotic confidence intervals)**.** Denote by Φ the cumulative distribution function of the standard normal distribution so that, for $\alpha \in (0,1)$, $\Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile. If we define

$$m_{n,N} := p^-_{n,N} - \frac{\Phi^{-1}(1-\alpha/2)\sigma^-_{n,N}}{\sqrt{N}}$$

as well as

$$M_{n,N} := p^+_{n,N} + \frac{\Phi^{-1}(1-\alpha/2)\sigma^+_{n,N}}{\sqrt{N}},$$

then $[m_{n,N}, 1]$ and $[0, M_{n,N}]$ are $100(1-\alpha/2)\%$ asymptotic confidence intervals for, respectively, p^-_n and p^+_n. Since $p^-_n \leqslant p \leqslant p^+_n$, the union bound ensures that $[m_{n,N}, M_{n,N}]$ is a $100(1-\alpha)\%$ asymptotic confidence interval for p.

4.2 Estimation error in practice

The purpose of this section is to prove Proposition 2.5. Recall from Definition 4.1 that each leaf probability

$$p(u) = \mathbb{P}(X \in Q(u)) = \prod_{v \in a(u)} \mathbb{P}(X \in Q(v) | X \in Q(\bar{v})) = \prod_{v \in a(u)} q(v)$$

is estimated by

$$p_N(u) = \prod_{v \in a(u)} q_N(v) \quad \text{where} \quad q_N(v) = \frac{C^v_N}{N} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{X_i^{\bar{v}} \in Q(v)\}}.$$

To apply the results of Theorem 4.4, this supposes that, for each $\bar{v}$, we have a sample of N i.i.d. random variables $X_i^{\bar{v}}$. In addition, for two vertices v and v' such that $\bar{v} \neq \bar{v}'$, these samples must be independent. The present section explains how to reach this goal, at least approximately.

Consider a fixed vertex $\bar{v}$, denote $\mu_{\bar{v}} = \mathcal{L}(X | X \in Q(\bar{v}))$, and $f_{\bar{v}}$ the corresponding probability density function, that is

$$\mu_{\bar{v}}(dx) = f_{\bar{v}}(x)dx = \frac{1}{\mathbb{P}(X \in Q(\bar{v}))} f_X(x) \mathbf{1}_{x \in Q(\bar{v})}.$$

Starting from a point $X_0 \sim \mathcal{U}_{\bar{v}}$ the uniform law on $Q(\bar{v})$, the Metropolis-Hastings algorithm allows us to construct a Markov chain (X_n) with asymptotic distribution $\mu_{\bar{v}}$.

We refer the interested reader to Tierney [18] for a thorough presentation as well as numerous theoretical results on Markov chain Monte Carlo methods. For our purpose, we just present the idea for a specific choice of the Markov dynamics, which turns out to be a particular case of independent Metropolis.

Here is the mechanism: starting from X_t, simulate $X'_t \sim \mathcal{U}_{\bar{v}}$ and set

$$X_{t+1} := X'_t \mathbf{1}_{U_{t+1} \leqslant f_X(X'_t)/f_X(X_t)} + X_t \mathbf{1}_{U_{t+1} > f_X(X'_t)/f_X(X_t)}, \tag{4.3}$$

where $(U_t)_{t \in \mathbb{N}^\star}$ is a sequence of i.i.d. random variables with uniform law on $[0,1]$. Needless to say, in the previous expression, X_t, X'_t, and U_{t+1} are also assumed independent. It is readily seen that, if we denote by $K_{\bar{v}}$ the transition kernel associated to this Markov chain, then $K_{\bar{v}}$ is $\mu_{\bar{v}}$-reversible so that, under appropriate assumptions, (X_t) goes in distribution to $\mu_{\bar{v}}$.

In order to make this convergence more precise, let us recall that the total variation distance between two probability measures μ and ν on $Q(\bar{v})$ is

$$\|\mu - \nu\|_{TV} := \sup_{B \in \mathcal{B}_{\bar{v}}} |\mu(B) - \nu(B)|,$$

where $\mathcal{B}_{\bar{v}}$ is the collection of all Borel sets on $Q(\bar{v})$. Denoting δ_x the Dirac measure at x and $\delta_x K_{\bar{v}}^t$ the law of X_t for the above Markov chain with initial condition $X_0 = x$, we say that the chain is uniformly ergodic on $Q(\bar{v})$ if there exist $A_{\bar{v}} > 0$ and $0 < r_{\bar{v}} < 1$ such that, for all $t \in \mathbb{N}^\star$,

$$\sup_{x \in Q(\bar{v})} \|\delta_x K_{\bar{v}}^t - \mu_{\bar{v}}\|_{TV} \leqslant A_{\bar{v}} r_{\bar{v}}^t.$$

Let $g_{\bar{v}}$ stand for the density of the uniform distribution on $Q(\bar{v})$ and

$$\beta_{\bar{v}}^{-1} := \sup_{x \in Q(\bar{v})} \frac{f_{\bar{v}}(x)}{g_{\bar{v}}(x)}, \tag{4.4}$$

then Corollary 4 in [18] ensures that the Markov chain (X_t) is uniformly ergodic with convergence rate $r_{\bar{v}} \leqslant 1 - \beta_{\bar{v}}$. In our context, notice that the latter is always strictly less than 1 if, for example, f_X is continuous and strictly positive on $\Omega = [0,1]^d$, hence our assumption in Proposition 2.5.

