A framework for adaptive Monte-Carlo procedures

Bernard Lapeyre[*] Jérôme Lelong[†]

July 9, 2010

Abstract

Adaptive Monte Carlo methods are recent variance reduction techniques. In this work, we propose a mathematical setting which greatly relaxes the assumptions needed by for the adaptive importance sampling techniques presented in [24, 23, 1, 2]. We establish the convergence and asymptotic normality of the adaptive Monte Carlo estimator under local assumptions which are easily verifiable in practice. We present one way of approximating the optimal importance sampling parameter using a randomly truncated stochastic algorithm. Finally, we apply this technique to some examples of valuation of financial derivatives.

1 Introduction

Monte-Carlo methods aim at computing the expectation $\mathbb{E}(Z)$ of a real-valued random variable Z using samples along the law of Z. In this work, we focus on cases where there exists a parametric representation of the expectation

$$\mathbb{E}(Z) = \mathbb{E}\left(H(\theta, X)\right) \quad \text{for all } \theta \in \mathbb{R}^d, \tag{1}$$

where X is a random vector with values in $\mathbb{R}^m$ and $H : \mathbb{R}^d \times \mathbb{R}^m \longmapsto \mathbb{R}$ is a measurable function satisfying $\mathbb{E}|H(\theta, X)| < \infty$ for all $\theta \in \mathbb{R}^d$. We also impose that

$$\theta \longmapsto v(\theta) = \mathrm{Var}(H(\theta, X)) \text{ is finite for all } \theta \in \mathbb{R}^d, \tag{2}$$

We want to make the most of this free parameter θ to settle an automatic variance reduction method, see [8] for a recent survey on adaptive variance reduction. It consists in first finding a minimiser $\theta^\star$ of the variance v and then in plugging it into a Monte Carlo method with a narrower confidence interval. This technique heavily relies on the ability to find a parametric representation and to effectively minimise the function v. Many papers have been written on how to construct parametric representations $H(\theta, X)$ for several kinds of random variables Z. We mainly have in mind examples based on control variates (see [4, 13, 12]) or importance sampling (see [24, 23, 1, 2]). We refer the reader to section 4 for a presentation of a few examples.

[*] Université Paris-Est, CERMICS, Projet MathFi ENPC-INRIA-UMLV, 6 et 8 avenue Blaise Pascal, 77455 Marne La Vallée, Cedex 2, France , e-mail : bernard.lapeyre@enpc.fr.

[†] Laboratoire Jean Kuntzmann, Université de Grenoble et CNRS, BP 53, 38041 Grenoble Cédex 9, France, e-mail : jerome.lelong@imag.fr

Assume we have a parametric representation of the form $H(\theta, X)$ satisfying Equations (1) and (2). Let $(X_n)_n$ be an independent and identically distributed sequence of random vectors following the law of X. Assume we know how to use the sequence $(X_n)_n$ to build an estimator θ_n of $\theta^\star$ adapted to the filtration $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$. Once such an approximation is available, there are at least two ways of using it to devise a variance reduction method.

The non-adaptive algorithm

Algorithm 1.1 (Non adaptive importance sampling (NADIS))**.** *Let n be the number of samples used for the Monte Carlo computation. Draw a second set of n samples $(X'_1, \dots, X'_n)$ independent of $(X_1, \dots, X_n)$ and compute*

$$\bar{\xi}_n = \frac{1}{n} \sum_{i=1}^{n} H(\theta_n, X'_i).$$

Since the sequence $(\theta_n)_n$ converges to θ^*, the convergence of $(\bar{\xi}_n)_n$ to $\mathbb{E}(Z)$ ensues from the strong law of large numbers and the sequence $(\bar{\xi}_n)_n$ satisfies a central limit theorem

$$\sqrt{n}(\bar{\xi}_n - \mathbb{E}(Z)) \xrightarrow[n \to \infty]{law} \mathcal{N}\left(0, v(\theta^\star)\right).$$

This algorithm has been studied in [23, 2] and required $2n$ samples. It may use less than $2n$ samples if the estimation of $\theta^\star$ is performed on a smaller number of samples but then it raises the question of how many samples to use.

The adaptive algorithm The adaptive approach is to use the *same samples* $(X_1, \dots, X_n)$ to compute θ_n and the Monte Carlo estimator. Compared to the sequential algorithm, the adaptive one uses half of the samples.

Algorithm 1.2 (Adaptive Importance Sampling (ADIS))**.** *Let n be the number of samples used for the Monte Carlo computation.*
For θ_0 fixed in $\mathbb{R}^d$, compute

$$\xi_n = \frac{1}{n} \sum_{i=1}^{n} H(\theta_{i-1}, X_i). \tag{3}$$

Note that the sequence $(\xi_i)_i$ can be written in a recursive manner so that it can be updated online each time a new iterate θ_i is drawn

$$\xi_{i+1} = \frac{i}{i+1}\xi_i + \frac{1}{i+1} H(\theta_i, X_{i+1}), \quad \text{with } \xi_0 = 0.$$

Being able to update the sequence $(\xi_i)_i$ online has the advantage that there is no need to store the whole sequence $(X_1, \dots, X_n)$ for computing ξ_n. This adaptive algorithm was first studied in [1] in which the author studied the convergence of the sequence $(\xi_n)_n$ under assumptions to be verified along the path $(\theta_n)_n$ which makes them hard to check in practise. In this article, we prove a new convergence result under local integrability conditions on the function H, namely we impose that for any compact subset K of $\mathbb{R}^n$, $\sup_{\theta \in K} \mathbb{E}(|H(\theta, X)|^2) < \infty$. We refer the reader to section 2.1 for a precise statement and proof of these results. We want to emphasize that such assumptions only involving properties of the function H and not of the sequence $(\theta_n)_n$ are far easier to check in practice.

Sofar, we have assumed that we knew how to devise a convergent estimator of $\theta^\star$, but this may not be so simple as when no closed form expression is available for $\mathbb{E}(H(\theta, X))$, there is hardly no chance that the function v can be computed explicitly. Henceforth, it is needed to approximate $\theta^\star$ without being able to compute the variance itself. In this work, we recall the methodology based on stochastic approximation developed in [23, 24, 2] to estimate $\theta^\star$ using some stochastic gradient style algorithms. We aim at applying this methodology to the evaluation of financial derivatives and the main difficulty in approximating $\theta^\star$ comes from the non-boundedness of the payoff functions usually considered and consequently the non-boundedness of the H functions. To encompass this problem, several authors as in [24, 23] restrict the parameter θ to lie in a compact set, which is obviously unknown in practice; therefore, this compact set will have to be quite large. Although, it permits to prove the theoretical convergence of the Robbins-Monro algorithm it does not help to build a numerically convergent estimator of $\theta^\star$. We all know that the true convergence of stochastic algorithms highly relies on the fine tuning of the gain sequence which reveals to be very difficult when dealing with an artificially bounded parameter set.

