Alea 7, 21–24 (2010)

Large deviations principle for the Adaptive Multilevel Splitting Algorithm in an idealized setting

Charles-Edouard Bréhier

Université de Lyon
Université Claude Bernard Lyon 1
CNRS UMR 5208
Institut Camille Jordan
43 blvd. du 11 novembre 1918
F-69622 Villeurbanne cedex
France
E-mail address: brehier@math.univ-lyon1.fr

Abstract. The Adaptive Multilevel Splitting (AMS) algorithm is a powerful and versatile method for the simulation of rare events. It is based on an interacting (via a mutation-selection procedure) system of replicas, and depends on two integer parameters: $n \in \mathbb{Z}_+^*$ the size of the system and the number $k \in \{1, \dots, n-1\}$ of the replicas that are eliminated and resampled at each iteration.

In an idealized setting, we analyze the performance of this algorithm in terms of a Large Deviations Principle when n goes to infinity, for the estimation of the (small) probability $\mathbb{P}(X > a)$ where a is a given threshold and X is real-valued random variable. The proof uses the technique introduced in Bréhier et al. (2015): in order to study the log-Laplace transform, we rely on an auxiliary functional equation.

Such Large Deviations Principle results are potentially useful to study the algorithm beyond the idealized setting, in particular to compute rare transitions probabilities for complex high-dimensional stochastic processes.

1. Introduction

1.1. *Splitting algorithms for rare event simulation.* In many problems from engineering, biology, chemistry, physics or finance, rare events are often critical and have a huge impact on the phenomena which are studied. From a general mathematical perspective, we may consider the following situation: let $(X_t)_{t\in\mathbb{T}}$, where $\mathbb{T} = \mathbb{Z}_+$ or $\mathbb{R}$, be a (discrete or continuous in time) stochastic process, taking values in $\mathbb{R}^d$. Assume that $A, B \subset \mathbb{R}^d$ are two *metastable regions*: starting from a

2000 *Mathematics Subject Classification.* 65C05, 65C35, 62G30, 60F10.

Key words and phrases. Monte-Carlo simulation, rare events, multilevel splitting, large deviations.

The author acknowledges support from the SNF grant 200020-149871/1, which founded his post-doctoral research at University of Neuchâtel (Switzerland).

neighborhood of A (resp. of B), the probability that the process reaches B (resp. A) before hitting A (resp. B) is very small (typically, less than 10^{-10}). As a consequence, a direct numerical Monte-Carlo with an ensemble of size N does not provide significant results when N is reasonably large (typically, less than 10^{10}) in real-life applications.

Even if theoretical asymptotic expansions on quantities of interest are (often, but not always) available - such as the Kramers-Arrhenius law given for instance by the Freidlin-Wentzell Large Deviations Theory or Potential Theory for the exit problem of a diffusion process in the small noise regime - in practice their explicit computation is not possible (for instance when the dimension is large) and numerical simulations are unavoidable.

It is thus essential to propose efficient and general methods, and to rigorously study their consistency and efficiency properties. Two main families of methods have been introduced in the 1950's and studied extensively since then, in order to improve the Monte-Carlo simulation algorithms, in particular for rare events: importance sampling and importance splitting (see for instance Asmussen and Glynn (2007), Rubino and Tuffin (2009) for general reviews of these methods and Kahn and Harris (1951) for the historical introduction of importance splitting). The main difference between these two methods is the following: the first one is *intrusive*, meaning that the dynamics of the stochastic process (more generally, the distribution of the random variable of interest) is modified so that the probability that the event of interest increases and in a Monte-Carlo simulation it is realized more often, while the second is not intrusive and can thus be used more directly for complex problems. Instead, for importance splitting strategies, the state space is decomposed as a nested sequence of regions which are visited sequentially and more easily by an interacting system of replicas.

In this paper, we focus on an importance splitting strategy which is known as the Multilevel Splitting approach and describe it in the following setting. Let $h : \mathbb{R}^d \to \mathbb{R}$ be a given function and assume we want to estimate the probability $p = \mathbb{P}(X > a)$ that a real-valued random variable $X = h(Y)$ (where Y is a $\mathbb{R}^d$-valued random variable) belongs to $(a, +\infty)$ for a given threshold $a \in \mathbb{R}$. This situation is not restrictive for many applications; indeed, we may take $X = \mathbb{1}_{\tau_B < \tau_A}$ and any $a \in (0, 1)$ in the situation described above, where τ_A and τ_B are the hitting times of A and B by the process X. A key assumption, for the analysis developed in the paper, on the distribution of X is the following: we assume that the cumulative distribution function F of X - *i.e.* $F(x) = \mathbb{P}(X \leq x)$ for any $x \in \mathbb{R}$ - is continuous; for convenience, we also assume that $F(0) = 0$ - *i.e.* $X > 0$ almost surely.

1.2. *Multilevel splitting.* The multilevel splitting approach (see Kahn and Harris (1951), Glasserman et al. (1999), Cérou et al. (2012) for instance) is based on the following decomposition of p as a telescoping product of conditional probabilities:

$$p = \mathbb{P}(X > a) = \prod_{i=1}^{N} \mathbb{P}(X > a_i \big| X > a_{i-1}), \tag{1.1}$$

where $a_0 = 0 < a_1 < \ldots < a_N = a$ is a sequence of non-decreasing *levels*. In other words, the realization of the event $\{X > a\}$ is split into the realizations of the N events $\{X > a_i\}$ conditionally on $\{X > a_{i-1}\}$; each event has a larger probability than the initial one and is thus much easier to realize. Then each of

the conditional probabilities is estimated separately, for instance with independent Monte-Carlo simulations, or using a Sequential Monte-Carlo technique with a splitting of successful trajectories. This approach has been studied with different points of view and variants under different names in the literature – nested sampling Skilling (2006), Skilling (2007), subset simulation Au and Beck (2001), RESTART (REpetitive Simulation Trials After Reaching Thresholds) Villén-Altamirano and Villén-Altamirano (1991), Villén-Altamirano and Villén-Altamirano (1994).

For future reference, we introduce the following (unbiased) estimator of p given by the multilevel splitting approach with N levels and n independent replicas (per level):

$$\hat{p}_n^N = \prod_{i=1}^{N} \frac{1}{n} \sum_{m=1}^{n} \mathbb{1}_{X_m^{(i)} > a_i}, \tag{1.2}$$

where the random variables $(X_m^{(i)})_{1 \le m \le N, 1 \le i \le N}$ are independent and the distribution of $X_m^{(i)}$ is $\mathcal{L}(X|X > a_{i-1})$. Thus $\hat{p}_n^N$ is a product of N independent Monte-Carlo estimators of the conditional probabilities $\mathbb{P}(X > a_i \big| X > a_{i-1})$ appearing in (1.1).

The efficiency of the algorithm crucially depends on the choice of the sequence of levels $(a_i)_{1 \le i \le N}$: for a fixed number of levels N, the asymptotic variance of the estimator (when $n \to +\infty$) is minimized when the conditional probabilities satisfy $\mathbb{P}(X > a_i \big| X > a_{i-1}) = p^{1/N}$ for all $1 \le i \le N$; moreover under this condition the variance is of size (to $-p^2 \log(p)/n$) when N goes to infinity - see for instance Cérou et al. (2012) for more details.

In general, finding a sequence of levels $(a_i)_{1 \le i \le N}$ such that the conditional probabilities $\mathbb{P}(X > a_i \big| X > a_{i-1})$ are equal, or at least of the same order, is not an easy task. Moreover, if the levels are not chosen that way, the performance of the algorithm may be very poor. Computing the levels adaptively, following the above guideline, is a fruitful approach, which gives Adaptive Multilevel splitting algorithm.

1.3. *The Adaptive Multilevel Splitting (AMS) algorithm.* Versions of multilevel splitting algorithms with an adaptive computation of the levels have been proposed in Cérou and Guyader (2007), and studied extensively in the last years, see for instance Bréhier et al. (2015), Bréhier et al. (2014), Cérou and Guyader (2014), Guyader et al. (2011), Simonnet (2014), Walter (2014). It is essential and non trivial to check that these adaptive versions still give reliable results, and to prove they do it efficiently. In addition to the original reference Cérou and Guyader (2007), we refer to the presentation given in Bréhier et al. (2015), in particular for notation and main properties of the algorithm.

Let us describe the algorithm we study in this paper; a precise statement is provided by Algorithm 1 below. Two integer parameters are required: $n \in \mathbb{Z}_+^*$ is the size of the system of interacting replicas; at each iteration a selection-mutation procedure leads to eliminate and resample $k \in \{1, \dots, n-1\}$ of them.

The deterministic sequence $(a_i)_{1 \le i \le N}$ of the fixed-level algorithm is replaced with a sequence of random variables, referred to as levels, $Z^0, Z^1, \dots$, which are computed as k-th order statistics of the n replicas at each iteration. the k replicas with values less than Z are eliminated, and are then resampled using the conditional distribution $\mathcal{L}(X|X > Z)$ of X conditionally on $\{X > Z\}$. The conditional

probabilities are replaced with $1-k/n$: indeed, the levels are estimators of empirical quantiles of the conditional distribution.

The number of iterations of the algorithm is a random variable: the algorithm stops when a level satisfies $Z^j \geq a$. At the end, one obtains the estimator

$$\hat{p}^{n,k} = \frac{1}{n}\mathrm{Card}\left\{i;\, X_i^{J^{n,k}} \geq a\right\}\left(1-\frac{k}{n}\right)^{J^{n,k}}, \tag{1.3}$$

where $J^{n,k}$ is the number of iterations and $\left(X_i^{J^{n,k}}\right)_{1\leq i\leq n}$ is the system of replicas at the final iteration.

The results proved in this article can be applied for practical rare event estimation problems only if one is able to sample according to the conditional distribution $\mathcal{L}(X|X>z)$ for any value of z: this is part of the idealized setting assumption (see Section 2.1). Even if it is rarely satisfied in real-life applications, the study of the algorithm in that idealized setting is already challenging and yields very interesting results, which give a guideline for a generalization beyond this simplified case, but at the price of a much more intricate analysis. We refer to Section 8 for elements on the non-idealized situation.

The adaptive multilevel splitting algorithm is non-intrusive: contrary to importance sampling techniques, the (conditional) distributions do not need to be modified during the simulation. The generation of random variables comes from a black-box, and this property is a strength of the approach. The limitation is in the requirement of an exact sampling.

Let us recall some important results obtained recently on the estimator (1.3), in the setting of this article. On the one hand, in Guyader et al. (2011) (see also Simonnet (2014), Walter (2014)), it has been proved that $\hat{p}^{n,1}$ is an unbiased estimator of p – meaning that $\mathbb{E}[\hat{p}^{n,1}] = p$ – for all values of $n \geq 2$. This result was extended to arbitrary $k \in \{1,\dots,n-1\}$ in Bréhier et al. (2015): $\hat{p}^{n,k}$ is an unbiased estimator of p. On the other hand, efficiency properties have been studied with the proof of Central Limit Theorems in two different kinds of regimes: either k is fixed and $n \to +\infty$ (see Guyader et al. (2011) and Simonnet (2014) for the case $k=1$ and Bréhier et al. (2014) for arbitrary k), or both k and n go to infinity, in such a way that k/n converges to $\alpha \in (0,1)$ - which gives a fixed proportion of resampled replicas at each iteration, see Cérou and Guyader (2007). We also mention recent works which go beyond the idealized setting considered here: an unbiasedness result in Bréhier et al. and a central limit theorem in Cérou and Guyader (2014).