To see the consequence of this result in our context, remember the coupling interpretation of the total variation distance, that is

$$\|\mu - \nu\|_{TV} = \inf_{(X,Y)} \mathbb{P}(X \neq Y),$$

where the infimum is over all couples of random variables on $Q(\bar{v}) \times Q(\bar{v})$ with marginal laws μ and ν. More precisely, given X with law ν, it is always

possible to construct a random variable Y with law μ such that the equality is achieved, i.e., $\mathbb{P}(X \neq Y) = \|\mu - \nu\|_{TV}$.

Hence, if we consider as above a Markov chain $(X_t^{\bar{v}})$ with arbitrary initial condition, for example $X_0^{\bar{v}}$ with uniform law $\mathcal{U}_{\bar{v}}$ on $Q(\bar{v})$, there exists a random variable $X_\infty^{\bar{v}}$ with law $\mu_{\bar{v}}$ such that

$$\mathbb{P}(X_t^{\bar{v}} \neq X_\infty^{\bar{v}}) = \|\mathcal{U}_{\bar{v}} K_{\bar{v}}^t - \mu_{\bar{v}}\|_{TV} \leqslant A_{\bar{v}} r_{\bar{v}}^t.$$

Therefore, if we start from N i.i.d. initial conditions $\mathbf{X}_0^{\bar{v}} := (X_0^{\bar{v},(1)}, \dots, X_0^{\bar{v},(N)})$ with uniform distribution on $Q(\bar{v})$, and run independently during t steps the previous Metropolis algorithm to obtain the sample $\mathbf{X}_t^{\bar{v}} := (X_t^{\bar{v},(1)}, \dots, X_t^{\bar{v},(N)})$, we deduce that

$$\mathbb{P}(\mathbf{X}_t^{\bar{v}} = \mathbf{X}_\infty^{\bar{v}}) \geqslant (1 - A_{\bar{v}} r_{\bar{v}}^t)^N,$$

where $\mathbf{X}_\infty^{\bar{v}} := (X_\infty^{\bar{v},(1)}, \dots, X_\infty^{\bar{v},(N)}) \sim \mu_{\bar{v}}^{\otimes n}$.

Next, apply the previous procedure to each vertex $\bar{v}$ of the considered tree Λ, denote by $\mathcal{X}_t := (\mathbf{X}_t^{\bar{v}})_{\bar{v}\in\Lambda}$ all the corresponding sets of N i.i.d. samples, and $\mathcal{X}_\infty := (\mathbf{X}_\infty^{\bar{v}})_{\bar{v}\in\Lambda}$ the corresponding sets of N i.i.d. "idealized" samples. Denoting $A_\Lambda := \max_{\bar{v}\in\Lambda} A_{\bar{v}}$ and $r_\Lambda := \max_{\bar{v}\in\Lambda} r_{\bar{v}} \in (0,1)$, we deduce that

$$\mathbb{P}(\mathcal{X}_t = \mathcal{X}_\infty) \geqslant (1 - A_\Lambda r_\Lambda^t)^{|\Lambda|N}.$$

For each vertex v and each leaf u, consider the estimators

$$\hat{p}_N(u) := \prod_{v\in a(u)} \hat{q}_N(v) \quad \text{where} \quad \hat{q}_N(v) := \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\left\{X_t^{\bar{v},(i)} \in Q(v)\right\}},$$

and, for any set $\mathcal{S}$ of leaves of the tree Λ,

$$\hat{p}_N(\mathcal{S}) := \sum_{u\in\mathcal{S}} \hat{p}_N(u).$$

Clearly, on the event $\{\mathcal{X}_t = \mathcal{X}_\infty\}$, we have $\hat{p}_N(\mathcal{S}) = p_N(\mathcal{S})$, which means that

$$\mathbb{P}(\hat{p}_N(\mathcal{S}) = p_N(\mathcal{S})) \geqslant (1 - A_\Lambda r_\Lambda^t)^{|\Lambda|N},$$

where $p_N(\mathcal{S})$ is the ideal estimator defined in Theorem 4.4. Finally, it suffices to consider $\mathcal{S} = \mathcal{I}_n$ and $\mathcal{S} = \mathcal{I}_n \cup \mathcal{U}_n$ to conclude the proof of Proposition 2.5.

Remark 4.7 (Confidence intervals in practice)**.** Mutatis mutandis, the result of Remark 4.6 is still valid. Specifically, if we denote

$$\hat{m}_{n,N} := \hat{p}_{n,N}^- - \frac{\Phi^{-1}(1-\alpha/2)\hat{\sigma}_{n,N}^-}{\sqrt{N}}$$

as well as

$$\hat{M}_{n,N} := \hat{p}^{+}_{n,N} + \frac{\Phi^{-1}(1-\alpha/2)\hat{\sigma}^{+}_{n,N}}{\sqrt{N}},$$

then on the event $\{\mathcal{X}_t = \mathcal{X}_\infty\}$, $[\hat{m}_{n,N}, \hat{M}_{n,N}]$ is a $100(1-\alpha)\%$ asymptotic confidence interval for p. This will be illustrated in Section 5.