In this work following [2], we would rather use a randomly truncated algorithm which is known to converge for a much wider class of functions. We give a unified framework with easily verifiable assumptions under which Algorithm 1.2 converges and satisfies a central limit theorem. Then, we combine this convergence result with the new results on randomly truncated stochastic algorithm from [16] to revisit the adaptive algorithm in the Gaussian framework studied in [1].

The paper is organised as follows. In Section 2, we focus on the mathematical foundation of the method and give both a strong law of large numbers and a central limit theorem for the adaptive estimator under weak assumptions. In Section 3, we present one way of constructing a convergent estimator of $\theta^\star$ and recall some recent results on stochastic approximation. Then, we give in Section 4 some examples of how to construct a parametric estimator using importance sampling or other more elaborate transformations. Finally, we illustrate the convergence results obtained in Section 2 on numerical examples coming from financial problems.

2 Mathematical foundations of the method

NOTATIONS:

- We encode any elements of $\mathbb{R}^m$ as column vectors.
- If $x \in \mathbb{R}^m$, x^* is a row vector. We use the "*" notation to denote the transpose operator for vectors and matrices.
- If $x, y \in \mathbb{R}^m$, $x \cdot y$ denotes the Euclidean scalar product of x and y and the associated norm is denoted by $|\cdot|$.

In this section, $(X_n)_{n\geq 1}$ is an i.i.d. sequence following the law of X and we introduce the $\sigma-$algebra $\mathcal{F}_n$ it generates $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$. For technical reasons, we assume that the variance v does not vanish, i.e. $\inf_{\theta\in\mathbb{R}^n} v(\theta) > 0$. If such is not the case, it means that we are actually in a better situation as far as variance reduction is concerned but it does not fit in our framework.

2.1 An adaptive strong law of large numbers

Theorem 2.1 (Adaptive strong law of large numbers)**.** *Assume Equation* (1) *and* (2) *hold. Let $(\theta_n)_{n\geq 0}$ be a $(\mathcal{F}_n)-$adapted sequence with values in $\mathbb{R}^d$ such that for all $n \geq 0$, $\theta_n < \infty$ a.s and for any compact subset $K \subset \mathbb{R}^d$, $\sup_{\theta\in K} \mathbb{E}(|H(\theta, X)|^2) < \infty$. If*

$$\inf_{\theta\in\mathbb{R}^d} v(\theta) > 0 \qquad \textit{and} \qquad \frac{1}{n}\sum_{k=0}^{n} v(\theta_k) < \infty \quad \textit{a.s.}, \tag{4}$$

then ξ_n converges a.s. to $\mathbb{E}(Z)$.

Proof. For any $p \geq 0$, we define $\tau_p = \inf\{k \geq 0; |\theta_k| \geq p\}$. The sequence $(\tau_p)_p$ is an increasing sequence of $(\mathcal{F}_n)-$stopping times such that $\lim_{p\to\infty} \tau_p \uparrow \infty$ a.s.. Let $M_n = \sum_{i=0}^{n-1} H(\theta_i, X_{i+1}) - \mathbb{E}(Z)$. We introduce $M_n^{\tau_p} = M_{\tau_p\wedge n}$ defined by

$$M_n^{\tau_p} = \sum_{i=0}^{n-1\wedge\tau_p} H(\theta_i, X_{i+1}) - \mathbb{E}(Z) = \sum_{i=0}^{n-1}(H(\theta_i, X_{i+1}) - \mathbb{E}(Z))\mathbf{1}_{\{i\leq\tau_p\}}.$$

$\mathbb{E}(|H(\theta_i, X_{i+1}) - \mathbb{E}(Z)|^2\mathbf{1}_{\{i\leq\tau_p\}}) \leq \mathbb{E}(\mathbf{1}_{\{i\leq\tau_p\}}\mathbb{E}(|H(\theta, X) - \mathbb{E}(Z)|^2)_{\theta=\theta_i})$. On the set $\{i \leq \tau_p\}$, the conditional expectation is bounded from above by $\sup_{|\theta|\leq p} v(\theta)$. Hence, the sequence $(M_n^{\tau_p})_n$ is square integrable and it is obvious that $(M_n^{\tau_p})_n$ is a martingale, which means that the sequence $(M_n)_n$ is a locally square integrable martingale (i.e. a local martingale which is locally square integrable).

$$\langle M\rangle_n = \sum_{i=0}^{n-1} \mathbb{E}((H(\theta_i, X_{i+1}) - \mathbb{E}(Z))^2|\mathcal{F}_i) = \sum_{i=0}^{n-1} v(\theta_i).$$

By Condition (4), we have a.s. $\limsup_n \frac{1}{n}\langle M\rangle_n < \infty$ and $\liminf_n \frac{1}{n}\langle M\rangle_n > 0$. Applying the strong law of large numbers for locally $\mathbb{L}^2$ martingales (see [18]) yields the result. □

The sequence $(\theta_n)_n$ can be any sequence adapted to $(X_n)_{n\geq 1}$ convergent or not. For instance, $(\theta_n)_n$ can be an ergodic Markov chain distributed around the minimizer $\theta^\star$ such as Monte Carlo Markov Chain algorithms.

Remark 2.2. *When the sequence $(\theta_n)_{n\geq 0}$ converges a.s. to a deterministic constant θ_∞, it is sufficient to assume that v is continuous at θ_∞ and $v(\theta_\infty) > 0$ to ensure that Condition* (4) *is satisfied. Note that there is no need to impose that $\theta_\infty = \theta^\star$ although it is undoubtedly wished in practice. For instance, θ_∞ can be an approximation of $\theta^\star$ obtained either by heuristic arguments such as large deviations.*

2.2 A Central limit theorem for the adaptive strong law of large numbers

To derive a central limit theorem for the adaptive estimator ξ_n, we need a central limit theorem for locally square integrable martingales, whose convergence rate has been extensively studied. We refer to the works of Rebolledo [21], Jacod and Shiryaev [7], Hall and

Heyde [6] and Whitt [25] to find different statements of central limit theorems for locally square integrable càdlàg martingales in continuous time, from which theorems can easily be deduced for discrete time locally square integrable martingales.

Theorem 2.3. *Assume Equation* (1) *and* (2) *hold. Let* $(\theta_n)_{n\geq 0}$ *be a* $\mathcal{F}_n-$*adapted sequence with values in* $\mathbb{R}^d$ *such that for all* $n \geq 0$, $\theta_n < \infty$ *a.s and converging to some deterministic value* θ_∞. *Assume there exists* $\eta > 0$ *such that the function* $s_{2+\eta} : \theta \in \mathbb{R}^d \longmapsto \mathbb{E}\left(|H(\theta, X)|^{2+\eta}\right)$ *is finite for all* $\theta \in \mathbb{R}^d$ *and continuous at* θ_∞. *Moreover, if* v *is continuous at* θ_∞ *and* $v(\theta_\infty) > 0$, *then,* $\sqrt{n}(\xi_n - \mathbb{E}(Z)) \xrightarrow{law} \mathcal{N}(0, v(\theta_\infty))$.