We recall the following observation, which ensures the efficiency of the adaptive multilevel splitting approach: the asymptotic variance is the same for both the adaptive and the non-adaptive versions – *i.e.* with the estimator (1.2) in the limit $N \to +\infty$ with an optimal choice of the levels a_i. Moreover, it is much smaller when the probability p is small than when using a crude Monte-Carlo estimator, *i.e.* the empirical average

$$\overline{p}_n = \frac{1}{n}\sum_{m=1}^{n} \mathbb{1}_{X_m>a}, \tag{1.4}$$

where the random variables $(X_m)_{1\leq m\leq n}$ are independent and identically distributed, with distribution $\mathcal{L}(X)$. Indeed, for the splitting algorithm, asymptotically (in $n\to\infty$) the variance is of the order $-p^2 log(p)/n$, while it is of the order p/n for the crude Monte-Carlo estimator.

1.4. *Large deviations principle for the AMS algorithm.* In this paper, we provide a similar comparison of the efficiency of adaptive and non-adaptive splitting algorithms, also with crude Monte-Carlo estimators. However, the comparison is based on a different criterion, which seems to be original compared with existing literature: we prove a Large Deviations Principle principle for the distribution of the estimator $\hat{p}^{n,k}$ given by (1.3) when k is fixed and $n \to +\infty$. Our main result is Theorem 3.1, which in particular yields that for any given $\epsilon > 0$

$$\frac{1}{n} \log\Big(\mathbb{P}\big(|\hat{p}^{n,k} - p| \geq \epsilon\big)\Big) \underset{n\to+\infty}{\to} -\min\big(I(p+\epsilon), I(p-\epsilon)\big) < 0.$$

The rate function I - see (3.1) - obtained in Theorem 3.1 does not depend on k. We then compare this rate function with the rate function $\mathcal{I}$ (see (6.2)) obtained for the crude Monte-Carlo estimator $\overline{p}_n$ given by (1.4) (thanks to Cramer's Theorem, see Dembo and Zeitouni (2010)) and show that for any $y \in (0,1) \setminus p$ we have $I(y) > \mathcal{I}(y)$. In addition, it is clear that $\mathcal{I}(p) = I(p) = 0$ and $\mathcal{I}(y) = I(y) = +\infty$ if $y \notin (0,1)$. Thus

$$\frac{\mathbb{P}(\hat{p}^{n,k} - p > \epsilon)}{\mathbb{P}(\overline{p}_n - p > \epsilon)} \underset{n\to+\infty}{\to} 0.$$

In other words, for large n, the probability that $\hat{p}^{n,k}$ deviates from p from above (and similarly from below) with threshold $\epsilon > 0$ decreases exponentially fast, at a faster rate than for $\overline{p}_n$.

Moreover, the non-adaptive, fixed-levels estimator $\hat{p}_n^N$ given by (1.2) satisfies a Large Deviations Principle when $n \to +\infty$ with rate function $\mathcal{I}_N$ for a fixed number of levels N and when the levels are chosen in an optimal way, namely such that $\mathbb{P}(X > a_i | X > a_{i-1}) = p^{1/N}$ does not depend on i. We then show that $\lim_{N\to+\infty} \mathcal{I}_N(y) \leq I(y)$ for any $y \in \mathbb{R}$: this inequality is sufficient to prove that asymptotically the adaptive algorithm performs (at least) as well as the non-adaptive version in this setting, in terms of large deviations.

Gathering this new result with those recalled above, in the idealized setting, the consistency (estimators remain unbiased) and the efficiency (asymptotic variance and rate function) are preserved (at least) when considering adaptive versions of the multilevel splitting algorithms. The main advantage of the adaptive version is the reduced number of parameters required *a priori*: only two integers n and k, instead of n and a well-chosen family of levels $(a_i)_{1\leq i\leq N}$.

The proof of Theorem 3.1 relies on the technique introduced in Bréhier et al. (2015), and which has also been used in Bréhier et al. (2014) to prove a central limit theorem in the same regime (fixed k and $n \to +\infty$). First, we restrict the study of the properties of the algorithm to the case when X is exponentially distributed with parameter 1 (this key remark was introduced first in Guyader et al. (2011) and used also in Simonnet (2014), Walter (2014)). Instead of directly studying $\hat{p}^{n,k}$, we focus on its logarithm $\log(\hat{p}^{n,k})$, and prove that when considering the algorithm as depending on an initial condition x, the Laplace transform of the latter is solution of a functional (integral) equation (with respect to the x variable) - thanks to a decomposition of the realizations of the algorithm according to the value of the first level. To study the equation in the asymptotic regime considered in this paper, we then derive a linear ordinary differential equation of order k and perform an asymptotic expansion. Note that we do not give all details for the derivation of the differential equations and the basic properties of its coefficients; for some points we

refer the reader to Bréhier et al. (2015) where all the arguments are proved with details and here we mainly focus on the proof of the new asymptotic results as well as on the interpretation of the Large Deviations Principle.

It seems that studying the performance of multilevel splitting algorithms via Large Deviations Principle is an original approach, which can complement the other studies which are all based on Central Limit Theorems. Note that Glasserman et al. (1998) also uses Large Deviations techniques to study the efficiency of (fixed-levels) multilevel splitting algorithms; however asymptotics are not in terms of parameters of the algorithm (like n), but are in terms of a parameter $\epsilon \to 0$ such that $X = X^\epsilon$ satisfies a Large Deviations Principle, and $p \to p^\epsilon \to 0$. In the present paper, we proved a result in a specific regime (k is fixed, $n \to +\infty$), in the idealized setting. To go further, it would be interesting to look at other regimes ($k, n \to +\infty$ with $k/n \to \alpha \in (0,1)$) and to go beyond the idealized setting. These issues are discussed in Section 8 and will be the subject of future investigation.

1.5. *Organization of the paper.* The paper is organized as follows. In Section 2, we introduce our main assumptions (Section 2.1), describe the Adaptive Multilevel Splitting algorithm (Section 2.2) and recall several of its fundamental properties used in the sequel of the article (Section 2.3). The main result of this paper is given in Section 3: it is the Large Deviations Principle for the estimator of the probability given by the AMS estimator, see Theorem 3.1. An important auxiliary result is stated in Section 4, and proofs are carried over in Section 5 - some technical estimates being proved in Section 7. We compare the performance in terms of the Large Deviations Principle of the AMS algorithm with two other methods in Section 6: a crude Monte-Carlo method and a fixed-level splitting method. Finally, we give some concluding remarks and perspectives in Section 8.

2. Description of the Adaptive Multilevel Splitting algorithm

2.1. *Assumptions.* Let X be some real random variable. For simplicity, we assume that $X > 0$ almost surely.

We want to estimate the probability $p = \mathbb{P}(X > a)$, where $a > 0$ is some threshold. When a goes to $+\infty$, p goes to 0 and we have to estimate the probability of a rare event.

We make a fundamental assumption on the distribution of X.

Assumption 2.1. *Let F denote the cumulative distribution function of X: we assume that F is continuous.*

More generally, for both theoretical and practical purpose, we introduce for $0 \leq x \leq a$ the conditional probability

$$P(x) = \mathbb{P}(X > a | X > x); \tag{2.1}$$

we also denote by $\mathcal{L}\big(X|X > x\big)$ the associated conditional distribution, and $F(\cdot; x)$ its cumulative distribution function: for any $y > x$ we have $F(y;x) = \frac{F(y)-F(x)}{1-F(x)}$ whenever $F(x) < 1$.

We emphasize on two important equalities: $P(a) = 1$, and the estimated probability is $p = P(0)$; in fact, the distribution of X is equal to $\mathcal{L}\big(X|X > 0\big)$.

The *idealized setting* refers to the following assumptions:

- Assumption 2.1 is satisfied (*theoretical condition*);

- it is possible to sample according to the conditional distribution $\mathcal{L}\big(X|X > x\big)$ for any $x \in [0, a)$ (*practical condition*).

In view of a practical implementation of the algorithm, the second condition is probably the most restrictive. One may rely on some approximation of the conditional distribution $\mathcal{L}\big(X|X > x\big)$ thanks to a Metropolis-Hastings algorithm: in that case (see Cérou and Guyader (2014) for instance), the analysis we develop here does not apply, but gives an interesting insight for the behavior in the case of a large number of steps in the Metropolis-Hastings auxiliary scheme (rigorously, we treat the case of an infinite number of steps).

2.2. *The algorithm.* We now present the Adaptive Multilevel Splitting algorithm, under the assumptions of Section 2.1 above.

The algorithm depends on two parameters:

- the number of replicas n;
- the number $k \in \{1, \dots, n-1\}$ of replicas that are resampled at each iteration.

The other necessary parameters are the initial condition x and the stopping threshold a: the aim is to estimate the conditional probability $P(x)$ introduced in (2.1). For future reference, we denote by $\mathrm{AMS}(n, k; a, x)$ the algorithm.

The dependence with respect to x allows us below to state fundamental functional equations on useful observables of the estimator computed at the end of the iterations of the algorithm, as a function of x. In practice, we are interested in the case $x = 0$; in this situation, the algorithm is denoted by $\mathrm{AMS}(n, k; a)$.

Before we detail the algorithm, we introduce important notation. First, when we consider a random variable X_i^j, the subscript i denotes the index in $\{1, \dots, n\}$ of a replica, while the superscript j denotes the iteration of the algorithm.

Moreover, we use the following notation for order statistics. Let $Y = (Y_1, \dots, Y_n)$ be independent and identically distributed (i.i.d.) real valued random variables with continuous cumulative distribution function; then there exists almost surely a unique (random) permutation σ of $\{1, \dots, n\}$ such that $Y_{\sigma(1)} < \dots < Y_{\sigma(n)}$. For any $k \in \{1, \dots, n\}$, we then denote by $Y_{(k)} = Y_{\sigma(k)}$ the so-called k-th order statistic of the sample Y. Sometimes we need to specify the size of the sample of which we consider the order statistics: we then use the notation $Y_{(k,n)}$.

We are now in position to write the $\mathrm{AMS}(n, k; a, x)$ algorithm.

Algorithm 1 (Adaptive Multilevel Splitting, $\mathrm{AMS}(n, k; a, x)$)**.**
Initialization: *Set the initial level* $Z^0 = x$.

Sample n *i.i.d. realizations* $X_1^0, \dots, X_n^0$, *with distribution* $\mathcal{L}(X|X > x)$.

Define $Z^1 = X^0_{(k)}$, *the* k*-th order statistics of the sample* $X^0 = (X_1^0, \dots, X_n^0)$, *and* σ^1 *the (a.s.) unique associated permutation:* $X^0_{\sigma^1(1)} < \dots < X^0_{\sigma^1(n)}$.

Set $j = 1$.

Iterations (on $j \geq 1$**):** *While* $Z^j < a$*:*

- *Conditional on* Z^j, *sample* k *new independent random variables* $(Y_1^j, \dots, Y_k^j)$, *according to the law* $\mathcal{L}(X|X > Z^j)$.
- *Set*

$$X_i^j = \begin{cases} Y^j_{(\sigma^j)^{-1}(i)} & \text{if } (\sigma^j)^{-1}(i) \leq k \\ X_i^{j-1} & \text{if } (\sigma^j)^{-1}(i) > k. \end{cases}$$

In other words, we resample exactly k out of the n replicas, namely those with index i such that $X_i^{j-1} \leq Z^j$, i.e. such that $i \in \{\sigma^j(1), \dots, \sigma^j(k)\}$ (which is equivalent to $(\sigma^j)^{-1}(i) \leq k$). They are resampled according the the conditional distribution $\mathcal{L}(X|X > Z^j)$. The other replicas are not modified.