Remark 4.8. Returning to (4.4), one can notice that the smaller the side length of $Q(\bar{v})$, the faster the convergence of the Metropolis algorithm. Indeed, denoting $c_{Q(\bar{v})}$ its center and $\lambda(Q(\bar{v}))$ its Lebesgue measure, the continuity of f_X ensures that, when $\lambda(Q(\bar{v})) \to 0$,

$$f_{\bar{v}}(x) = \frac{f_X(x)\mathbf{1}_{x\in Q(\bar{v})}}{\mathbb{P}(X \in Q(\bar{v}))} \approx \frac{f_X(c_{Q(\bar{v})})\mathbf{1}_{x\in Q(\bar{v})}}{f_X(c_{Q(\bar{v})})\lambda(Q(\bar{v}))} = g_{\bar{v}}(x),$$

which means that $\beta_{\bar{v}}$ goes to 1 or, equivalently, that $r_{\bar{v}}$ goes to 0.

5 Numerical illustration

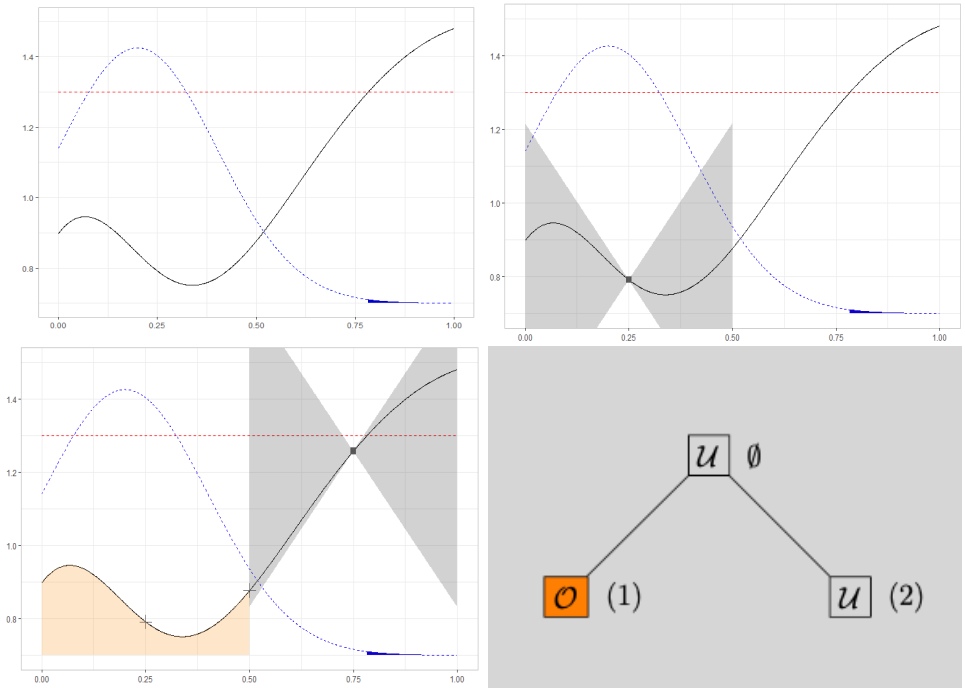

Figure 3: Representation of the function g (black), the pdf f_X (blue), the threshold T (red), the probability p (blue region), and illustration of the first step of the algorithm.

To illustrate our algorithm, we consider a toy example which is just a variant of the one proposed in Section 5.1 of [3]. For all $x \in [0,1]$, we set

$$g(x) = (0.8x - 0.3) + \exp\left(-11.534x^{1.95}\right) + \exp\left(-2(x-0.9)^2\right),$$

which is L-Lipschitz with $L = \sup_{x\in[0,1]} |g'(x)| \approx 1.61$. The law of X is the restriction of a Gaussian distribution $\mathcal{N}(1/5, 1/25)$ to $[0,1]$, i.e.,

$$f_X(x) \propto \exp\left\{-\frac{25}{2}\left(x - \frac{1}{5}\right)^2\right\} \mathbf{1}_{[0,1]}(x).$$

Finally, we take $T = 1.3$, so that a standard numerical integration gives $p \approx 2.08 \times 10^{-3}$. This is illustrated on Figure 3, together with the first step of the algorithm. Recall that the evaluation of g at point $x = 1/2$ is useless. Indeed, since $0 < p < 1$, the interval $\Omega = [0,1]$ is necessarily classified as uncertain (i.e., $\mathcal{U}$). Therefore, the first step consists in computing $g(1/4)$ and $g(3/4)$, which correspond respectively to vertices (1) and (2) of the tree. From this figure, it is easy to see that (1) is classified as out (i.e., $\mathcal{O}$) while (2) is classified as uncertain (i.e., $\mathcal{U}$). Therefore, there is no need to further investigate the interval $[0, 1/2]$.