Proof. We know from the proof of Theorem 2.1 that $M_n = \sum_{i=0}^{n-1} H(\theta_i, X_{i+1}) - \mathbb{E}(Z)$ is a locally square integrable martingale and that $\frac{1}{n}\langle M\rangle_n$ converges a.s. to $v(\theta_\infty)$.

$$\frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}(|H(\theta_i, X_{i+1}) - \mathbb{E}(Z)|^{2+\eta}|\mathcal{F}_i) \leq c\left(\frac{1}{n}\sum_{i=0}^{n-1} s_{2+\eta}(\theta_i) + \mathbb{E}(Z)^{2+\eta}\right).$$

The term on the r.h.s is bounded thanks to the continuity of $s_{2+\eta}$ at θ_∞. Hence, the local martingale $(M_n)_n$ satisfies Lindeberg's condition. The result ensues from the central limit theorem for locally $\mathbb{L}^2$ martingales. □

Corollary 2.4 (Effective central limit theorem with confidence interval)**.** *Assume Equation* (1) *and* (2) *hold. Let* $(\theta_n)_{n\geq 0}$ *be a* $\mathcal{F}_n-$*adapted sequence with values in* $\mathbb{R}^d$ *such that for all* $n \geq 0$, $\theta_n < \infty$ *a.s and converging to some deterministic value* θ_∞. *Assume there exists* $\eta > 0$ *such that the function* $s_{4+\eta} : \theta \in \mathbb{R}^d \longmapsto \mathbb{E}\left(|H(\theta, X)|^{4+\eta}\right)$ *is finite for all* $\theta \in \mathbb{R}^d$ *and continuous at* θ_∞. *Then,* $\sigma_n^2 = \frac{1}{n}\sum_{i=0}^{n-1} H(\theta_i, X_{i+1})^2 - \xi_n^2 \xrightarrow{a.s.} v(\theta_\infty)$. *If moreover* $v(\theta_\infty) > 0$, *then* $\frac{\sqrt{n}}{\sigma_n}(\xi_n - \mathbb{E}(Z)) \xrightarrow[n\to+\infty]{law} \mathcal{N}(0, 1)$.

Remark 2.5. *Even if* $v(\theta_\infty) > 0$, σ_n *may take negative values for n small. This corollary is really essential from a practical point of view because it proves that confidence intervals can be built as in the case of a crude Monte Carlo procedure. The only difference lies in the way of approximating the asymptotic variance.*

The assumptions of Theorem 2.3 are fairly easy to check in practice since they are formulated independently of the sequence $(\theta_n)_n$. When $\theta_\infty = \theta^\star$, which is nonetheless not required, the limiting variance is optimal in the sense that a crude Monte Carlo computation with the optimal parameter $\theta^\star$ would have lead to the same limiting variance. These assumptions are satisfied in the frameworks introduced in Section 4.

3 Estimation of the optimal variance parameter

From Theorem 2.1 and Theorem 2.3, we know that if we can construct a convergent estimator $(\theta_n)_n$ of $\theta^\star$, the adaptive estimator ξ_n is a convergent and asymptotically normal estimator of the expectation $\mathbb{E}(Z)$. The challenging issue is now to propose an automatic way of approximating the minimiser $\theta^\star$ of $v(\theta) = \mathbb{E}(H(\theta, X)^2) - \mathbb{E}(Z)^2$. In the following, we will assume that v is strictly convex, goes to infinity at infinity and is continuously differentiable. Moreover, we assume that ∇v admits a representation as an expectation

$$\nabla v(\theta) = \mathbb{E}(U(\theta, X)),$$

where $U : \mathbb{R}^d \times \mathbb{R}^m \longmapsto \mathbb{R}^d$ is a measurable and integrable function. We could see in the examples developed in Section 4 that these conditions are very easily satisfied. Stochastic algorithms such as the Robbins Monro algorithm (see [22]) are perfectly well suited to estimate quantities defined as the root of an expectation. Because for the applications we are targeting we cannot impose that $\mathbb{E}(|U(\theta, X)|^2) < c(1 + |\theta|^2)$, the Robbins-Monro algorithm will fail to converge and we need a more robust algorithm. This will naturally lead us to consider randomly truncated stochastic algorithms as introduced by Chen et al. [3]. When dealing with stochastic approximations, the idea of averaging the iterates comes out quite naturally to smooth the trajectories, see Section 3.2.

3.1 Randomly truncated stochastic algorithms

Let $(X_n)_{n\geq 1}$ be an i.i.d sequence of random variables following the law of X and $(\gamma_n)_{n\geq 1}$ be a decreasing sequence of a positive real numbers satisfying

$$\sum_n \gamma_n = \infty \quad \text{and} \quad \sum_n \gamma_n^2 < \infty. \tag{5}$$

The sequence $(\gamma_n)_n$ is often called the gain sequence or the step sequence. We define the $\sigma-$field $\mathcal{F}_n = \sigma(X_k,\ k \leq n)$. We introduce an increasing sequence of compact sets $(K_j)_j$ of $\mathbb{R}^d$

$$\bigcup_{n=0}^{\infty} K_n \ = \ \mathbb{R}^d \quad \text{and} \quad K_n \subsetneq \overset{\circ}{K}_{n+1} \tag{6}$$

Now, we can present the randomly truncated stochastic algorithm introduced in [3], which essentially consists in a truncation of the Robbins Monro algorithm on an increasing sequence of compact sets. For $\theta_0 \in K_0$ and $\alpha_0 = 0$, we define the sequences of random variables $(\theta_n)_n$ and $(\alpha_n)_n$ by

$$\begin{cases} & \theta_{n+\frac{1}{2}} = \theta_n - \gamma_{n+1} U(\theta_n, X_{n+1}), \\ \text{if } \theta_{n+\frac{1}{2}} \in \mathcal{K}_{\alpha_n} & \theta_{n+1} = \theta_{n+\frac{1}{2}} \quad \text{and} \quad \alpha_{n+1} = \alpha_n, \\ \text{if } \theta_{n+\frac{1}{2}} \notin \mathcal{K}_{\alpha_n} & \theta_{n+1} = \theta_0 \quad \text{and} \quad \alpha_{n+1} = \alpha_n + 1. \end{cases} \tag{7}$$

$\theta_{n+\frac{1}{2}}$ is the new sample we draw, either we accept it and set $\theta_{n+1} = \theta_{n+\frac{1}{2}}$ or we reject it and reset the algorithm to θ_0 when it tries to jump too far ahead in a single step. Note that $\theta_{n+\frac{1}{2}}$ is actually drawn along the dynamics of the Robbins Monro algorithm and either we accept it as the new iterate or we reject it when the algorithm tries to jump to far ahead and in this case we reset the new iterate to θ_0. In the following, we write Equation (7) in a more condensed form

$$\theta_{n+1} = \mathcal{T}_{K_{\alpha_n}} \left(\theta_n - \gamma_{n+1} U(\theta_n, X_{n+1}) \right), \tag{8}$$

where $\mathcal{T}_{K_{\alpha_n}}$ denotes the truncation on the compact sets K_{α_n}.