- *Define $Z^{j+1} = X^j_{(k)}$, the k-th order statistics of the sample $X^j = (X_1^j, \dots, X_n^j)$, and σ^{j+1} the (a.s.) unique associated permutation: $X^j_{\sigma^{j+1}(1)} < \dots < X^j_{\sigma^{j+1}(n)}$.*
- *Finally increment $j \leftarrow j+1$.*

End of the algorithm: *Define $J^{n,k}(x) = j-1$ as the (random) number of iterations. Note that $J^{n,k}(x)$ is such that $Z^{J^{n,k}(x)} < a$ and $Z^{J^{n,k}(x)+1} \geq a$.*

Observe for instance that $J^{n,k}(x) = 0$ if and only if $Z^1 > a$: we mean that in this case the algorithm has required 0 iteration, since the stopping condition at the beginning of the loop (on j) is satisfied without entering into the loop.

The estimator of the probability $P(x)$ is defined by

$$\hat{p}^{n,k}(x) = C^{n,k}(x) \left(1 - \frac{k}{n}\right)^{J^{n,k}(x)}, \tag{2.2}$$

with

$$C^{n,k}(x) = \frac{1}{n} \mathrm{Card} \left\{ i;\, X_i^{J^{n,k}(x)} \geq a \right\}. \tag{2.3}$$

The interpretation of the factor $C^{n,k}(x)$ is the following: it is the proportion of the replicas $X_i^{J^{n,k}(x)}$ which satisfy $X_i^j \geq a$: since $X_{(k)}^{J^{n,k}(x)} = Z^{J^{n,k}(x)+1} \geq a$, we have $C^{n,k}(x) \geq \frac{n-k+1}{n}$. Note that $C^{n,1}(x) = 1$.

When $x = 0$, to simplify notations we set $\hat{p}^{n,k} = \hat{p}^{n,k}(0)$.

2.3. *Properties of the AMS Algorithm 1.*

Well-posedness. We first recall some important results on the well-posedness of the algorithm. For more detailed statements and complete proofs, see Section 3.2 in Bréhier et al. (2015), in particular Proposition 3.2 there.

First, at each iteration j of the algorithm, conditional on the level Z^j, the resampling produces a family of n random variables $\left(X_i^j\right)_{1 \leq i \leq n}$ which are independent and identically distributed, with distribution $\mathcal{L}(X|X > Z^j)$. By Assumption 2.1, conditional on Z^j the latter conditional distribution also admits a continuous cumulative distribution function $F(\cdot; Z^j)$; as a consequence, almost surely the permutation σ^{j+1} is unique, and the level Z^{j+1} is well-defined.

Moreover, if we assume that $P(x) > 0$, almost surely the algorithm stops after a finite number of steps, for any values of k and n such that $1 \leq k \leq n-1$: the random variable $J^{n,k}(x)$ almost surely takes values in $\mathbb{Z}_+$, and the estimator $\hat{p}^{n,k}(x)$ is well-defined and takes values in $(0, 1]$.

Reduction to the exponential case. We now state properties that are essential for our theoretical study of the algorithm below.

One of the main tools in Bréhier et al. (2015) and Bréhier et al. (2014), which was also used in Guyader et al. (2011) in the case $k = 1$, is the restriction to the case where the random variables are exponentially distributed. More precisely, assume that $P(x) > 0$, and denote by $\mathcal{E}(1)$ the exponential distribution

with mean 1. Then in distribution the algorithm $\mathrm{AMS}(n,k;a)$ is equal to the algorithm $\mathrm{AMS}_{\mathrm{expo}}(n,k;-\log(p))$ in which we assume that the distribution is $\mathcal{E}(1)$; a similar result holds for $\mathrm{AMS}(n,k;a,x)$ when $x\in[0,a)$. In particular, the associated estimators are equally distributed. The main argument is the well-known equality of distribution $F(X)=U$ where U is uniformly distributed on $(0,1)$.

In the sequel, we state in Section 3 our results in the general setting - *i.e.* for $\mathrm{AMS}(n,k;a)$, with the probability p and the estimator $\hat{p}^{n,k}$ - but in the remaining of the paper we give proofs in the exponential case, namely for $\mathrm{AMS}_{\mathrm{expo}}(n,k;a_{\mathrm{expo}},x)$ with $a_{\mathrm{expo}}=-\log(p)$, and we omit the reference to the exponential case to simplify the notation. Whether we consider the general or the exponential case will be clear from the context.

3. The Large Deviations Principle result for the AMS algorithm

The main result of this article is the following Theorem 3.1, which states a Large Deviations Principle (in the sense of Dembo and Zeitouni (2010)) for the distribution $\mu^{n,k}=\mathcal{L}\big(\hat{p}^{n,k}\big)$ of $\hat{p}^{n,k}$ for fixed probability $p>0$ and $k\in\mathbb{Z}_+^*$, in the limit $n\to+\infty$.

Theorem 3.1. *Assume that $p\in(0,1)$ and $k\in\mathbb{Z}_+^*$ are fixed. Then the sequence $\left(\mu^{n,k}\right)_{n\in\mathbb{Z}_+,n>k}$ of distributions of the estimator $\hat{p}^{n,k}$ of p obtained by the $\mathrm{AMS}(n,k;a)$ algorithm satisfies a Large Deviations Principle with the rate function I defined by*

$$I(y)=\begin{cases}+\infty \text{ if } y\notin(0,1)\\ \log(y)\log(\frac{\log(p)}{\log(y)})+\log(\frac{y}{p}) \text{ if } y\in(0,1).\end{cases}\tag{3.1}$$

Observe that the rate function does not depend on k.

Note that the statement above is restricted to $p\in(0,1)$. Indeed, when $p=1$, we have almost surely $\hat{p}^{n,k}=1$ (the algorithm stops after 0 iteration). Moreover, we always estimate the probability of events which have a positive probability (otherwise the algorithm does not stop after a finite number of iterations).

The following Proposition describes some properties of the rate function I.

Proposition 3.2. *The rate function I is of class $\mathcal{C}^\infty$ on its domain $(0,1)$.*

Moreover, p is the unique minimizer of I: we have $I(p)=I'(p)=0$, $I''(p)=\frac{1}{-p^2\log(p)}>0$.

Finally, for any $y\in(0,1)\setminus\{p\}$ we have $I(y)>0$; I is decreasing on $(0,p)$ and is increasing on $(p,1)$.

Proof: Straightforward computations yield that for $y\in(0,1)$ we have

$$\frac{dI(y)}{dy}=\frac{\log(\log(p))-\log(\log(y))}{y},$$
$$\frac{d^2I(y)}{dy^2}=-\frac{\log(\log(p))-\log(\log(y))}{y^2}-\frac{1}{y^2\log(y)}.$$

□

Let $\epsilon\in(0,\max(p,1-p))$; then from Theorem 3.1 we have when $n\to+\infty$

$$\frac{1}{n}\log\Big(\mathbb{P}\big(|\hat{p}^{n,k}-p|\geq\epsilon\big)\Big)\underset{n\to+\infty}{\to}-\min\big(I(p+\epsilon),I(p-\epsilon)\big)<0.\tag{3.2}$$

Applying the Borel-Cantelli Lemma, we get the almost sure convergence $\hat{p}^{n,k} \to p$.

Remark 3.3. The almost sure limit is consistent with the unbiasedness result ($\mathbb{E}[\hat{p}^{n,k}] = p$) from Bréhier et al. (2015). There only the convergence in probability of $\hat{p}^{n,k}$ to p has been proved.

Note also that a Central Limit Theorem has been proved in Bréhier et al. (2014), in the same regime:

$$\sqrt{n}\big(\hat{p}^{n,k} - p\big) \to \mathcal{N}\big(0, -p^2\log(p)\big).$$

The asymptotic variance is given by $I''(p)$.

We conclude this section with a result showing that the choice of the regime p (and k) fixed and $n \to +\infty$ is crucial to get Theorem 3.1. Indeed, set $k = 1$, and for a given $\sigma > 0$ assume that n and p are related though the following formula: $-\log(p) = \sigma^2 n$. Then $\frac{p^{n,k}}{p}$ converges (in law) to a log-normal distribution, as stated in the following proposition.

Proposition 3.4. *If $-\log(p) = \sigma^2 n$, we have the convergence in distribution*

$$\lim_{n\to\infty} \frac{\hat{p}^{n,1}}{p} = \exp(\sigma Z - \sigma^2/2),$$

where $Z \sim \mathcal{N}(0,1)$.

The proof is postponed to Section 5.1, since it uses the same arguments as the proof of Theorem 3.1 in the case $k = 1$.

Let $\epsilon > 0$. Then (compare with (3.2) with ϵp instead of ϵ)

$$\mathbb{P}\big(|\frac{\hat{p}^{n,1}}{p} - 1| \geq \epsilon\big) \underset{n=-\frac{\log(p)}{\sigma^2}\to+\infty}{\longrightarrow} \mathbb{P}_{Z\sim\mathcal{N}(0,1)}\big(|\exp(\sigma Z - \sigma^2/2) - 1| \geq \epsilon\big)\Big) > 0,$$

where the limit is positive, while owing to (3.2) when p fixed, $\mathbb{P}\big(|\frac{\hat{p}^{n,1}}{p} - 1| \geq \epsilon\big)$ converges to 0 exponentially fast when $n \to +\infty$.

4. Strategy of the proof

To prove Theorem 3.1, we in fact first prove a Large Deviations Principle for $\tilde{\mu}^{n,k} = \mathcal{L}\big(\log(\hat{p}^{n,k})\big)$, with rate function J given below.

Proposition 4.1. *Assume that $p \in (0,1)$ and $k \in \mathbb{Z}_+^*$ are fixed. Then the sequence $\big(\tilde{\mu}^{n,k}\big)_{n\in\mathbb{Z}_+, n>k}$ of distributions of $\log(\hat{p}^{n,k})$ obtained by the $\mathrm{AMS}(n,k;a)$ algorithm satisfies a Large Deviations Principle with the rate function J defined by*

$$J(z) = \begin{cases} +\infty \text{ if } z \geq 0 \\ z - \log(p) - z\log(\frac{z}{\log(p)}) \text{ if } z < 0. \end{cases} \tag{4.1}$$

Then Theorem 3.1 immediately follows from Proposition 4.1 and the application of the contraction principle (see Dembo and Zeitouni (2010), Theorem 4.2.1): we have $\hat{p}^{n,k} = \exp\big(\log(\hat{p}^{n,k})\big)$, and we obtain the rate function with the identity $I(y) = J(\log(y))$.

The proof of Proposition 4.1 relies on the use of the Gärtner-Ellis Theorem (see Theorem 2.3.6 in Dembo and Zeitouni (2010)) and the asymptotic analysis when $n \to +\infty$ of the log-Laplace transform of $\tilde{\mu}^{n,k}$.