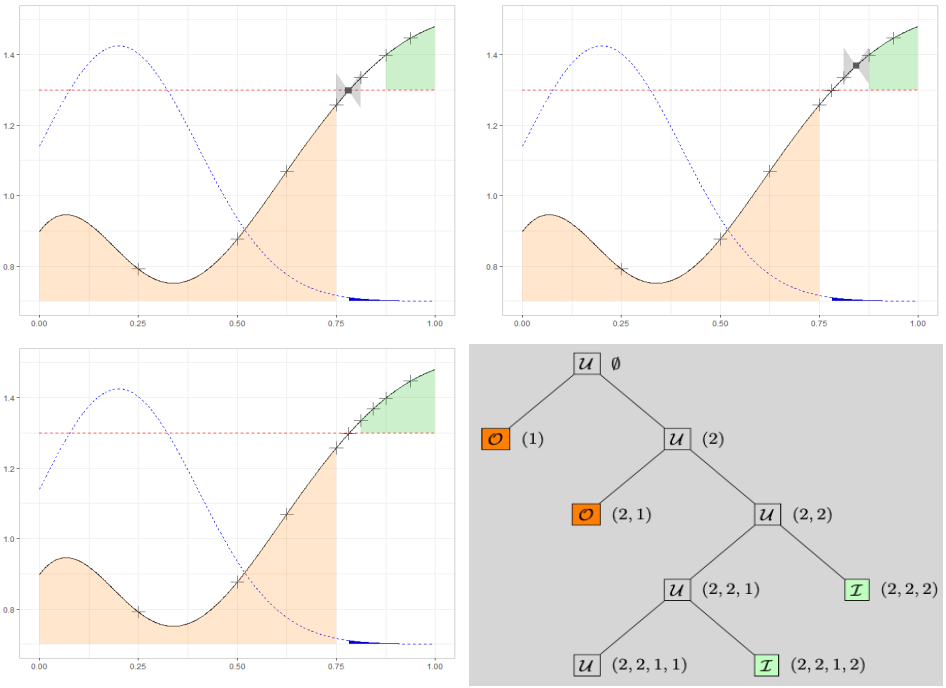

Figure 4: Step 4 of the algorithm.

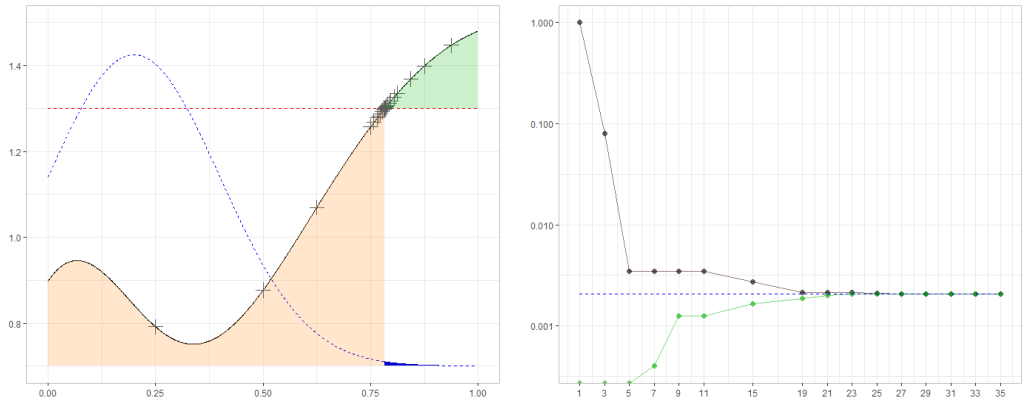

Figure 5: Convergence of p_n^- and p_n^+.

Figure 4 represents step 4 of the algorithm, which consists in evaluating g at points $x = 25/32$ (i.e., vertex $(2, 2, 1, 1)$) and $x = 27/32$ (i.e., vertex $(2, 2, 1, 2)$). These evaluations lead to classify the interval $[24/32, 26/32]$ as uncertain (i.e., $\mathcal{U}$) and the interval $[26/32, 28/32]$ as included in the failure domain (i.e., $\mathcal{I}$). At this point, the deterministic lower and upper bounds for p are thus

$$p^-(4) = \mathbb{P}(X \in [13/16, 1]) \approx 1.3 \times 10^{-3} \leqslant p,$$

and

$$p \leqslant p^+(4) = \mathbb{P}(X \in [12/16, 1]) \approx 2.5 \times 10^{-3},$$

and the approximation error is simply

$$p^+(4) - p^-(4) = \mathbb{P}(X \in [12/16, 13/16]) \approx 2.2 \times 10^{-3}.$$

Unsurprisingly, one may notice that the upper bound given by Lemma 3.2 is very pessimistic. Indeed, since $d = 1$ we know that $C \geqslant 1$ (see Section 2) and this upper bound can be minorized as follows:

$$CK2^{-k} \geqslant \frac{2^{-4}}{\displaystyle\int_0^1 \exp\left\{-\frac{25}{2}\left(x - \frac{1}{5}\right)^2\right\} dx} \approx 0.148 \gg 2.2 \times 10^{-3}.$$

On this toy example, since the law of X is simply the restriction of a Gaussian distribution, it is easy to have a very precise numerical approximation of $\mathbb{P}(X \in Q)$ for any dyadic interval Q and, in turn, for the lower and upper bounds at each step of the algorithm. In other words, we can easily compute

the (deterministic) approximation error. The evolution of these bounds p_n^- and p_n^+ as the number of evaluation points grows is given in Figure 5 for a total budget of $n = 35$ calls to g.

However, in practice, this is usually not possible, hence the use of MCMC techniques as explained in Section 4.2. On our example, up to a normalizing constant, the pdf f_X is defined by

$$f_X(x) \propto \exp\left\{-\frac{25}{2}\left(x - \frac{1}{5}\right)^2\right\} \mathbf{1}_{[0,1]}(x).$$

Thus, for any couple of points (x, x'), the Metropolis ratio $f_X(x')/f_X(x)$ that appears in (4.3) is very easy to compute. We have applied this idea for a sample size $N = 10^5$ with $t = 25$ Markov transitions for each probability estimation. In this respect, Figure 6 shows that when N is much larger than the approximation error, then the latter is much larger than the estimation error. In order to illustrate Remark 4.6, the asymptotic confidence intervals are also given.