The use of truncations enables to relax the hypotheses required to ensure the convergence. From the recent results of [16], we can state the following convergence result

Theorem 3.1. *Assume that*

(A1) *∇v is continuous and there exists a unique $\theta^\star$ s.t. $\nabla v(\theta^\star) = 0$ and $\forall\ \theta \neq \theta^\star$, $(\nabla v(\theta) \,|\, \theta - \theta^\star) > 0$.*

(A2) *$\forall q > 0$, $\sup_{|\theta| \leq q} \mathbb{E}(|U(\theta, Z)|^2) < \infty$.*

Then, the sequence $(\theta_n)_n$ defined by (7) converges a.s. to $\theta^\star$ for any sequence of compact sets satisfying (6) and moreover the sequence $(\alpha_n)_n$ is a.s. finite.

Note that the assumptions required to ensure the convergence are very weak and are formulated independently of the algorithm trajectories, which makes them easy to check. Since the variance reduction technique we settle here aims at being automatic in the sense that it does not require any fiddling with the gain sequence depending on the function U, it is quite natural to average the procedure defined by Equation (7).

3.2 Averaging a stochastic algorithm

This section is based on the remark that Cesaro type averages tend to smooth the behaviour of convergent estimators at least from a theoretical point of view. Such averaging techniques have already been studied and proved to provide asymptotically efficient estimators (see for instance [20], [14] or [19]).

At the same time, it is well known that true Cesaro averages are not so efficient from a practical point of view because the rate at which the impact of the first iterates vanishes in the average is too slow and it induces some kind of a numerical bias which in turn dramatically slows down the convergence. Combining these two facts has led us to consider a moving window average of Algorithm (7).
In this section, we restrict to gain sequences of the form $\gamma_n = \frac{\gamma}{(n+1)^a}$ with $\frac{1}{2} < a < 1$. Let $\tau > 0$ be the length of the window used for averaging. For $n \geq 1$, we introduce

$$\hat{\theta}_n(\tau) = \frac{\gamma_p}{\tau} \sum_{i=p}^{p+\lfloor \tau/\gamma_p \rfloor} \theta_i \quad \text{with } p = \sup\{k \geq 1 : k + \tau/\gamma_k \leq n\} \wedge n. \tag{9}$$

We use the convention $\sup \emptyset = +\infty$. The almost sure convergence of $(\hat{\theta}_n(\tau))_n$ can easily be deduced from Theorem 3.1. The asymptotic normality of the sequence $(\hat{\theta}_n(\tau))_n$ has been studied in [15]. The definition of $\hat{\theta}_n$ is a little different from the one used in [15] because we want to ensure that the sequence $(\hat{\theta}_n)_n$ is adapted to the filtration $\mathcal{F}_n$ in view of the use of $\hat{\theta}_n$ as an estimator of $\theta^\star$ in Algorithm 1.2.

4 Examples of parametric Monte-Carlo settings

In this section, we give various examples of cases in which a parametric representation of the expectation of interest is available

$$\mathbb{E}(Z) = \mathbb{E}(H(\theta, X)).$$

In each example, we highlight the strong convexity and the regularity of the function $\theta \longmapsto \mathbb{E}(H^2(\theta, X))$ such that the minimiser $\theta^\star$ is uniquely defined as the one root of $\theta \longmapsto \nabla_\theta \mathbb{E}(H^2(\theta, X))$.

4.1 Importance sampling for normal random variables

Let $G = (G_1, \dots, G_d)$ be a $d-$dimensional standard normal random vector. For any measurable function $h : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that $\mathbb{E}(|h(G)|) < \infty$, one has for all $\theta \in \mathbb{R}^d$

$$\mathbb{E}\left(h(G)\right) = \mathbb{E}\left(e^{-\theta\cdot G - \frac{|\theta|^2}{2}} h(G+\theta)\right). \tag{10}$$

Assume we want to compute $\mathbb{E}(f(G))$ for a measurable function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that $f(G)$ is integrable. By applying equality (10) to $h = f$ and $h(x) = f^2(x)\,\mathrm{e}^{-\theta\cdot x + \frac{|\theta|^2}{2}}$, one obtains that the expectation and the variance of the random variable $f(G+\theta)\,\mathrm{e}^{-\theta\cdot G\ -\frac{|\theta|^2}{2}}$ are respectively equal to $\mathbb{E}(f(G))$ and $v(\theta) - \mathbb{E}^2(f(G))$ where

$$v(\theta) = \mathbb{E}\left(f^2(G)\,\mathrm{e}^{-\theta\cdot G + \frac{|\theta|^2}{2}}\right).$$

The strict convexity of the function v is already known from [23] for instance. For the sake of completeness, we prove here a slightly improved version of this result.

Proposition 4.1. *Assume that*

$$\mathbb{P}(f(G) \neq 0) > 0, \tag{11}$$
$$\exists \varepsilon > 0,\ \mathbb{E}(|f(G)|^{2+\varepsilon}) < \infty \tag{12}$$

Then, v is infinitely continuously differentiable and strongly convex.

Proof. The function $\theta \mapsto f^2(G)e^{-\theta\cdot G + \frac{|\theta|^2}{2}}$ is infinitely continuously differentiable. Since,

$$\sup_{|\theta|\le M} |\partial_{\theta^j} f^2(G)e^{-\theta\cdot G+\frac{|\theta|^2}{2}}| \le e^{\frac{M^2}{2}} f^2(G)\left(M + (e^{G^j} + e^{-G^j})\right)\prod_{k=1}^d (e^{MG^k} + e^{-MG^k})$$

where the right hand side is integrable because by Hölder's inequality and Equation (12), we have $\forall\theta \in \mathbb{R}^d, \mathbb{E}\left(f^2(G)\,\mathrm{e}^{-\theta\cdot G}\right) < \infty$. Lebesgue's theorem ensures that v is continuously differentiable with $\frac{\partial}{\partial_{\theta^j}} v(\theta) = \mathbb{E}\left(f^2(G)(\theta^j - G^j)e^{-\theta\cdot G+\frac{|\theta|^2}{2}}\right)$. Higher order differentiability properties are obtained by similar arguments and in particular the Hessian matrix writes

$$\nabla^2 v(\theta) = \mathbb{E}\left(f^2(G)e^{-\theta\cdot G+\frac{|\theta|^2}{2}})\right) I + \mathbb{E}\left((\theta - G)(\theta - G)^* f^2(G)e^{-\theta\cdot G+\frac{|\theta|^2}{2}}\right)$$