Proposition 4.2. *Set for any* $1 \leq k \leq n-1$ *and any* $\lambda \in \mathbb{R}$

$$\Lambda_{n,k}(\lambda) = \log\Big(\mathbb{E}\Big[\exp\big(\lambda \log(\hat{p}^{n,k})\big)\Big]\Big). \tag{4.2}$$

Then for any fixed $k \in \mathbb{Z}_+^*$ *and any* $\lambda \in \mathbb{R}$ *we have the convergence*

$$\frac{1}{n}\Lambda_{n,k}(n\lambda) \to \Lambda(\lambda) = -\log(p)(\exp(-\lambda) - 1). \tag{4.3}$$

The Fenchel-Legendre transform Λ^* *of* Λ *satisfies:*

$$\begin{aligned} \Lambda^*(z) &= \sup_{\lambda \in \mathbb{R}} \big(\lambda z - \Lambda(\lambda)\big) \\ &= \begin{cases} +\infty \text{ if } z \geq 0 \\ z - \log(p) - z\log\big(\frac{z}{\log(p)}\big) \text{ if } z < 0. \end{cases} \end{aligned} \tag{4.4}$$

Then for any $k \in \mathbb{Z}_+^*$, the sequence of distributions $\big(\tilde{\mu}^{n,k}\big)_{n \in \mathbb{Z}_+, n>k}$ satisfies a Large Deviations Principle, with the rate function $J = \Lambda^*$.

The proof of (4.3) is the main task of this paper. In Section 5.1, we give a first easy proof in the case $k = 1$, relying on the knowledge of the distribution of $J^{n,1}$: it is a Poisson distribution with mean $-n\log(p)$. We can then compute explicitly $\Lambda_{n,1}(\lambda)$ and prove (4.3). In Section 5.2, we study the general case $k \geq 1$ with the method introduced in Bréhier et al. (2015), in the exponential case: for the algorithm $\mathrm{AMS}_{\mathrm{expo}}(n, k; a, x)$, we derive a functional equation on the Laplace transform $\exp\big(\Lambda_{n,k}(\lambda)\big)$ as a function of the initial condition x, for fixed parameter λ.

For completeness, we close this Section with the computation of the Fenchel-Legendre transform $J = \Lambda^*$ of Λ in Proposition 4.2.

Proof: First, assume that $z \geq 0$. Then $\lambda z - \Lambda(\lambda) \to +\infty$ when $\lambda \to +\infty$: thus $\Lambda^*(z) = +\infty$. This result is not surprising, since $\log(\hat{p}^{n,k}) < 0$ almost surely.

If $z < 0$, the map $\lambda \in \mathbb{R} \mapsto \lambda z - \Lambda(\lambda)$ admits the limit $-\infty$ for $z \to \pm\infty$, and attains its maximum at the unique solution λ_z of the equation $z - \frac{d\Lambda(\lambda)}{d\lambda}(\lambda_z) = 0$, which is given by $\lambda_z = -\log\big(\frac{z}{\log(p)}\big)$. Then $\Lambda^*(z) = \lambda_z z - \Lambda(\lambda_z)$, which gives (4.4). □

5. Proof of Proposition 4.2

5.1. *The case* $k = 1$. We start with a proof of Theorem 3.1 when $k = 1$: in this case, we have $C^{n,1} = 1$ almost surely, and the number of iterations $J^{n,1}$ follows a Poisson distribution $\mathcal{P}(-n\log(p))$ (see for instance Bréhier et al. (2015), Guyader et al. (2011)).

As a consequence, it is very easy to prove Proposition 4.1. Let $\lambda \in \mathbb{R}$. Then

$$\begin{aligned} \Gamma_{n,1}(\lambda) &= \exp\big(\Lambda_{n,1}(\lambda)\big) \\ &= \mathbb{E}\big[\exp\big(\lambda \log(\hat{p}^{n,1})\big)\big] \\ &= \mathbb{E}\big[\exp\big(\lambda \log(1-1/n) J^{n,1}\big)\big] \\ &= \exp\Big(-n\log(p)\big(\exp(\lambda\log(1-1/n)) - 1\big)\Big). \end{aligned}$$

It is now easy to conclude: when $n \to +\infty$

$$\frac{1}{n}\log\big(\Lambda(n\lambda)\big) = -\log(p)\big(\exp(n\lambda\log(1-1/n))-1\big) \underset{n\to+\infty}{\to} -\log(p)\big(\exp(-\lambda)-1\big).$$

We have performed explicit calculations, using the knowledge of the distribution of $J^{n,1}$. However for $k>1$, we cannot rely on such simple arguments and we need other tools.

The connexion with the Poisson distribution gives an interpretation of the rate functions I and J. More precisely, I is obtained from J by the contraction principle ($I(y) = J(\log(y))$), and J is the rate function obtained in the Cramer theorem where the distribution R is such that $-R \sim \mathcal{P}\big(-\log(p)\big)$. Indeed, let $(R_m)_{m\in\mathbb{Z}_+^*}$ be independent, with the same distribution as X; if we denote by $\overline{R}_n = \frac{1}{n}\sum_{m=1}^n R_m$ the empirical average, we compute for any $\lambda\in\mathbb{R}$

$$\mathbb{E}\big[\exp\big(n\lambda\overline{R}_n\big)\big] = \Big(\mathbb{E}\big[\exp\big(\lambda R\big)\Big)^n = \Big(\exp\big(-\log(p)\big(\exp(-\lambda)-1\big)\big)\Big)^n.$$

To conclude this section on the case $k=1$, we prove Proposition 3.4. We use again the explicit knowledge of the distribution of $J^{n,1}$ and use a Central Limit Theorem on exponential distributions to conclude.

Proof of Proposition 3.4: We write (with $a = -\log(p) = \sigma^2 n$)

$$\frac{\hat{p}^{n,1}}{p} = \exp(J^{n,1}\log(1-1/n)+a) = \exp\left(\frac{J^{n,1}-na}{\sqrt{na}}\sqrt{na}\log(1-1/n)+a+na\ln(1-1/n)\right).$$

By the Central Limit Theorem on the Poisson distribution, one gets, in the limit $n\to+\infty$, the following convergence in distribution

$$\frac{J^{n,1}-na}{\sqrt{na}} \to \mathcal{N}(0,1).$$

Moreover, when $n\to+\infty$, we have $\sqrt{na}\log(1-1/n) = n\sigma\log(1-1/n) \to -\sigma$ and $a+na\log(1-1/n) = \sigma^2\big(n+n^2\ln(1-1/n)\big)$ tends to $\frac{-\sigma^2}{2}$. This concludes the proof.

□

5.2. *The general case.* In this section, we give the main arguments used to prove Proposition 4.2 in the general case $k\in\mathbb{Z}_+^*$. In particular, we want to show that the rate function we obtain does not depend on k. The proof of some important but technical results is postponed to Section 7.

Even if in Section 5.1 above we have proved Proposition 4.2 in the case $k=1$, we include this case in our general framework, and obtain an alternative proof.

To this aim, we make use of the strategy introduced in Bréhier et al. (2015) to study the properties of the AMS$(n,k;a)$ algorithm. First, as explained in Section 2.3, we are allowed to restrict the study to the case where X is exponentially distributed: it is enough to study the AMS$_{\mathrm{expo}}(n,k;a_{\mathrm{expo}})$ algorithm, where $a_{\mathrm{expo}} = -\log(p)$.

Moreover, one of the main ideas is to consider the initial condition of the algorithm as an extra variable: for $x \in [0,a)$, we study the $\mathrm{AMS}_{\mathrm{expo}}(n,k;a_{\mathrm{expo}},x)$ algorithm. From now on, in this Section, and in Section 7, we only consider the exponential case and we omit the dependence.

Definition 5.1. We use the following notation: for any $(x,y) \in \mathbb{R}^2$

$$f(y) = \exp(-y)\mathbb{1}_{y\geq 0} \quad , \quad F(y) = \big(1-\exp(-y)\big)\mathbb{1}_{y\geq 0} = \int_{-\infty}^{y} f(z)dz;$$

$$f(y;x) = \frac{f(y)}{1-F(x)}\mathbb{1}_{y\geq x} \quad , \quad F(y;x) = \frac{F(y)-F(x)}{1-F(x)}\mathbb{1}_{y\geq x} = \int_{-\infty}^{y} f(z;x)dz;$$

$$f_{n,k}(y;x) = k\binom{n}{k}F(y;x)^{k-1}f(y;x)\big(1-F(y;x)\big)^{n-k},$$

$$F_{n,k}(y;x) = \int_x^y f_{n,k}(z;x)dz.$$

Let X be exponentially distributed with parameter 1. Then f (resp. F)) is the density (resp. the c.d.f.) of $\mathcal{L}(X)$. For $x \geq 0$, $f(\cdot;x)$ (resp. $F(\cdot;x)$) is the density (resp. the c.d.f.) of the conditional distribution $\mathcal{L}\big(X|X>x\big)$.

Finally, let $(X_1,\dots,X_n)$ be i.i.d. with the distribution of $\mathcal{L}(X)$, with the associated order statistics $X_{(1)} < \dots < X_{(n)}$. Then $f_{n,k}(\cdot;x)$ (resp. $F_{n,k}(\cdot;x)$) is the density (resp. the c.d.f.) of the k-th order statistic $X_{(k)}$.

The main object we need to study is the following function $\Gamma_{n,k}$ of $\lambda \in \mathbb{R}$ (considered as a fixed parameter) and the initial condition $x \in [0,a]$

$$\begin{aligned}\Gamma_{n,k}(\lambda;x) &= \mathbb{E}\Big[\exp\big(n\lambda\log(\hat{p}^{n,k}(x))\big)\Big] \\ &= \exp\big(\Lambda_{n,k}(n\lambda;x)\big).\end{aligned} \tag{5.1}$$

Notice that we include $x=a$ in the domain of definition of the functions $\Gamma_{n,k}$ and $\Lambda_{n,k}$ (defined by (4.2)). It is also important to remark that we evaluate $\Lambda_{n,k}$ at $(n\lambda;x)$.

We state several fundamental results which together yield Proposition 4.2 in the x-dependent case; to get (4.3) it is then enough to take $x=0$.

First, Proposition 5.2 gives a functional equation satisfied by $\Gamma_{n,k}(\lambda;\cdot)$ on $[0,a]$, for any value of the parameters $1 \leq k < n$ and $\lambda \in \mathbb{R}$.

We use the following auxiliary function:

$$\Theta_{n,k}(\lambda;x) = \sum_{\ell=0}^{k-1} \exp\big(n\lambda\log(1-\frac{\ell}{n})\big)\Big(F_{n,\ell}(a;x) - F_{n,\ell+1}(a;x)\Big), \tag{5.2}$$

with the convention $F_{n,0}(y;x) = \mathbb{1}_{y\geq x}$.

Proposition 5.2. *For any $n \in \mathbb{Z}_+^*$, $k \in \{1,\dots,n-1\}$, and $\lambda \in \mathbb{R}$, the function $\Gamma_{n,k}(\lambda;\cdot)$ is solution on the interval $[0,a]$ of the functional equation (with the unknown Γ):*

$$\Gamma(x) = \int_x^a \exp\Big(n\lambda\log(1-\frac{k}{n})\Big)\Gamma(y)f_{n,k}(y;x)dy + \Theta_{n,k}(\lambda;x). \tag{5.3}$$

Observe that for the moment, it is not clear that $\Gamma_{n,k}$ is the unique solution of the functional equation (5.3). We will prove this property below.

For completeness, we include a proof of this fundamental result, even if follows the same lines as Proposition 4.2 in Bréhier et al. (2015).