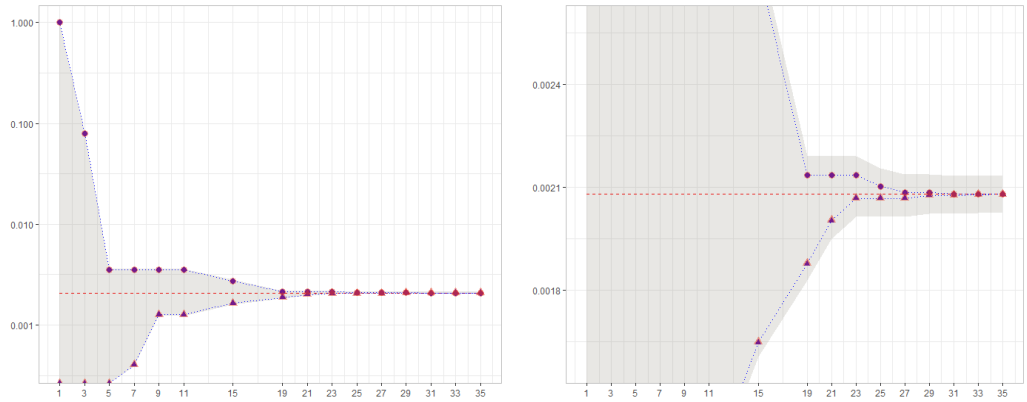

Figure 6: Estimators $\hat{p}_{n,N}^-$ and $\hat{p}_{n,N}^+$ of p_n^- and p_n^+, together with asymptotic confidence intervals.

6 Optimality

We have established in Theorem 2.2 that after n evaluations of the function g, the approximation error of our algorithm is of polynomial order $n^{-\frac{1}{d-1}}$ when $d \geqslant 2$, and of exponential order $2^{-\beta n}$ when $d = 1$. The aim of this section is to show that these bounds are optimal, meaning that they cannot

be improved by any other algorithm under the sole general assumptions that we have made on the function g.

6.1 The case $d \geqslant 2$

When $d \geqslant 2$, we consider the following particular case:

- The random variable X is uniformly distributed on $\Omega = [0,1]^d$;
- The function g is defined by $g(x) = -x^1$ for $x = (x^1, \ldots, x^d) \in \Omega$;
- The threshold T is equal to 0.

Thus, in this setting, the failure probability $p := \mathbb{P}(g(X) > 0)$ is equal to 0. Clearly the function g satisfies the Lipschitz Assumption 2 with $L = 1$ and the level set Assumption 3 with $M = 1$.

Let us fix an integer $n = 2^{j(d-1)-1}$ for some $j \in \mathbb{N}^\star$ and n arbitrary points $x_1, \ldots, x_n$ in Ω. In the sequel, we construct a function $\tilde{g}$ on Ω such that

- $g(x_i) = \tilde{g}(x_i)$ for $1 \leqslant i \leqslant n$;
- $\tilde{p} := \mathbb{P}(\tilde{g}(X) > 0) \geqslant c n^{-\frac{1}{d-1}}$;
- $\tilde{g}$ satisfies Assumptions 2 and 3 with L and M independent of n.

The first fact ensures that any algorithm based on the points $(x_i)_{1 \leqslant i \leqslant n}$ leads to the same estimation for p and $\tilde{p}$. The second one ensures that $\tilde{p} - p$ is (at least) of order $n^{-\frac{1}{d-1}}$.

First, let us define the face

$$\mathcal{C} := \{x \in \Omega \,:\, g(x) = 0\} = \left\{x \in \Omega \,:\, x^1 = 0\right\}.$$

Consider the set

$$\mathcal{D}_j^\star = \{Q \in \mathcal{D}_j \,:\, \mathrm{dist}(Q, \mathcal{C}) = 0\}$$

of dyadic cubes with side length 2^{-j} which intersect $\mathcal{C}$ and the set

$$\tilde{\mathcal{D}}_j = \{Q \in \mathcal{D}_j^\star \,:\, \; x_i \notin Q,\; i = 1, \ldots, n\}$$

of dyadic cubes that intersect $\mathcal{C}$ and do not contain any point x_i. Since the cardinal of $\mathcal{D}_j^\star$ is equal to $2^{j(d-1)} = 2n$, the cardinal of $\tilde{\mathcal{D}}_j$ is at least n.

Second, for any cube Q and any $x \in \Omega$, let us introduce the piecewise affine function

$$h_Q(x) = \mathrm{dist}(x, Q^c) = \inf_{y \in Q^c} \|x - y\|_\infty$$

where $Q^c := \Omega \setminus Q$. The function h_Q is thus supported on Q and it is 1-Lipschitz for the ℓ^∞ norm.

Finally, consider the function $\tilde{g}$ defined as follows on Ω:

$$\tilde{g} = g + 2 \sum_{Q \in \tilde{\mathcal{D}}_j} h_Q.$$

By construction, the functions g and $\tilde{g}$ coincide on the cubes that do not belong to $\tilde{\mathcal{D}}_j$. In particular, $g(x_i) = \tilde{g}(x_i)$ for any $1 \leqslant i \leqslant n$.

Additionally, since $\sum_{Q \in \tilde{\mathcal{D}}_j} h_Q$ is 1-Lipschitz, the function $\tilde{g}$ is 3-Lipschitz, and therefore Assumption 2 holds with $L = 3$.