The second term in the above equation is a positive semi-definite matrix, hence

$$\nabla^2 v(\theta) \ge \mathbb{E}(f^2(G)e^{-\theta\cdot G+\frac{|\theta|^2}{2}}) = \mathbb{E}(f^2(G)e^{-\theta\cdot G})\mathbb{E}(e^{\theta\cdot G}) \ge \mathbb{E}(|f(G)|)^2.$$

Assumption (11) ensures that $\mathbb{E}(|f(G)|) > 0$. Then, the Hessian matrix is uniformly bounded from below by the positive definite matrix $\mathbb{E}(|f(G)|)^2 I$. This yields the strong convexity of the function v. □

Proposition 4.1 implies that v has a unique minimiser $\theta^\star$ characterised by $\nabla v(\theta^\star) = 0$, i.e. $\mathbb{E}\left((\theta^\star - G)e^{-\theta^\star\cdot G+\frac{|\theta^\star|^2}{2}} f^2(G)\right) = 0$.

4.2 Importance sampling for processes

Equality (10) can actually be extended to the Brownian motion framework using Girsanov's theorem. Let $(W_t, 0 \leq t \leq T)$ be a $d-$dimensional Brownian motion and $\mathcal{F}$ its natural filtration. For any measurable and $\mathcal{F}-$predictable process $(\theta_t, 0 \leq t \leq T)$ such that $\mathbb{E}\left(e^{\frac{1}{2}\int_0^T |\theta_t|^2 dt}\right) < \infty$, one has

$$\mathbb{E}\left(f(W_t, 0 \leq t \leq T)\right) = \mathbb{E}\left(e^{-\int_0^T \theta_t \cdot dW_t - \frac{1}{2}\int_0^T |\theta_t|^2 dt} f\left(W_t + \int_0^t \theta_s ds, 0 \leq t \leq T\right)\right).$$

Assume $\theta_t = \theta \in \mathbb{R}^d$, for all $t \in [0, T]$. The variance of $e^{-\theta \cdot W_T - \frac{\theta^2 T}{2}} f(W_t, 0 \leq t \leq T)$ writes down $v(\theta) - \mathbb{E}\left(f^2(W_t, 0 \leq t \leq T)\right)$ with

$$v(\theta) = \mathbb{E}\left(e^{-\theta \cdot W_T + \frac{|\theta|^2}{2}T} f^2\left(W_t + \theta t,\, 0 \leq t \leq T\right)\right).$$

A similar result to Proposition 4.1 holds; in particular v is infinitely continuously differentiable, strictly convex and goes to infinity at infinity.

For more general processes $(\theta_t, 0 \leq t \leq T)$, we refer the reader to [17].

4.3 The exponential change of measure

The idea of tilting some probability measure to find the ones that minimises the variance is a very common idea which can be also be applied to a wide range of distribution, see for instance the recent results of Kawai [11, 10] in which he applied an exponential change of measure to Lévy processes, also known as the Esscher transform.

Consider a random variable X with values in $\mathbb{R}^d$ and cumulative generating function $\psi(\theta) = \log \mathbb{E}\left(e^{\theta \cdot X}\right)$. We assume that $\psi(\theta) < \infty$ for all $\theta \in \mathbb{R}^d$. Let p denote the density of X. We define the density p_θ by

$$p_\theta(x) = p(x)\, e^{\theta \cdot x - \psi(\theta)}, \quad x \in \mathbb{R}^d.$$

Let $X^{(\theta)}$ have p_θ as a density, then

$$\mathbb{E}(f(X)) = \mathbb{E}\left[f(X^{(\theta)})\frac{p(X^{(\theta)})}{p_\theta(X^{(\theta)})}\right].$$

The variance of $f(X^{(\theta)})\frac{p(X^{(\theta)})}{p_\theta(X^{(\theta)})}$ writes $v(\theta) - \mathbb{E}\left(f(X^{(\theta)})^2 \frac{p(X^{(\theta)})^2}{p_\theta(X^{(\theta)})^2}\right)$ with

$$v(\theta) = \mathbb{E}\left(\left|f(X^{(\theta)})\frac{p(X^{(\theta)})}{p_\theta(X^{(\theta)})}\right|^2\right) = \mathbb{E}\left(f(X)^2\, e^{-\theta \cdot X + \psi(\theta)}\right).$$

Obviously, this change of measure is only valuable as a variance reduction technique if $X^{(\theta)}$ can be simulated at approximately the same cost as X.

Proposition 4.2. *Assume that*

$$\exists \varepsilon > 0,\ \mathbb{E}(|f(G)|^{2+\varepsilon}) < \infty \tag{13}$$

$$\lim_{|\theta| \longrightarrow \infty} p_\theta(x) = 0 \textit{ for all } x \textit{ in } \mathbb{R}^d \tag{14}$$

Then, v is infinitely continuously differentiable, convex and $\lim_{|\theta| \longrightarrow \infty} v(\theta) = \infty$.

Proof. To prove the differentiability of v, it suffices to reproduce the first part of the proof of Proposition 4.1. The convexity of v comes from the log-convexity of ψ. Moreover,

$$v(\theta) = \mathbb{E}\left(f(X)^2 \, \mathrm{e}^{-\theta\cdot X+\psi(\theta)}\right) = \mathbb{E}\left(f(X)^2 \frac{p(X)}{p_\theta(X)}\right)$$

Combining Equation (14) with Fatou's Lemma yields that $\lim_{|\theta|\longrightarrow\infty} v(\theta) = \infty$. □

Remark 4.3. *If X is a random standard normal vector, $p_\theta(x) = p(x-\theta)$ and $X^{(\theta)}$ is a random normal vector with mean θ and identity covariance matrix. Hence, we recover Equality* (10)*.*

5 Application to the Gaussian random vector framework

5.1 Presentation of the problem

We consider a $D-$multidimensional local volatility model in which each asset is supposed to be driven by the following dynamics under the risk neutral measure.