Proof of Propositon 5.2: We decompose the expectation according to the value of the (random) number of iterations $J^{n,k}(x)$ in the algorithm starting from x:

$$\begin{aligned}\Gamma_{n,k}(\lambda;x) &= \mathbb{E}\Big[\exp\big(n\lambda\log(\hat{p}^{n,k}(x))\big)\Big]\\ &= \mathbb{E}\Big[\exp\big(n\lambda\log(\hat{p}^{n,k}(x))\big)\mathbb{1}_{J^{n,k}(x)=0}\Big] + \mathbb{E}\Big[\exp\big(n\lambda\log(\hat{p}^{n,k}(x))\big)\mathbb{1}_{J^{n,k}(x)\geq 1}\Big].\end{aligned}$$

First, since $\big\{J^{n,k}(x)=0\big\} = \big\{Z^1\geq a\big\} = \bigcup_{\ell=0}^{k-1}\big\{X_{(\ell+1)}\geq a > X_{(\ell)}\big\}$, we have

$$\begin{aligned}\mathbb{E}\Big[\exp\big(n\lambda\log(\hat{p}^{n,k}(x))\big)\mathbb{1}_{J^{n,k}(x)=0}\Big] &= \mathbb{E}\Big[\exp\big(n\lambda\log(C^{n,k}(x))\big)\mathbb{1}_{J^{n,k}(x)=0}\Big]\\ &= \sum_{\ell=0}^{k-1}\exp\Big(n\lambda\log(1-\frac{\ell}{n})\Big)\Big(F_{n,\ell}(a;x)-F_{n,\ell+1}(a;x)\Big)\\ &= \Theta_{n,k}(\lambda;x).\end{aligned}$$

Second, we use $\big\{J^{n,k}(x)\geq 1\big\} = \big\{Z^1\leq a\big\}$ and condition with respect to Z^1:

$$\begin{aligned}\mathbb{E}\Big[\exp\big(n\lambda\log(\hat{p}^{n,k}(x))\big)\mathbb{1}_{J^{n,k}(x)\geq 1}\Big] &= \mathbb{E}\Big[\mathbb{E}\big[\exp\big(n\lambda\log(\hat{p}^{n,k}(x))\big)\big|Z^1\big]\mathbb{1}_{Z^1<a}\Big]\\ &= \mathbb{E}\Big[\mathbb{E}\big[\exp\big(n\lambda\log((1-k/n)^{J^{n,k}(x)-1}C^{n,k}(x)) + n\lambda\log(1-k/n)\big)\big|Z^1\big]\mathbb{1}_{Z^1<a}\Big]\\ &= \exp\Big(n\lambda\log(1-\frac{k}{n})\Big)\mathbb{E}\Big[\mathbb{E}\big[\exp\big(n\lambda\log((1-k/n)^{J^{n,k}(Z^1)}C^{n,k}(Z^1))\big)\big|Z^1\big]\mathbb{1}_{Z^1<a}\Big]\\ &= \exp\Big(n\lambda\log(1-\frac{k}{n})\Big)\mathbb{E}\Big[\Gamma_{n,k}(Z^1;x)\mathbb{1}_{Z^1<a}\Big]\\ &= \int_x^a \exp\Big(n\lambda\log(1-\frac{k}{n})\Big)\Gamma_{n,k}(\lambda;y)f_{n,k}(y;x)dy.\end{aligned}$$

We have used a kind of Markov property for the algorithm: up to taking into account for one more iteration, the algorithm behaves the same starting from x or from $Z^1\in(x,a]$. □

Notice that the functional equation (5.3) involves a simple factor depending only on λ, n and k in the integral, and that on both the left and the right-hand sides the function Γ is evaluated at the same value of the parameter λ. These observations are consequences of the choice to prove a Large Deviations Principle for $\log(\hat{p}^{n,k})$ (instead of $\hat{p}^{n,k}$) thanks to the Gärtner-Ellis Theorem, and to conclude with the use of the contraction principle; the same trick was used in Bréhier et al. (2014) to prove the Central Limit Theorem, thanks to the delta-method and the use of Levy Theorem. If one replaces $\log(\hat{p}^{n,k}(x))$ with $\hat{p}^{n,k}(x)$ in (5.1), then one obtains a more complicated functional equation where the observations above do not hold, and which is not easily exploitable. In particular, one does not obtain a nice counterpart of the fundamental result, Proposition 5.3 below.

We now state in Proposition 5.3 that solutions Γ of the functional equation (5.3) are in fact solutions of a linear Ordinary Differential Equation (ODE) of order k, with constant coefficients.

Proposition 5.3. *For any $n\in\mathbb{Z}_+^*$, $k\in\{1,\dots,n-1\}$, and $\lambda\in\mathbb{R}$, let Γ be a solution of the functional equation (5.3). Then it is solution of the following linear*

ODE of order k:

$$\frac{d^k}{dx^k}\Gamma_{n,k}(\lambda;x) = \exp\Big(n\lambda\log(1-\frac{k}{n})\Big)\mu^{n,k}\Gamma_{n,k}(\lambda;x) + \sum_{m=0}^{k-1} r_m^{n,k}\frac{d^m}{dx^m}\Gamma_{n,k}(\lambda;x). \quad (5.4)$$

The coefficients $\mu^{n,k}$ *and* $(r_m^{n,k})_{0\le m\le k-1}$ *satisfy the following properties:*

$$\begin{aligned} \mu^{n,k} &= (-1)^k n\dots(n-k+1) \\ \nu^k - \sum_{m=0}^{k-1} r_m^{n,k}\,\nu^m &= (\nu-n)\dots(\nu-n+k-1) \quad \text{for all } \nu\in\mathbb{R}. \end{aligned} \quad (5.5)$$

A sketch of proof of this result is postponed to Section 7. It uses the same arguments as to prove the corresponding functional equation in Bréhier et al. (2015). For the proof of (5.5) in particular, we refer to that article.

To conclude on uniqueness of the solution of (5.3), and then prove asymptotic expansions on $\Gamma_{n,k}$, we prove the following Lemma.

Lemma 5.4. *For any fixed* $k\in\{1,\dots,\}$ *and any* $\lambda\in\mathbb{R}$*, we have for any* $m\in\{0,\dots,k-1\}$

$$\begin{aligned} \frac{d^m}{dx^m}\Gamma_{n,k}(\lambda;x)\Big|_{x=a} &= \tfrac{d^m}{dx^m}\Theta_{n,k}(\lambda;x)\Big|_{x=a} \\ &\underset{n\to\infty}{\sim} n^m\big(1-\exp(-\lambda)\big)^m. \end{aligned} \quad (5.6)$$

By Cauchy-Lipschitz theory, the linear ODE (5.4) with the conditions (5.6) at $x=a$ admits a unique solution; therefore it is clear that $\Gamma_{n,k}$ is the unique solution of (5.3).

Remark 5.5. A similar result, although in a weaker form, is used to prove the Central Limit Theorem in Bréhier et al. (2014): it was sufficient to establish $\frac{d^m}{dx^m}\Theta_{n,k}(\lambda;x)\Big|_{x=a} = \mathrm{O}(n^m)$. Here we require a more precise asymptotic result in order to prove that the coefficient $\gamma_{n,k}^1(\lambda)$ defined in Proposition 5.6 below converges to 1 (in fact, we only need that it is bounded from below by a positive constant).

We finally explain how to obtain asymptotic knowledge on $\Gamma_{n,k}(\lambda;x)$ and $\Lambda_{n,k}(n\lambda,x)$ when $n\to+\infty$. First, the k roots $\big(\nu_{n,k}^\ell(\lambda)\big)_{1\le\ell\le k}$ of the polynomial equation associated with the linear ODE (5.4) are pairwise distinct for n large enough (the other parameters λ and k being fixed), and more precisely they satisfy (5.8). As a consequence, the solution $\Gamma_{n,k}$ can be written (see (5.7)) as a linear combination of exponential functions $x\mapsto\exp\Big(\nu_{n,k}^\ell(\lambda)\,(x-a)\Big)$. Finally, using the asymptotic expression for the derivatives of order $0,\dots,k-1$ at $x=a$, we obtain a linear system of equations, solve it using the Cramer's formulae and obtain the asymptotic expression (5.9). The proof is postponed to Section 7.

Proposition 5.6. *Let* $k\in\{1,\dots,\}$ *and* $\lambda\in\mathbb{R}$ *be fixed. Then for* n *large enough, we have for any* $x\in[0,a]$

$$\Gamma_{n,k}(\lambda,x) = \sum_{\ell=1}^{k}\gamma_{n,k}^\ell(\lambda)\exp\Big(\nu_{n,k}^\ell(\lambda)\,(x-a)\Big), \quad (5.7)$$

where

$$\nu_{n,k}^{\ell}(\lambda) \sim n\Big(1 - e^{-\lambda} e^{i2\pi \frac{(\ell-1)}{k}}\Big) \tag{5.8}$$

and

$$\gamma_{n,k}^{\ell}(\lambda) \to \mathbb{1}_{\ell=1}. \tag{5.9}$$

We now conclude and prove Proposition 4.1, namely the Large Deviations Principle for $\big(\mathcal{L}(\log(\hat{p}^{n,k}))\big)_{n>k}$.

We start with the case $k > 1$. Then for any $\ell \in \{2, \dots, k\}$ we have for any $\lambda \in \mathbb{R}$

$$\mathrm{Re}\Big(1 - e^{-\lambda} e^{i2\pi(\ell-1)/k}\Big) > \mathrm{Re}\Big(1 - e^{-\lambda}\Big).$$

As a consequence, for $x < a$ we have when $n \to +\infty$

$$e^{\nu_{n,k}^{\ell}(\lambda)(x-a)} = \mathrm{o}\Big(e^{1-\exp(-\lambda))(x-a)}\Big),$$

and thus

$$\frac{1}{n}\Lambda_{n,k}(n\lambda; x) = \frac{1}{n}\log(\Gamma_{n,k}(\lambda; x)) \underset{n\to+\infty}{\sim} \nu_{n,k}^{1}(\lambda)(x-a) \underset{n\to+\infty}{\to} (1 - e^{-\lambda})(x-a).$$

When $k = 1$, the linear ODE (5.4) is of order 1, and it is easy to check that

$$\Gamma_{n,1}(\lambda; x) = \exp\Big(\nu_{n,k}^{1}(\lambda)(x-a)\Big),$$

so that the same asymptotic result as above holds.

It remains to take $x = a$, and to recall that $a = -\log(p)$ if $p = \mathbb{P}(X > a)$ and X is exponentially distributed with parameter 1.

This concludes the proof of Proposition 4.1.

6. Comparison with other algorithms

We propose a comparison (in terms of large deviations) of the Adaptive Multilevel Splitting algorithm with the two other methods described in the introduction: a direct, naive Monte-Carlo method, based on a non-interacting system of replicas with the same size (see the estimator (1.4)) , and a non-adaptive version of multilevel splitting (see the estimator (1.2)).

In the first case, we obtain that large deviations are much less likely for the AMS algorithm than for the crude Monte-Carlo method. In the second case, we show that the AMS estimator is more efficient than the non-adaptive one taken in the limit of a large number N of fixed levels.

These results are consistent with the cost analysis and the comparison based on the central limit theorem, see Bréhier et al. (2015), Bréhier et al. (2014), Cérou et al. (2012), Cérou and Guyader (2014).

6.1. *Crude Monte-Carlo.* We compare the performance of the AMS algorithm with the use of a Crude Monte-Carlo estimation in the large n limit.

Let $(X_m)_{m\in\mathbb{Z}_+^*}$ a sequence of independent and identically distributed random variables, each one being equal in law with X.

Then for any $n \in \mathbb{Z}_+^*$

$$\overline{p}_n = \frac{1}{n}\sum_{m=1}^{n} \mathbb{1}_{X_m > a} \tag{6.1}$$

is an unbiased estimator of p.