For any $Q \in \tilde{\mathcal{D}}_j$, if x_Q is the center of Q one has

$$\tilde{g}(x_Q) = g(x_Q) + 2h_Q(x_Q) = -2^{-j-1} + 2^{-j} = 2^{-j-1}.$$

Therefore $\tilde{g}(x) > 0$ for any $x \in Q$ such that $\|x - x_Q\|_\infty \leqslant \frac{2^{-j-1}}{3}$. As a consequence, since X has a uniform distribution on Ω, the failure probability associated to $\tilde{g}$ satisfies

$$\mathbb{P}(\tilde{g}(X) > 0) \geqslant n 3^{-d} 2^{-dj} = c n^{-\frac{1}{d-1}},$$

where $c = 3^{-d} 2^{-\frac{d}{d-1}}$.

Finally, let us prove the validity of the level set Assumption 3 for the function $\tilde{g}$. Just like g, the absolute value of $\tilde{g}$ is smaller than 2^{-j} on the cubes $Q \in \mathcal{D}_j^\star$ and larger elsewhere. Therefore, when $\delta \geqslant 2^{-j}$, it is readily seen that

$$\lambda(\{x \in \Omega \,:\, |\tilde{g}(x)| \leqslant \delta\}) \leqslant \delta.$$

For the values $\delta \leqslant 2^{-j}$, we know that $\{x \in \Omega \,:\, |\tilde{g}(x)| \leqslant \delta\}$ is contained in the union of the cubes $Q \in \mathcal{D}_j^\star$. If $Q \notin \tilde{\mathcal{D}}_j$, then

$$\lambda(\{x \in Q \,:\, |\tilde{g}(x)| \leqslant \delta\}) \leqslant \delta 2^{-j(d-1)}.$$

The cubes $Q \in \tilde{\mathcal{D}}_j$ are treated by noticing that on such a cube, the function $\tilde{g}(x) = -x^1 + 2h_Q(x)$ is a rescaled version of the function $g^*(x) = -x^1 + 2h_\Omega(x)$ defined on Ω. The gradient of this function is piecewise constant with $\|\nabla g^*(x)\|_1 \geqslant 1$ and therefore $|\nabla g^*(x)| \geqslant \frac{1}{\sqrt{d}}$ almost everywhere on Ω. In addition g^* vanishes on a polyhedral shaped set S of $(d-1)$-dimensional measure $1 < H < \infty$ since in particular $g^*(x) = 0$ if $x^1 = 0$. Using the coarea formula (2.2), this yields

$$\lambda(\{x \in \Omega \,:\, |g^*(x)| \leqslant \delta\}) \leqslant 2\sqrt{d} H \delta,$$

for $\delta > 0$ small enough, and therefore

$$\lambda(\{x \in \Omega \,:\, |g^*(x)| \leqslant \delta\}) \leqslant B\delta,$$

for all value of $\delta > 0$ up to possibly taking a constant B larger than $2\sqrt{d}H$. By rescaling

$$\lambda(\{x \in Q \,:\, |\tilde{g}(x)| \leqslant \delta\}) \leqslant B\delta 2^{-j(d-1)}.$$

for all $\delta \leqslant 2^{-j}$. Summing on all $Q \in \mathcal{D}_j^\star$, since $H > 1$ and $|\mathcal{D}_j^\star| = 2n$, we find that

$$\lambda(\{x \in \Omega \,:\, |\tilde{g}(x)| \leqslant \delta\}) \leqslant B\delta.$$

This shows that Assumption 3 holds with $M = B$ independent of n.

This proves the optimality of the approximation error rate of our algorithm.

6.2 The case $d = 1$

The idea is the same as for the case $d \geqslant 2$. More precisely, we consider the following setting:

- The random variable X is uniformly distributed on $\Omega = [0, 1]$;
- The function g is defined by $g(x) = -x$;
- The threshold T is equal to 0.

As in the previous subsection, the failure probability $p := \mathbb{P}(g(X) > 0)$ is thus equal to 0, and the function g satisfies Assumption 2 with $L = 1$ and Assumption 3 with $M = 1$.

Let us fix an integer $n \in \mathbb{N}^\star$ and n points $x_1, \ldots, x_n$ in Ω. As before, the idea is to construct a function $\tilde{g}$ on Ω such that

- $g(x_i) = \tilde{g}(x_i)$ for $1 \leqslant i \leqslant n$;
- $\tilde{p} := \mathbb{P}(\tilde{g}(X) > 0) \geqslant c2^{-n}$.
- $\tilde{g}$ satisfies Assumptions 2 and 3 with L and M independent of n.

First, we define $I_{n+1} := [0, 2^{-n}]$ and, for $1 \leqslant j \leqslant n$, $I_j := [2^{-j}, 2^{-(j-1)}]$. To mimic the previous notation, this set of $(n + 1)$ intervals is denoted $\mathcal{D}_n^\star$ and, accordingly,

$$\tilde{\mathcal{D}}_n = \{I \in \mathcal{D}_n^\star \,:\, \forall i = 1, \ldots, n,\ x_i \notin I\}$$

stands for the set of intervals that do not contain any point x_i. Since the cardinal of $\mathcal{D}_n^\star$ is equal to $(n + 1)$, the cardinal of $\tilde{\mathcal{D}}_n$ is at least equal to 1.