$$dS_t^i = S_t^i(rdt + \sigma(t, S_t^i)\cdot dW_t^i), \; S_0^i = s^i.$$

$W = (W^1, \dots, W^D)^*$ is a vector of correlated standard Brownian motions. The covariance structure of the Brownian motions is given by $\langle W, W\rangle_t = \Gamma t$ where Γ is a definite positive matrix with a diagonal filled with ones. In our numerical examples, we take $\Gamma_{ij} = \mathbf{1}_{\{i=j\}} + \rho\mathbf{1}_{\{i\neq j\}}$ with $\rho \in (\frac{-1}{D-1}, 1)$ to ensure that the matrix Γ is positive definite. The function σ is the local volatility function, r is the instantaneous interest rate and the vector $(s^1, \dots, s^D)$ is the vector of the spot values. In this model, we want to price path-dependent options whose payoffs can be written as a function of $(S_t, t \leq T)$. Hence, the price is given by the discounted expectation $\mathrm{e}^{-rT}\,\mathbb{E}(\psi(S_t, t\leq T))$. Most of the time, this expectation must be computed by Monte Carlo methods and one has to consider an approximation of $\psi(S_t, t\leq T)$ on a time grid $0 = t_0 < t_1 < \dots < t_N = T$. Then, the quantity of interest becomes

$$\mathrm{e}^{-rT}\,\mathbb{E}(\hat{\psi}(S_{t_0}, S_{t_1}, \dots, S_{t_N})).$$

The discretisation of the asset S can for instance be obtained using an Euler scheme, which means that the function $\hat{\psi}$ can be expressed in terms of the Brownian increments or equivalently using a random normal vector. These remarks finally turn the original pricing problem into the computation of an expectation of the form $\mathbb{E}(\phi(G))$ where G is a standard normal random vector in $\mathbb{R}^{ND}$ and $\phi : \mathbb{R}^{N\times D} \longmapsto \mathbb{R}$ is a measurable and integrable function. Using Equation (10), we have for all $\theta \in \mathbb{R}^d$,

$$\mathbb{E}(\phi(G)) = \mathbb{E}\left(\phi(G + A\theta)e^{-A\theta\cdot G - \frac{|A\theta|^2}{2}}\right), \tag{15}$$

where A is $d \times ND$ matrix. The particular choice $d = ND$ and $A = I_d$ corresponds to Equation (10). When $D = 1$, the choice $d = 1$ and $A = (\sqrt{t_1}, \sqrt{t_2 - t_1}, \dots, \sqrt{t_N - t_{N-1}})^*$ corresponds to adding a linear drift to the one dimensional standard Brownian motion W and we recover the Cameron-Martin formula.

Transformation (15) actually relies on an importance sampling change of measure. Other strategies may be applicable such as stratification for instance as it is explained by Glasserman et al. in [5].

5.2 Bespoke estimators for the optimal variance parameter

It ensues from Proposition 4.1, that the second moment

$$v(\theta) = \mathbb{E}\left(\psi(G + A\theta)^2 e^{-2A\theta\cdot G - |A\theta|^2}\right) = \mathbb{E}\left(\phi(G)^2 e^{-A\theta\cdot G + \frac{|A\theta|^2}{2}}\right) \tag{16}$$

is strongly convex, infinitely differentiable and

$$\nabla v(\theta) = \mathbb{E}\left(A^*(A\theta - G)\phi(G)^2 e^{-A\theta\cdot G + \frac{|A\theta|^2}{2}}\right). \tag{17}$$

If we apply Equation (10) again, we obtain an other expression for

$$\nabla v(\theta) = \mathbb{E}\left(-A^* G\phi(G + A\theta)^2 e^{-2A\theta\cdot G + |A\theta|^2}\right). \tag{18}$$

Let us introduce the following two functions

$$U^1(\theta, G) = A^*(A\theta - G)\phi(G)^2 e^{-A\theta\cdot G + \frac{|A\theta|^2}{2}}, \tag{19}$$
$$U^2(\theta, G) = -A^* G\phi(G + A\theta)^2 e^{-2A\theta\cdot G + |A\theta|^2}. \tag{20}$$

Using either Equation (17) or (18), we can write $\nabla v(\theta) = \mathbb{E}(U^2(\theta, G)) = \mathbb{E}(U^1(\theta, G))$ and these two functions U^1 and U^2 fit in the framework of Section 3 and enable to construct two estimators of $\theta^\star$ $(\theta^1_n)_n$ and $(\theta^2_n)_n$ following Equation (7)

$$\theta^1_{n+1} = \mathcal{T}_{K_{\alpha_n}}\left(\theta^1_n - \gamma_{n+1} U^1(\theta^1_n, G_{n+1})\right), \tag{21}$$
$$\theta^2_{n+1} = \mathcal{T}_{K_{\alpha_n}}\left(\theta^1_n - \gamma_{n+1} U^2(\theta^2_n, G_{n+1})\right), \tag{22}$$

where G_n is an i.i.d sequence of random variables following the law of G. We also introduce their corresponding averaging versions $(\widehat{\theta^1}_n)_n$ and $(\widehat{\theta^2}_n)_n$ following Equation (9). Based on Equation (15), we define

$$H(\theta, G) = \phi(G + A\theta)e^{-A\theta\cdot G - \frac{|A\theta|^2}{2}}.$$

Corresponding to the different estimators of $\theta^\star$ listed above, we can define as many approximations of $\mathbb{E}(\phi(G))$ following Equation (3)

$$\xi^1_n = \frac{1}{n}\sum_{i=1}^n H(\theta^1_{i-1}, G_i), \quad \xi^2_n = \frac{1}{n}\sum_{i=1}^n H(\theta^2_{i-1}, G_i),$$
$$\hat{\xi}^1_n = \frac{1}{n}\sum_{i=1}^n H(\hat{\theta}^1_{i-1}, G_i), \quad \hat{\xi}^2_n = \frac{1}{n}\sum_{i=1}^n H(\hat{\theta}^2_{i-1}, G_i),$$

where the sequence G_i has already been used to build the $(\theta_n)_n$ estimators. From Proposition 4.1 and Theorems 3.1, 2.1 and 2.3, we can deduce the following result.

Theorem 5.1. *If there exists $\varepsilon > 0$ such that $\mathbb{E}(\phi(G)^{4+\varepsilon}) < \infty$ then, the sequences $(\theta^1_n)_n$, $(\theta^2_n)_n$, $(\widehat{\theta^1}_n)_n$ and $(\widehat{\theta^2}_n)_n$ defined by Equations (7) or (9) converge a.s. to $\theta^\star$ for any increasing sequence of compact sets $(K_j)_j$ satisfying (6) and the adaptive estimator $(\xi^1_n)_n, (\xi^2_n)_n, (\hat{\xi}^1_n)_n, (\hat{\xi}^2_n)_n$ converge to $\mathbb{E}(\phi(G))$ and are asymptotically normal with optimal limiting variance $v(\theta^\star)$.*

Proof. We only do the proof for $(\theta_n^1)_n$ and $(\widehat{\theta^1}_n)_n$ as the same ideas can be applied to $(\theta_n^2)_n$ and $(\widehat{\theta^2}_n)_n$. We know from Proposition 4.1, that the function v defined by Equation (16) is strongly convex and infinitely differentiable, hence ∇v satisfies Assumption (A1). Let $q > 0$. For any θ satisfying $|\theta| \leq q$, we have

$$\mathbb{E}\left|U^1(\theta, G)\right|^2 \leq \mathbb{E}\left((\|A\|(\|A\|q + |G|)^2 \phi(G)^4 e^{-2A\theta\cdot G} e^{\|A\|^2 q^2}\right).$$