It is a classical result (Theorem 2.2.3 in Dembo and Zeitouni (2010)) due to Cramer that the sequence $\big(\mathcal{L}(\overline{p}_n)\big)_{n\in\mathbb{Z}_+^*}$ satisfies a Large Deviations Principle with the rate function (case of Bernoulli random variables, see Exercice 2.2.23 in Dembo and Zeitouni (2010)):

$$\mathcal{I}(y) = \begin{cases} +\infty \text{ if } y \notin (0,1) \\ y\log\left(\frac{y}{p}\right) + (1-y)\log\left(\frac{1-y}{1-p}\right) \text{ if } y \in (0,1). \end{cases} \tag{6.2}$$

The comparison between the algorithms is based on the following result:

Proposition 6.1. *For any $p \in (0,1)$ and any $y \in (0,1)$, we have*

$$I(y) \geq \mathcal{I}(y),$$
$$I(y) = \mathcal{I}(y) \quad \textit{if and only if} \quad y = p.$$

Proof: We explicitly mention the dependence of I and of $\mathcal{I}$ with respect to p, and we define

$$D(y,p) = I(y,p) - \mathcal{I}(y,p).$$

It is clear that $D(p,p) = 0$ for any $p \in (0,1)$. We compute that

$$\frac{\partial D(y,p)}{\partial p} = \frac{1-y}{p\log(p)}\Big(\frac{\log(y)}{1-y} - \frac{\log(p)}{1-p}\Big);$$

since the function $t \mapsto \frac{\log(t)}{1-t}$ is strictly decreasing on $(0,1)$ (as can be seen by computing its first and second order derivatives), we see that for any $y,p \in (0,1)^2$ we have the inequalities

$$\frac{\partial D(y,p)}{\partial p} > 0 \quad \text{if} \quad y > p \quad \text{and} \quad \frac{\partial D(y,p)}{\partial p} < 0 \quad \text{if} \quad y < p.$$

Using $D(p,p) = 0$, it is easy to conclude. □

Now let $\epsilon \in (0, \max(p, 1-p))$; then for n large we have

$$\frac{\mathbb{P}(\hat{p}^{n,k} - p > \epsilon)}{\mathbb{P}(\overline{p}_n - p > \epsilon)} = \exp\Big(n\Delta(\epsilon, n)\Big) \to 0,$$

exponentially fast, since we have by the Large Deviations Principles $\Delta(\epsilon, n) \to \mathcal{I}(p+\epsilon) - I(p+\epsilon) < 0$ when $n \to +\infty$ (indeed both $\mathcal{I}$ and I are increasing on $(p,1)$).

The same arguments apply to get

$$\frac{\mathbb{P}(\hat{p}^{n,k} - p < -\epsilon)}{\mathbb{P}(\overline{p}_n - p < -\epsilon)} \to 0.$$

As a consequence, the probability of observing large deviations from the mean p is much smaller for the AMS algorithm than when using a crude Monte-Carlo estimator, in the large n limit. This statement is a new way of expressing the efficiency of the AMS algorithm.

In the discussion above we have not assumed that we are estimating a probability in a rare event regime: the conclusion holds for any $p \in (0,1)$. Now it is also

instructive to compare $I((1+\epsilon)p)$ and $\mathcal{I}((1+\epsilon)p)$ for a given $\epsilon \in (0,1)$ and when $p \to 0$: it amounts at looking at deviations of the relative error, and we have

$$\lim_{n\to+\infty} \frac{1}{n} \log\Big(\mathbb{P}(\frac{\hat{p}^{n,k}-p}{p} > \epsilon)\Big) = -I\big(p(1+\epsilon)\big) \sim_{p\to 0} -\frac{\big(\log(1+\epsilon)\big)^2}{-2\log(p)}$$
$$\lim_{n\to+\infty} \frac{1}{n} \log\Big(\mathbb{P}(\frac{\overline{p}_n-p}{p} > \epsilon)\Big) = -\mathcal{I}\big(p(1+\epsilon)\big) \sim_{p\to 0} -p\big((1+\epsilon)\log(1+\epsilon)-\epsilon\big).$$

Given $\delta > 0$, in order to have a probability lower than δ that the relative error is larger than ϵ, in the small p limit, one thus needs a number of replicas n which scales like $1/p$ when using the crude Monte-Carlo method, while it scales like $-\log(p)$ (which is much smaller) when using the AMS algorithm. Moreover, since the expected workload is of size n when using the Monte-Carlo method and of size $-n\log(p)$ when using the AMS algorithm, it is clear that in terms of large deviations from the mean the AMS algorithm is more efficient than the crude Monte-Carlo method.

This discussion is consistent with the conclusions coming from the Central Limit Theorem, where in the regime $p \to 0$ the asymptotic variance is equivalent to p when using the crude Monte-Carlo method and $-p^2\log(p)$ when using the AMS algorithm: to obtain reliable confidence intervals on the relative error, the number of replicas n scales in the same way.

6.2. *Non-adaptive Multilevel Splitting.* We now compare the rate function I obtained for the Large Deviations Principle on the AMS algorithm, with the one we obtain when using a deterministic (non-adaptive) sequence of levels.

Namely, using Assumption 2.1, we decompose the probability as a telescoping product of $N \in \mathbb{Z}_+^*$ conditional probabilities

$$p = \mathbb{P}(X > a) = \prod_{i=1}^{N} \mathbb{P}(X > a_i \big| X > a_{i-1}), \tag{6.3}$$

associated with a given non-decreasing sequence of levels $a_0 = 0 < a_1 < \ldots < a_N = a$. We denote by $p^{(i)} = \mathbb{P}(X > a_i \big| X > a_{i-1})$ the i-th conditional probability. The sequence is of size N and we study the asymptotic regime $N \to +\infty$.

We can define an unbiased estimator of p as follows: let $n \in \mathbb{Z}_+^*$ and set

$$\hat{p}_n^N = \prod_{i=1}^{N} \overline{p}_n^{(i)}, \tag{6.4}$$

where $\big(\overline{p}_n^{(i)}\big)_{1\le i\le N}$ is a family of independent random variables, where each $\overline{p}_n^{(i)}$ is a Crude Monte-Carlo estimator (as defined in the section above) for the probability $p^{(i)}$ with n realizations. More precisely, let $\big(X_m^{(i)}\big)_{1\le m\le n, 1\le i\le N}$ be independent random variables, such that $\mathcal{L}\big(X_m^{(i)}\big) = \mathcal{L}(X|X > a_{i-1})$, and set

$$\overline{p}_n^{(i)} = \frac{1}{n} \sum_{m=1}^{n} \mathbb{1}_{X_m^{(i)} > a_i}. \tag{6.5}$$

From a practical point of view, notice that the computation of these estimators requires the sampling of random variables according to the conditional distribution $\mathcal{L}(X|X > a_{i-1})$ for each $i \in \{1,\ldots,N\}$, just like for the adaptive version.

Here n thus denotes the number of replicas used for the estimation of the probabilities in both the adaptive and the non-adaptive versions. We needed the extra parameter N to denote the number of iterations (*i.e.* the length of the sequence of levels) of the algorithm, while we know that the average number of iterations is of the order $-\frac{n\log(p)}{k}$ in the adaptive case. Therefore, to study the non-adaptive version, we first let $n \to +\infty$, and then analyze the behavior of the asymptotic quantities with respect to N (in the limit $N \to +\infty$), while for the adaptive version we need to pass to the limit only once, namely $n \to +\infty$.

Clearly, by the independence properties of the random variables introduced here we have

$$\mathbb{E}[\hat{p}_n^N] = p.$$

Moreover, it is well-known that, for a given value of N (the length of the sequence of levels) the asymptotic variance (when n goes to $+\infty$) is minimized when $p^{(i)} = p^{1/N}$ for all $i \in \{1, \dots, N\}$ (i.e. the conditional probabilities in (6.3) are equal); moreover the asymptotic variance is a decreasing function of N, which converges to $\frac{-p^2\log(p)}{n}$ when $N \to +\infty$. From a practical point of view, the computation of the associated sequence of levels $a_1, \dots, a_{N-1}$ is *a priori* difficult: the adaptive version overcomes this issue, and in the regime $N \to +\infty$ both the non-adaptive and the adaptive version have the same statistical properties.

As a consequence, from now on we assume that $p^{(i)} = p^{1/N}$ for any $i \in \{1, \dots, N\}$.

For any $i \in \{1, \dots, N\}$, $\left(\mathcal{L}(\overline{p}_n^{(i)})\right)_{n\in\mathbb{Z}_+^*}$ satisfies a Large Deviations Principle with the rate function (see (6.2))

$$\mathcal{I}_N(y) = \begin{cases} +\infty \text{ if } y \notin (0,1) \\ y\log\left(\frac{y}{p^{1/N}}\right) + (1-y)\log\left(\frac{1-y}{1-p^{1/N}}\right) \text{ if } y \in (0,1). \end{cases} \tag{6.6}$$

Since for any $n \in \mathbb{Z}_+^*$ the random variables $\left(\overline{p}_n^{(i)}\right)_{1\le i\le N}$ are independent, it is easy to generalize this statement as follows. The sequence $\left(\mathcal{L}(\overline{p}_n^{(1)}, \dots, \overline{p}_n^{(N)})\right)_{n\in\mathbb{Z}_+^*}$ satisfies a Large Deviations Principle in $\mathbb{R}^N$ with the rate function (with abuse of notation $\mathcal{I}_N$ refers both to the function depending on a 1-dimensional or a N-dimensional variable)

$$\mathcal{I}_N(y_1, \dots, y_N) = \sum_{i=1}^N \mathcal{I}_N(y_i). \tag{6.7}$$

Now as a consequence of the contraction principle, since $\hat{p}_n^N = \prod_{i=1}^N \overline{p}_n^{(i)}$, the sequence $\left(\hat{p}_n^N\right)_{n\in\mathbb{Z}_+^*}$ also satisfies a Large Deviations Principle with the rate function

$$I_N(y) = \inf\left\{\mathcal{I}_N(y_N, \dots, y_N) \ ; \ y = \prod_{i=1}^N y_i\right\}. \tag{6.8}$$

On the one hand, it is clear that if $y \notin (0,1)$, then $I_N(y) = +\infty$. Indeed, for any $(y_1, \dots, y_N)$ satisfying the constraint $y = \prod_{i=1}^N y_i \notin (0,1)$, at least one of the y_i's satisfies $y_i \notin (0,1)$, which yields $\mathcal{I}_N(y_i) = \mathcal{I}_N(y_1, \dots, y_n) = +\infty$.

On the other hand, by definition of I_N, we have for any $y \in (0,1)$

$$\begin{aligned} I_N(y) &\leq \mathcal{I}_N(y^{1/N}, \dots, y^{1/N}) = N\mathcal{I}_N(y^{1/N}) \\ &= Ny^{1/N} \log\Big(\frac{y^{1/N}}{p^{1/N}}\Big) + N(1-y^{1/N}) \log\Big(\frac{1-y^{1/N}}{1-p^{1/N}}\Big) \\ &\underset{N\to\infty}{\to} \log(y) - \log(p) - \log(y) \log\Big(\frac{\log(y)}{\log(p)}\Big) = I(y). \end{aligned}$$

For our purpose, this inequality is sufficient.