Second, for any interval I and any $x \in \Omega$, we consider the 1-Lipschitz function

$$h_I(x) = \mathrm{dist}(x, I^c) = \inf_{y \in I^c} |x - y|.$$

Finally, we pick one interval $J \in \tilde{\mathcal{D}}_n$ and define the function $\tilde{g}$ defined as follows

$$\tilde{g} = g + 4h_J.$$

As before, the functions g and $\tilde{g}$ coincide on $\Omega \backslash J$. In particular, $g(x_i) = \tilde{g}(x_i)$ for any $1 \leqslant i \leqslant n$.

Additionally, the function $\tilde{g}$ is 5-Lipschitz, and therefore Assumption 2 holds with $L = 5$. Since $\tilde{g}$ vanishes at $x = 0$ and (at most) at two other points inside J where its gradient is larger than 3, it is also easily seen that Assumption 3 holds with $M = 7/3$.

If x_J denotes the center of J, then one has

$$\tilde{g}(x_J) = g(x_J) + 4h_J(x_J) = -\frac{3}{4}2^{-(j-1)} + 4 \times 2^{-(j+1)} = 2^{-(j+1)} > 0,$$

in the case $J = I_j = [2^{-j}, 2^{-(j-1)}]$, $1 \leqslant j \leqslant n$, and

$$\tilde{g}(x_J) = g(x_J) + 4h_J(x_J) = -\frac{1}{2}2^{-n} + 4 \times 2^{-(n+1)} = 3 \times 2^{-(n+1)} > 0,$$

in the case $J = I_{n+1} = [0, 2^{-n}]$. Since $\tilde{g}$ has Lipschitz constant 5, it follows that $\{x \in Q : \tilde{g}(x) > 0\}$ always contains an interval of length larger than $\frac{1}{5}2^{-n}$. As a consequence, since X has a uniform distribution on Ω, the failure probability associated to $\tilde{g}$ is such that

$$\mathbb{P}(\tilde{g}(X) > 0) \geqslant c2^{-n},$$

with $c = \frac{1}{5}$.

This proves the optimality of the approximation error rate of our algorithm.

7 Proof of Theorem 4.4

Consistency and unbiasedness are clear by Remark 4.2. The asymptotic normality is a consequence of the delta method. Remember that for $N \geqslant 1$ and $u, v \in \mathcal{T}$, we denote

$$C_N^v := \sum_{i=1}^{N} \mathbf{1}_{\left\{X_i^{\bar{v}} \in Q(v)\right\}}, \quad q_N(v) := \frac{C_N^v}{N} \quad \text{and} \quad p_N(u) := \prod_{v \in a(u)} q_N(v).$$

First of all, let us recall the (classical) multidimensional CLT.

Lemma 7.1 (Multidimensional CLT)**.** *For all* $w \in \Lambda$*,*

$$q_N(w) = \frac{C_N^w}{N} \xrightarrow[N\to\infty]{a.s.} q(w).$$

Let us denote by $\mathbf{C}_N$ *the random vector* $(C_N^w)_{w\in\Lambda}$ *and* $\mathbf{q}$ *the vector* $(q(w))_{w\in\Lambda}$*. We have*

$$\sqrt{N}\left[\frac{\mathbf{C}_N}{N} - \mathbf{q}\right] \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(0,\Gamma),$$

where the covariance matrix Γ *is given by*

$$\Gamma(v,w) = \begin{cases} q(v)(1-q(v)) & \text{if } v = w, \\ -q(v)q(w) & \text{if } v \neq w \text{ and } \bar{v} = \bar{w}, \\ 0 & \text{otherwise.} \end{cases}$$

From the latter we immediately deduce that

$$p_N(\mathcal{S}) = \sum_{u\in\mathcal{S}} p_N(u) = \sum_{u\in\mathcal{S}} \prod_{v\leqslant u} q_N(v) \xrightarrow[N\to\infty]{a.s.} p(\mathcal{S}).$$

Next, we may rewrite $p(\mathcal{S})$ as a function of $\mathbf{q} = (q(v))_{v\in\Lambda}$ as follows:

$$p(\mathcal{S}) = F(\mathbf{q}) := \sum_{u\in\mathcal{S}} \prod_{v\leqslant u} q(v).$$

The partial derivative of F with respect to $q(v)$, denoted $\partial_v F$, is given by

$$\partial_v F(\mathbf{q}) = \sum_{\substack{u\in\mathcal{S} \\ v\leqslant u}} \prod_{\substack{w\in a(u) \\ w\neq v}} q(w) = \sum_{\substack{u\in\mathcal{S} \\ v\leqslant u}} \frac{p(u)}{q(v)}.$$

If $\nabla F = (\partial_v F)_{v\in\Lambda}$ is seen a row vector, the delta method ensures that

$$\sqrt{N}(F(\mathbf{q}_N) - F(\mathbf{q})) \xrightarrow[N\to\infty]{} \mathcal{N}(0,\sigma^2),$$