Using Hölder's inequality, it can easily be proved that the expectation on the right hand side is uniformly bounded for $|\theta| \leq q$. Hence Assumption (A2) is satisfied. Therefore, θ_n^1 and $\hat{\theta}_n^1$ both converge to $\theta^\star$. Let $\eta > 0$,

$$\mathbb{E}\left|H(\theta, G)\right|^{2+\eta} = \mathbb{E}\left(|\phi(G)|^{2+\eta} e^{-(1+\eta)A\theta\cdot G} e^{\frac{|A\theta|^2(1+\eta)}{2}}\right).$$

Using the integrability of $\phi(G)$ and Hölder's inequality, one can prove that the expectation on the .r.h.s is bounded for θ in a ball. Moreover, combining this with Lebesgue's theorem, we obtain that the functions $\theta \longmapsto \mathbb{E}|H(\theta, G)|^2$ and $\theta \longmapsto \mathbb{E}|H(\theta, G)|^{2+\eta}$ are continuous. Therefore, the convergence and asymptotic normality of ξ_n issues from Theorems (2.1) and (2.3). □

Remark 5.2. *Theorem 5.1 extends the result of [2, Theorem 4]. Our result is valid for any increasing sequences of compact sets $(K_j)_j$ satisfying (6) whereas Arouna needed a condition on the compact sets to ensure the convergence of the $(\theta_n)_n$ estimators. The only condition required is some integrability on the payoff function and nothing has to be checked along the algorithm paths, which is a great improvement from a practical point of view.*

For the vast majority of payoff functions commonly used, the assumptions of Theorem 5.1 are always satisfied.

5.3 Numerical results

5.3.1 Complexity of the different approximations

In the introduction, we have presented two different strategies for implementing a variance reduction method based on the approximation of the optimal variance parameter. We know from Theorem 5.1, that the adaptive and non-adaptive algorithms both converges at the same rate and the same limiting variance. Therefore, to decide which one is the better, we have to compare their computational costs. In this section, we assume that the computational cost of the different algorithms is determined by the number of evaluations of the function ϕ. We will see in the examples later that this assumption is realistic and therefore it becomes obvious that the averaging and non-averaging estimators of $\theta^\star$ all have the same computational costs when implemented with expertise.

The non-adaptive algorithm We know from [2, 23] that the sequential algorithm converges with a rate of $\sqrt{v(\theta^\star)/n}$ if we have $2n$ samples at hand and want to implement the sequential algorithm by using the first n samples for approximating $\theta^\star$ and the last n samples for actually computing the Monte Carlo estimator with the previously computed approximation of $\theta^\star$. Whatever approximation of $\theta^\star$ is used, be it $(\theta_n^1)_n$, $(\theta_n^2)_n$, $(\widehat{\theta^1}_n)_n$ or

$(\widehat{\theta^2}_n)_n$, this algorithm requires $2n$ evaluations of the function ϕ whereas a crude Monte Carlo method only evaluates the function ϕ n times and achieves a convergence rate of $\sqrt{v(0)/n}$, hence this method only becomes efficient when $v(\theta^\star) \leq v(0)/2$.

The adaptive algorithm From Theorem 5.1, we know that the adaptive estimators $(\xi_n^1)_n, (\xi_n^2)_n, (\hat{\xi}_n^1)_n, (\hat{\xi}_n^2)_n$ all converge with the same rate $\sqrt{v(\theta^\star)/n}$ but as we will see it they do not have the same computational cost. First, let us concentrate on $(\xi_n^1)_n$ and $(\hat{\xi}_n^1)_n$, at each iteration i, the function ϕ has to be computed twice : once at the point $G_{i+1} + \theta_i^1$ (or $G_{i+1} + \hat{\theta}_i^1$) to update the Monte Carlo estimator and once at the point G_{i+1} to update θ_{i+1}^1 or $\hat{\theta}_{i+1}^1$. Hence, the computation of ξ_n^1 or $\hat{\xi}_n^1$ requires $2n$ evaluations of the function ϕ. Similarly, the computation of $\hat{\xi}_n^2$ requires at each step 2 evaluations of the function ϕ : one at the point $G_{i+1} + \hat{\theta}_i^2$ to update the Monte Carlo estimator and one at the point $G_{i+1} + \theta_i^2$ to update the stochastic algorithm. So the overall cost is still $2n$ evaluations of the function ϕ. But looking closely at the computation of $(\xi_n^2)_n$ immediately highlights the benefit of having put the parameter θ back into the function ϕ in the expression of ∇v : the updates of ξ_{i+1}^2 and θ_{i+1}^2 both use the evaluation of the function ϕ at the same point $G_{i+1} + \theta_i^2$. Hence, the computation of ξ_n^2 only needs n evaluations of the function ϕ instead of $2n$ for all the others algorithms. Obviously, the computational costs of the different estimators cannot really be reduced to the number of times the function ϕ is evaluated so one should not expect that computing ξ_n^2 is twice less costly than the other estimators but we will see in the examples below that the estimator ξ_n^2 is indeed faster than the others.

To shortly conclude on the complexity of the different algorithms, be they sequential or adaptive, one should bear in mind that all the estimators except $(\xi_n^2)_n$ roughly require twice the computational time of the crude Monte-Carlo method.

5.3.2 Practical implementation

The choice of using Equations (7) or (9) to build an estimator of $\theta^\star$ becomes really important when one has to implement the variance reduction procedure either by using Algorithms (1.1) or (1.2). Both the averaging and non-averaging strategies have pros and cons. The averaging algorithm theoretically converges a little slower but has a much smoother behaviour with respect to the proper adjustment of the gain sequence $(\gamma_n)_n$. Then, to have a robust estimator — in the sense that the numerical convergence of the estimator does not depend too much on the choice of of the gain sequence — the averaging procedure proves to be better in practice. The non averaging algorithm should converge a little faster even though we do not notice it in practise as the convergence oscillates too much and is far more sensitive to the proper choice of the sequence $(\gamma_n)_n$. Eventually, both algorithms produce very similar results regarding variance reduction; the averaging one is easier to tune but requires more computational time.