We now interpret the previous inequality in terms of asymptotic estimates for deviations of $\hat{p}_n^N$ and of $\hat{p}^{n,k}$ with respect to their expected value p. Let $\epsilon > 0$, then we have by definition of the Large Deviations Principle with rate function I_N

$$\begin{aligned} \liminf_{n\to+\infty} \frac{1}{n} \log\Big(\mathbb{P}\big(|\hat{p}_n^N - p| > \epsilon\big)\Big) &\geq -\inf\left\{I_N(y)\ ; |y-p| \geq \epsilon\right\} \\ &\geq -\inf\Big\{N\mathcal{I}_N(y^{1/N})\ ;\ |y-p| \geq \epsilon\Big\} \\ &\geq -\min\Big\{N\mathcal{I}_N((p+\epsilon)^{1/N}), N\mathcal{I}_N((p-\epsilon)^{1/N})\Big\}, \end{aligned}$$

using that $\mathcal{I}_N$ is non-increasing on $(-\infty, p^{1/N})$ and non-decreasing on $(p^{1/N}, +\infty)$.

To conclude, observe that

$$\begin{aligned} \lim_{N\to+\infty} -\min\Big\{N\mathcal{I}_N((p+\epsilon)^{1/N}), N\mathcal{I}_N((p-\epsilon)^{1/N})\Big\} &= -\min\left\{I(p+\epsilon), I(p-\epsilon)\right\} \\ &= \lim_{n\to+\infty} \frac{1}{n} \log\Big(\mathbb{P}\big(|\hat{p}^{n,k} - p| > \epsilon\big)\Big). \end{aligned}$$

We can thus assess that the Adaptive Multilevel Splitting algorithm is more efficient (in a large sense) than the non-adaptive version in terms of large deviations when the number of replicas n goes to $+\infty$ and in the limit of large number N if levels.

7. Proof of the technical estimates

In this section, we give detailed proofs for the technical auxiliary results used in Section 5.2.

Proof of Proposition 5.3: We proceed by recursion, like in the proof of Proposition 6.4 in Bréhier et al. (2015) and Lemma 2 in Bréhier et al. (2014). We fix the values of $1 \leq k < n$ and of $\lambda \in \mathbb{R}$.

Differentiating recursively with respect to x, for any $0 \leq l \leq k-1$ and for any $0 \leq x \leq a$ we have (for a family of coefficients described by (7.3) below)

$$\begin{aligned} \frac{d^l}{dx^l}\left(\Gamma_{n,k}(\lambda;x) - \Theta_{n,k}(\lambda;x)\right) &= \mu_l^{n,k} \exp\Big(n\lambda \log(1-\frac{k}{n})\Big) \int_x^a \Gamma_{n,k}(\lambda;y) f_{n,k-l}(y;x) dy \\ &\quad + \sum_{m=0}^{l-1} r_{m,l}^{n,k} \frac{d^m}{dx^m}\left(\Gamma_{n,k}(\lambda;x) - \Theta_{n,k}(\lambda;x)\right), \end{aligned} \tag{7.1}$$

and that differentiating once more we get

$$\frac{d^k}{dx^k}\left(\Gamma_{n,k}(\lambda;x)-\Theta_{n,k}(\lambda;x)\right) = \mu^{n,k}\exp\Big(n\lambda\log(1-\frac{k}{n})\Big)\Gamma_{n,k}(\lambda;x) + \sum_{m=0}^{k-1} r_m^{n,k}\frac{d^m}{dx^m}\left(\Gamma_{n,k}(\lambda;x)-\Theta_{n,k}(\lambda;x)\right) \quad (7.2)$$

with $\mu^{n,k} := \mu_k^{n,k}$ and $r_m^{n,k} := r_{m,k}^{n,k}$.

The coefficients satisfy

$$\mu_0^{n,k}=1, \mu_{l+1}^{n,k}=-(n-k+l+1)\mu_l^{n,k};$$
$$\begin{cases} r_{0,l+1}^{n,k} = -(n-k+l+1)r_{0,l}^{n,k}, & \text{if } l>0,\\ r_{m,l+1}^{n,k} = r_{m-1,l}^{n,k} - (n-k+l+1)r_{m,l}^{n,k}, & 1\le m\le l,\\ r_{l,l}^{n,k} = -1. \end{cases} \quad (7.3)$$

Note that these coefficients do not depend on λ, and are the same as in Bréhier et al. (2015) and Bréhier et al. (2014). Properties (5.5) are proved in Bréhier et al. (2015).

Thanks to (5.5), for all $j\in\{0,\dots,k-1\}$ and any $x\in[0,a]$ we have

$$\frac{d^k}{dx^k}\exp\left((n-k+j+1)(x-a)\right) = \sum_{m=0}^{k-1} r_m^{n,k}\frac{d^m}{dx^m}\exp\left((n-k+j+1)(x-a)\right).$$

Using the expression of $F_{n,k}$, straightforward computations show that $\Theta_{n,k}(\lambda;\cdot)$ is a linear combination of the exponential functions $z\mapsto\exp(-nz),\dots,\exp(-(n-k+1)z)$; therefore

$$\frac{d^k}{dx^k}\Theta_{n,k}(t,x) = \sum_{m=0}^{k-1} r_m^{n,k}\frac{d^m}{dx^m}\Theta_{n,k}(t,x),$$

and thus (7.2) gives (5.4).

□

Proof of Lemma 5.4: From (7.1), the equality in (5.6) is clear.

We claim that for any $0\le m\le k-1$ and any $0\le\ell\le k-1$

$$\frac{d^m}{dx^m}\Big(F_{n,\ell}(a;x)-F_{n,\ell+1}(a;x)\Big)\Big|_{x=a} \underset{n\to\infty}{\sim} n^m\binom{m}{\ell}(-1)^\ell. \quad (7.4)$$

In particular, $\frac{d^m}{dx^m}\Big(F_{n,\ell}(a;x)-F_{n,\ell+1}(a;x)\Big)\Big|_{x=a}=0=\binom{m}{\ell}$ for n large enough as soon as $\ell>m$. Conclusion is then straightforward: using the definition (5.2) of $\Theta_{n,k}$, we get

$$\begin{aligned}\frac{1}{n^m}\frac{d^m}{dx^m}\Theta_{n,k}(\lambda;x)\big|_{x=a} &= \frac{1}{n^m}\sum_{\ell=0}^{k-1}\frac{d^m}{dx^m}\exp\big(n\lambda\log(1-\frac{\ell}{n})\big)\Big(F_{n,\ell}(a;x)-F_{n,\ell+1}(a;x)\Big)\Big|_{x=a}\\ &\underset{n\to\infty}{\to}\sum_{\ell=0}^{m}\binom{m}{\ell}(-1)^\ell\exp\big(-\ell\lambda\big)\\ &=\Big(1-\exp\big(-\lambda\big)\Big)^m.\end{aligned}$$

We now prove (7.4) by induction on m.

We first consider $m = 0$. Then for any $\ell \in \mathbb{Z}_+^*$ we have $F_{n,\ell}(a;a) = 0$ and $F_{n,0}(a;a) = 1$ (by the convention $F_{n,0}(y;x) = \mathbb{1}_{y \geq x}$), and (7.4) holds.

Let us also consider $m = 1$, when $k \geq 2$. Then $\frac{d}{dx} F_{n,0}(a;x)\big|_{x=a} = 0$, while for any $x \leq a$

$$\frac{d}{dx} F_{n,\ell}(a;x) = \frac{d}{dx} F_{n,\ell}(a-x;0) = -f_{n,\ell}(a-x;0) = -f_{n,\ell}(a;x)$$

as a consequence of the absence of memory property of the exponential distribution.

Now since $f_{n,\ell}(a,a) = n\mathbb{1}_{\ell=1}$, we get (7.4) for $m = 1$.

The induction is based on the following relations (deduced from elementary computations; for a proof see Bréhier et al. (2015), Section 6.3)

$$\left\{ \begin{array}{c} \dfrac{d}{dx} f_{n,1}(y;x) = n f_{n,1}(y;x). \\ \text{for } \ell \in \{2,\dots,n-1\},\ \dfrac{d}{dx} f_{n,\ell}(y;x) = (n-\ell+1)\big(f_{n,\ell}(y;x) - f_{n,\ell-1}(y;x)\big). \end{array} \right. \tag{7.5}$$

Thanks to the first formula in (7.5), we easily get (7.4) for $\ell = 0$ by induction on m.

If now $\ell \in \{1,\dots,k-1\}$, we have the recursive formula for $m \geq 1$

$$\begin{aligned} \frac{d^{m+1}}{dx^{m+1}} \Big(F_{n,\ell}(a;x) - F_{n,\ell+1}(a;x) \Big)\big|_{x=a} &= \frac{d^m}{dx^m} \Big(f_{n,\ell+1}(a;x) - f_{n,\ell}(a;x) \Big)\big|_{x=a} \\ &= (n-\ell) \frac{d^{m-1}}{dx^{m-1}} \Big(f_{n,\ell+1}(a;x) - f_{n,\ell}(a;x) \Big)\big|_{x=a} \\ &\quad -(n-\ell+1) \frac{d^{m-1}}{dx^{m-1}} \Big(f_{n,\ell}(a;x) - f_{n,\ell-1}(a;x) \Big)\big|_{x=a} \\ &= (n-\ell) \frac{d^m}{dx^m} \Big(F_{n,\ell}(a;x) - F_{n,\ell+1}(a;x) \Big)\big|_{x=a} \\ &\quad -(n-\ell+1) \frac{d^m}{dx^m} \Big(F_{n,\ell-1}(a;x) - F_{n,\ell}(a;x) \Big)\big|_{x=a} \end{aligned}$$

Finally using the induction hypothesis and obtain

$$\begin{aligned} \frac{1}{n^{m+1}} \frac{d^{m+1}}{dx^{m+1}} \Big(F_{n,\ell}(a;x) - F_{n,\ell+1}(a;x) \Big)\big|_{x=a} &\underset{n\to+\infty}{\to} (-1)^\ell \binom{m}{\ell} - (-1)^{\ell-1} \binom{m}{\ell-1} \\ &= (-1)^\ell \binom{m+1}{\ell}. \end{aligned}$$

This concludes the proof of Lemma 5.4.

□

Proof of Proposition 5.6: The $\nu_{n,k}^\ell(\lambda)$ are the roots of the caracteristic equation associated with the linear ODE (5.4):

$$\frac{(n-\nu)\dots(n-k+1-\nu)}{n\dots(n-k+1)} - \exp\Big(n\lambda \log(1 - \frac{k}{n}) \Big) = 0,$$

which can be rewritten as a polynomial equation of degree k with respect to the variable $\overline{\nu}_n = \frac{\nu}{n}$:

$$\frac{(1-\overline{\nu}_n)\dots(1-\frac{k-1}{n}-\overline{\nu}_n)}{1\dots(1-\frac{k-1}{n})} - \exp\Big(n\lambda \log(1 - \frac{k}{n}) \Big) = 0,$$

where $\exp\Big(n\lambda\log(1-\frac{k}{n})\Big) \underset{n\to+\infty}{\to} \exp(-k\lambda)$.

By continuity of the roots of polynomials of degree k with respect to the coefficients, we get that for all $\ell \in \{1,\dots,k\}$ (with an appropriate ordering of the roots)

$$\frac{\nu^\ell_{n,k}(\lambda)}{n} \to \overline{\nu}^\ell_{\infty,k}(\lambda)$$

where $(1-\overline{\nu}^\ell_{\infty,k}(\lambda))^k = e^{-k\lambda}$. This identity immediately yields (5.8).