where

$$\begin{aligned}
\sigma^2 = (\nabla F)\Gamma(\nabla F)^T &= \sum_{v,w\in\Lambda} (\partial_v F)\Gamma(v,w)(\partial_w F) \\
&= \sum_{v\in\Lambda} \Gamma(v,v)(\partial_v F)^2 + \sum_{\substack{v\neq w\in\Lambda \\ \bar{v}=\bar{w}}} \Gamma(v,w)(\partial_v F)(\partial_w F) \\
&= \sum_{v\in\Lambda} q(v)(1-q(v)) \left[\sum_{\substack{u\in\mathcal{S} \\ v\leqslant u}} \frac{p(u)}{q(v)}\right]^2 - \sum_{\substack{v\neq w\in\Lambda \\ \bar{v}=\bar{w}}} q(v)q(w) \left[\sum_{\substack{u\in\mathcal{S} \\ v\leqslant u}} \frac{p(u)}{q(v)}\right] \left[\sum_{\substack{u'\in\mathcal{S} \\ w\leqslant u'}} \frac{p(u')}{q(w)}\right] \\
&= \sum_{v\in\Lambda} \frac{1-q(v)}{q(v)} \left[\sum_{\substack{u\in\mathcal{S} \\ v\leqslant u}} p(u)\right]^2 - \sum_{\substack{v\neq w\in\Lambda \\ \bar{v}=\bar{w}}} \left[\sum_{\substack{u\in\mathcal{S} \\ v\leqslant u}} p(u)\right] \left[\sum_{\substack{u'\in\mathcal{S} \\ w\leqslant u'}} p(u')\right].
\end{aligned}$$

Let us define

$$A := \sum_{v\in\Lambda} \frac{1-q(v)}{q(v)} \left[\sum_{\substack{u\in\mathcal{S} \\ v\leqslant u}} p(u)\right]^2 \quad \text{and} \quad B := \sum_{\substack{v\neq w\in\Lambda \\ \bar{v}=\bar{w}}} \left[\sum_{\substack{u\in\mathcal{S} \\ v\leqslant u}} p(u)\right] \left[\sum_{\substack{u'\in\mathcal{S} \\ w\leqslant u'}} p(u')\right].$$

We have, since $(v \leqslant u) \Leftrightarrow v \in a(u)$,

$$\begin{aligned}
A &= \sum_{v\in\Lambda} \sum_{\substack{u\in\mathcal{S} \\ v\leqslant u}} p(u)^2 \frac{1-q(v)}{q(v)} + \sum_{v\in\Lambda} \sum_{\substack{u,u'\in\mathcal{S} \\ v\leqslant u \\ v\leqslant u'}} p(u)p(u') \frac{1-q(v)}{q(v)} \\
&= \sum_{u\in\mathcal{S}} \sum_{v\in a(u)} p(u)^2 \frac{1-q(v)}{q(v)} + \sum_{u\neq u'\in\mathcal{S}} \sum_{v\in a(u)\cap a(u')} p(u)p(u') \frac{1-q(v)}{q(v)}.
\end{aligned}$$

Similarly, we get

$$B = \sum_{\substack{v\neq w\in\Lambda \\ \bar{v}=\bar{w}}} \sum_{\substack{u,u'\in\mathcal{S} \\ v\leqslant u \\ w\leqslant u'}} p(u)p(u') = \sum_{u\neq u'\in\mathcal{S}} p(u)p(u').$$

Since $a(u) \cap a(u') = a(u \wedge u')$, this finally yields the claimed expression for σ^2.

References

[1] S.K. Au and J.L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263–277, 2001. 1

[2] S.K. Au and J.L. Beck. Subset simulation and its application to seismic risk based on dynamic analysis. *Journal of Engineering Mechanics*, 129(8):901–917, 2003. 1

[3] J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Stat. Comput.*, 22(3):773–793, 2012. 1, 5

[4] J. Bect, L. Li, and E. Vazquez. Bayesian subset simulation. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):762–786, Jan 2017. 1

[5] J.A. Bucklew. *Introduction to rare event simulation*. Springer Series in Statistics. Springer-Verlag, New York, 2004. 1

[6] F. Cérou, P. Del Moral, T. Furon, and A. Guyader. Sequential Monte Carlo for rare event estimation. *Stat. Comput.*, 22(3):795–808, 2012. 1

[7] F. Cérou and A. Guyader. Fluctuation analysis of adaptive multilevel splitting. *Ann. Appl. Probab.*, 26(6):3319–3380, 2016. 1

[8] F. Cérou, A. Guyader, and M. Rousset. Adaptive multilevel splitting: Historical perspective and recent results. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(4):043108, 2019. 1

[9] A. Cohen, R. Devore, G. Petrova, and P. Wojtaszczyk. Finding the minimum of a function. *Methods Appl. Anal.*, 20(4):365–381, 2013. (document), 1

[10] O. Ditlevsen and H.O. Madsen. *Structural reliability methods*, volume 178. Wiley New York, 1996. 1

[11] L.C. Evans and R.F. Gariepy. *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992. 2.1

[12] A. Guyader, N. Hengartner, and E. Matzner-Løber. Simulation and estimation of extreme quantiles and extreme probabilities. *Applied Mathematics and Optimization*, 64:171–196, 2011. 1

[13] A. Harbitz. An efficient sampling method for probability of failure calculation. *Structural Safety*, 3(2):109–115, 1986. 1

[14] H. Kahn and T.E. Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards Appl. Math. Series*, 12:27–30, 1951. 1

[15] J. Neveu. Arbres et processus de Galton-Watson. *Annales de l'I.H.P*, 22(2):199–207, 1986. 3.1

[16] C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2006. 1

[17] R. Schöbi, B. Sudret, and J. Wiart. Polynomial-chaos-based Kriging. *Int. J. Uncertain. Quantif.*, 5(2):171–193, 2015. 1

[18] L. Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1701–1762, 1994. With discussion and a rejoinder by the author. 4.2, 4.2