As a consequence, for n large enough the roots $\nu^\ell_{n,k}(\lambda)$ are pairwise distinct. Then (5.7) holds for some complex numbers $\gamma^\ell_{n,k}(\lambda)$, where $\ell \in \{1,\dots,k\}$. Thanks to (5.7) evaluated at $x=a$, these coefficients are solution of the following linear system of equations:

$$\begin{cases} \gamma^1_{n,k}(\lambda) + \dots + \gamma^k_{n,k}(\lambda) = \Gamma_{n,k}(\lambda;x)\big|_{x=a}, \\ \gamma^1_{n,k}(\lambda)\nu^1_{n,k}(\lambda) + \dots + \gamma^k_{n,k}(\lambda)\nu^k_{n,k}(\lambda) = \frac{d}{dx}\Gamma_{n,k}(\lambda;x)\big|_{x=a}, \\ \vdots \\ \gamma^1_{n,k}(\lambda)\left(\nu^1_{n,k}(\lambda)\right)^{k-1} + \dots + \gamma^k_{n,k}(\lambda)\left(\nu^k_{n,k}(\lambda)\right)^{k-1} = \frac{d^{k-1}}{dx^{k-1}}\Gamma_{n,k}(\lambda;x)\big|_{x=a}. \end{cases} \tag{7.6}$$

This system is equivalent with

$$\begin{cases} \gamma^1_{n,k}(\lambda) + \dots + \gamma^k_{n,k}(\lambda) = \Gamma_{n,k}(\lambda;x)\big|_{x=a} \underset{n\to+\infty}{\to} 1, \\ \gamma^1_{n,k}(\lambda)\overline{\nu}^1_{n,k}(\lambda) + \dots + \gamma^k_{n,k}(\lambda)\overline{\nu}^k_{n,k}(\lambda) = \frac{1}{n}\frac{d}{dx}\Gamma_{n,k}(\lambda;x)\big|_{x=a} \underset{n\to+\infty}{\to} \overline{\nu}^1_{\infty,k}(\lambda), \\ \vdots \\ \gamma^1_{n,k}(\lambda)\overline{\nu}^1_{n,k}(\lambda)^{k-1} + \dots + \gamma^k_{n,k}(\lambda)\overline{\nu}^k_{n,k}(\lambda)^{k-1} = \frac{1}{n^{k-1}}\frac{d^{k-1}}{dx^{k-1}}\Gamma_{n,k}(\lambda;x)\big|_{x=a} \underset{n\to+\infty}{\to} \overline{\nu}^1_{\infty,k}(\lambda)^{k-1}, \end{cases} \tag{7.7}$$

thanks to (5.6) and (5.8), where $\overline{\nu}^\ell_{n,k}(\lambda) = \frac{\nu^\ell_{n,k}(\lambda)}{n} \underset{n\to+\infty}{\to} \overline{\nu}^\ell_{\infty,k}(\lambda)$.

It is now easy to get (5.9), which concludes the proof of Proposition 5.6.

□

8. Conclusion and perspectives

We have established (Theorem 3.1) a Large Deviations Principle result for the Adaptive Multilevel Splitting AMS(n,k) Algorithm in an idealized setting, when the number of replicas n goes to infinity while the parameter k and the threshold a remain fixed. The rate function does not depend on k: when $k=1$, the proof is very simple and uses an interpretation of the algorithm with a Poisson process (the number of iterations follows a Poisson distribution). When $k>1$, we rely on a functional equation technique which was already used to prove unbiasedness and asymptotic normality of the estimator in the previous works Bréhier et al. (2015) and Bréhier et al. (2014).

We were able to relate the efficiency of the algorithm with this Large Deviations result, with a comparison with two algorithms (see Section 6): a crude Monte-Carlo method and a non-adaptive version. More generally, in other situations Large Deviations could be a powerful tool to compare adaptive or non-adaptive multilevel splitting algorithms, instead of resorting only on comparison of asymptotic variances associated with central limit theorems.

Let us mention a few open directions for future works. First, it should be interesting to look at the regime where k also goes to infinity, with k/n converging to a proportion $\alpha \in (0,1)$. We expect to prove that the optimal rate function is obtained for α decreasing to 0: indeed, the asymptotic variance is minimized in this regime. A comparison with a non-adaptive version of the algorithm is expected to show that the adaptive algorithm behaves (in terms of large deviations) like the non-adaptive version when the number of replicas and of levels goes to infinity, like in the regime we have studied in this paper.

A severe restriction is given by the so-called idealized setting: we need to know how to sample according to the conditional distribution $\mathcal{L}(X|X > x)$. In practice, and especially when computing crossing probabilities for high dimensional metastable stochastic processes, it is not satisfied and the multilevel splitting algorithm needs to use an importance function to define appropriate levels, and at each step the computation of the new sample uses the one at the previous iteration (thanks to a branching procedure of the successful trajectories). A natural question is whether one can prove a Large Deviations Principle in such a framework, and study quantitatively how the rate function depends on the importance function.

Another general strategy is as follows: one can use an approximation of the conditional distribution $\mathcal{L}(X|X > x)$, based for instance on a Metropolis chain. It is expected that a Large Deviations Principle still holds for associated algorithms, but the rate function is expected to depend on the proposition kernel used in the Metropolis chain. Indeed, this observation is made in the case of the Central Limit Theorem in Cérou and Guyader (2014): the variance (which is the same both for adaptive and fixed-levels splitting algorithms) is the sum of the variance for the idealized version of the algorithm (exact sampling of conditional distribution) and of a nonnegative term which takes into account the mixing properties of the Metropolis chain. Large deviations properties of adaptive splitting algorithms in the setting of Cérou and Guyader (2014)(Metropolis scheme) or of Bréhier et al. (discrete-time dynamics), with a precise analysis of rate functions, will be investigated in future works.

When using both non-adaptive (see Garvels et al. (2002), Glasserman et al. (1998)) and adaptive (Bréhier et al.) multilevel splitting algorithms, one may observe a very large difference between the value of the estimator (averaged over a number M of independent realizations) and the true result, or between the results obtained for different choices of the importance function. Even if the estimator of the probability is unbiased, in such situations one observes an *apparent bias* toward smaller values if M is not sufficiently large. This phenomenon is explained by specificity of the models: there are several channels to reach the region B from A (in the case of the estimation of crossing probabilities between metastable states of a Markov process), which may be sampled very differently when the importance function changes. It should be interesting to investigate the relation between this phenomenon and the Large Deviations Principle for the associated estimator.

Acknowledgements

The author acknowledges support from the SNF grant 200020-149871/1, which founded his post-doctoral research at University of Neuchâtel (Switzerland), January to August 2015. He would like to thank B. Bercu and A. Richou for suggesting this work, and F. Cérou, A. Guyader and M. Rousset for helpful discussions and

advice. He also would like to thank the referee for comments which helped improve the presentation of the paper.

References

S. Asmussen and P. W. Glynn. *Stochastic simulation: algorithms and analysis*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer, New York (2007). ISBN 978-0-387-30679-7.

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

C. E. Bréhier, M. Gazeau, L. Goudenège, T. Lelièvre and M. Rousset. Unbiasedness for some generalized adaptive multilevel splitting algorithms. *preprint* (). URL http://arxiv.org/abs/1505.02674.

C. E. Bréhier, L. Goudenège and L. Tudela. Central limit theorem for adaptive multilevel splitting estimators in an idealized setting. *preprint* (2014). URL http://arxiv.org/abs/1501.01399.

C. E. Bréhier, T. Lelièvre and M. Rousset. Analysis of adaptive multilevel splitting algorithms in an idealized setting. *ESAIM Probability and Statistics* **to appear** (2015).

F. Cérou, P. Del Moral, T. Furon and A. Guyader. Sequential Monte Carlo for rare event estimation. *Stat. Comput.* **22** (3), 795–808 (2012). ISSN 0960-3174. doi:10.1007/s11222-011-9231-6. URL http://dx.doi.org/10.1007/s11222-011-9231-6.

F. Cérou and A. Guyader. Adaptive multilevel splitting for rare event analysis. *Stoch. Anal. Appl.* **25** (2), 417–443 (2007). ISSN 0736-2994. doi:10.1080/07362990601139628. URL http://dx.doi.org/10.1080/07362990601139628.

F. Cérou and A. Guyader. Fluctuations of adaptive multilevel splitting. *preprint* (2014).

A. Dembo and O. Zeitouni. *Large deviations techniques and applications. 2nd ed., corrected 2nd printing.* Berlin: Springer, 2nd ed., corrected 2nd printing edition (2010). ISBN 978-3-642-03310-0/pbk; 978-3-642-03311-7/ebook. doi:10.1007/978-3-642-03311-7.

M. J. J. Garvels, D. P. Kroese and J. C. W. van Ommeren. On the importance function in splitting simulation. *European Transactions on Telecom- munications* **13** (4), 363–371 (2002).

P. Glasserman, P. Heidelberger, P. Shahabuddin and T. Zajic. A large deviations perspective on the efficiency of multilevel splitting. *IEEE Trans. Automat. Control* **43** (12), 1666–1679 (1998). ISSN 0018-9286. doi:10.1109/9.736061. URL http://dx.doi.org/10.1109/9.736061.

P. Glasserman, P. Heidelberger, P. Shahabuddin and T. Zajic. Multilevel splitting for estimating rare event probabilities. *Oper. Res.* **47** (4), 585–600 (1999). ISSN 0030-364X. doi:10.1287/opre.47.4.585. URL http://dx.doi.org/10.1287/opre.47.4.585.

A. Guyader, N. Hengartner and E. Matzner-Løber. Simulation and estimation of extreme quantiles and extreme probabilities. *Appl. Math. Optim.* **64** (2), 171–196 (2011). ISSN 0095-4616. doi:10.1007/s00245-011-9135-z. URL http://dx.doi.org/10.1007/s00245-011-9135-z.

H. Kahn and T. E. Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards* **12**, 27–30 (1951).

G. Rubino and B. Tuffin. Introduction to rare event simulation. In *Rare event simulation using Monte Carlo methods*, pages 1–13. Wiley, Chichester (2009). doi:10.1002/9780470745403.ch1. URL http://dx.doi.org/10.1002/9780470745403.ch1.

E. Simonnet. Combinatorial analysis of the adaptive last particle method. *Statistics and Computing* pages 1–20 (2014). ISSN 0960-3174. doi:10.1007/s11222-014-9489-6. URL http://dx.doi.org/10.1007/s11222-014-9489-6.

J. Skilling. Nested sampling for general Bayesian computation. *Bayesian Anal.* **1** (4), 833–859 (electronic) (2006). ISSN 1931-6690. doi:10.1214/06-BA127. URL http://dx.doi.org/10.1214/06-BA127.

J. Skilling. Nested sampling for Bayesian computations. In *Bayesian statistics 8*, Oxford Sci. Publ., pages 491–524. Oxford Univ. Press, Oxford (2007).

M. Villén-Altamirano and J. Villén-Altamirano. Restart: A method for accelerating rare events simulations. In *Proceeding of the thirteenth International Teletraffic Congress*, volume Copenhagen, Denmark, June 19-26 of *Queueing, performance and control in ATM: ITC-13 workshops*, pages 71–76. North-Holland, Amsterdam-New York (1991).

M. Villén-Altamirano and J. Villén-Altamirano. Restart: a straightforward method for fast simulation of rare events. In *Proceedings of the 1994 Winter Simulation Conference*, volume Orlando 1994, December 1994, pages 282–289 (1994).

C. Walter. Moving particles: a parallel optimal multilevel splitting method with applications in quantiles estimation and meta-model-based algorithms. *preprint* (2014).