In the numerical experiments of this section, we compare the different algorithms on multi-asset options. The quantity "Var MC" denotes the variance of the crude Monte Carlo estimator computed on-line on a single run of the algorithm. The variance denoted "Var ξ^2" (resp. "Var $\hat{\xi}^2$") is the variance of the ADIS algorithm (see Algorithm 1.2) which uses $(\theta_n^2)_n$ (resp. $(\hat{\theta}_n^2)_n$) to estimate $\theta^\star$. These variances are computed using the on-line estimator given by Corollary 2.4. These adaptive algorithms are also compared to the sequential strategy described by Algorithm 1.1 denoted by "θ^2+MC" or "$\hat{\theta}^2$+MC" depending on how $\theta^\star$ is approximated. In all these algorithms, the matrix A introduced in Equation (15) is chosen as the identity matrix. When A is not the identity matrix, its purpose is to reduce the dimension of the space in which the optimal $\theta^\star$ is searched and in such cases the algorithms will be call "reduced". Note that for the comparison to be fair between the different strategies, we have used n samples for the adaptive algorithms but $2n$ for the sequential algorithms so that they all satisfy a central limit theorem with the rate $\sqrt{n}$.

Basket options We consider options with payoffs of the form $(\sum_{i=1}^d \omega^i S_T^i - K)_+$ where $(\omega^1, \ldots, \omega^d)$ is a vector of algebraic weights (enabling us to consider exchange options).

ρ	K	γ	Price	Var MC	Var ξ^2	Var $\hat{\xi}^2$
0.1	45	1	7.21	12.24	1.59	1.10
	55	10	0.56	1.83	0.19	0.14
0.2	50	0.1	3.29	13.53	1.82	1.76
0.5	45	0.1	7.65	43.25	6.25	4.97
	55	0.1	1.90	14.74	1.91	1.4
0.9	45	0.1	8.24	69.47	10.20	7.78
	55	0.1	2.82	30.87	2.7	2.6

Table 1: Basket option in dimension $d = 40$ with $r = 0.05$, $T = 1$, $S_0^i = 50$, $\sigma^i = 0.2$, $\omega^i = \frac{1}{d}$ for all $i = 1, \ldots, d$ and $n = 100\,000$.

Estimators	MC	ξ^2	$\hat{\xi}^2$
CPU time	0.85	0.9	1.64

Table 2: CPU times for the option of Table 1.

The results of Table 1 indicate that the adaptive algorithm using an averaging stochastic approximation outperforms not only the crude Monte Carlo approach but also the adapted algorithms using non-averaging stochastic approximation. The better performance of the algorithms using averaging estimators of $\theta^\star$ comes from the better smoothness of the averaging algorithm (see Equation (9)). Nonetheless, these good results in terms of variance reduction must be considered together with their computation costs reported in Table 2. As explained in Section 5.3.1, we notice that the computational cost of the estimator ξ^2 is very close to the one of the crude Monte Carlo estimator because the implementation made the most of the fact that the updates of ξ_{i+1}^2 and θ_{i+1}^2 both need to evaluate the function ϕ at the same point. Note that, because this implementation trick cannot be applied to $\hat{\xi}^2$, the adaptive algorithm using an averaging stochastic approximation is twice slower. For a given precision, the adaptive algorithm is between 5 and 10 times faster.

Barrier Basket Options We consider basket options in dimension D with a discrete barrier on each asset. For instance, if we consider a Down and Out Call option, the payoff writes down $(\sum_{i=1}^{D} \omega^i S_T^i - K)_+ \mathbf{1}_{\{\forall i \leq D,\, \forall j \leq N,\, S_{t_j}^i \geq L^i\}}$ where $\omega = (\omega^1, \dots, \omega^D)$ is a vector of positive weights, $L = (L^1, \dots, L^D)$ is the vector of barriers, $K > 0$ the strike value and $t_N = T$. We consider one time step per month, which means that for an option with maturity time $T = 2$, the number of time steps is $N = 24$. From now on, we fix $D = 5$. Hence if we use the identity matrix A, the parameter θ is of size $N \times D = 120$. Here, we propose to reduce the dimension of θ and we will in Table 3 that it achieves almost the same variance reduction. Of course the matrix A cannot be chosen independently of the structure of the problem. Remember that the vector G actually corresponds to the increments of the Brownian motion B with values in $\mathbb{R}^d$ on the grid $(t_k = kT/N, k = 0, \dots, N)$. We recall that we can simulate the Brownian motion B on the time grid $(t_k)_k$ by using the following equality in law

$$\begin{pmatrix} B_{t_1} \\ B_{t_2} \\ \vdots \\ B_{t_{N-1}} \\ B_{t_N} \end{pmatrix} = \begin{pmatrix} \sqrt{t_1} I_D & 0 & 0 & \dots & 0 \\ \sqrt{t_1} I_D & \sqrt{t_2 - t_1} I_D & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sqrt{t_{N-1} - t_{N-2}} I_D & 0 \\ \sqrt{t_1} & \sqrt{t_2 - t_1} I_D & \dots & \sqrt{t_{N-1} - t_{N-2}} I_D & \sqrt{t_N - t_{N-1}} I_D \end{pmatrix} G$$

where I_D is the identity matrix in dimension D. If we choose

$$A = \begin{pmatrix} \sqrt{t_1} I_D \\ \sqrt{t_2 - t_1} I_D \\ \vdots \\ \vdots \\ \sqrt{t_N - t_{N-1}} I_D \end{pmatrix}$$

then the transformation $G + A\theta$ corresponds to the transformation $(B_{t_1} + \theta t_1, B_{t_2} + \theta t_2, \dots, B_{t_N} + \theta t_N)^*$ and it reduces the effective dimension of the importance sampling parameter to $D = 5$ rather than $DN = 120$.

K	γ	Price	Var MC	Var ξ^2	Var $\hat{\xi}^2$	Var $\hat{\theta}^2$+MC	Var ξ^2 reduced	Var $\hat{\xi}^2$ reduced	Var $\hat{\theta}^2$+MC reduced
45	0.5	2.37	22.46	4.92	3.52	2.59	2.64	2.62	2.60
50	1	1.18	10.97	1.51	1.30	0.79	0.80	0.80	0.79
55	1	0.52	4.85	0.39	0.38	0.19	0.24	0.23	0.19

Table 3: Down and Out Call option in dimension $I = 5$ with $\sigma = 0.2$, $S_0 = (50, 40, 60, 30, 20)$, $L = (40, 30, 45, 20, 10)$, $\rho = 0.3$, $r = 0.05$, $T = 2$, $\omega = (0.2, 0.2, 0.2, 0.2, 0.2)$ and $n = 100\,000$.

First, we note from Table 3 that the reduced and non-reduced ADIS algorithm achieve almost the same variance reduction. Actually, it is even advisable to reduce the size of

Estimators	MC	ξ^2	$\hat{\xi}^2$	θ^2 + MC	ξ^2	$\hat{\xi}^2$	θ^2 + MC
					reduced	reduced	reduced
CPU time	1.86	1.93	3.34	4.06	1.89	2.89	3.90

Table 4: CPU times for the option of Table 3.

the importance sampling parameter to reduce the noise in the stochastic approximation and therefore in the adaptive Monte Carlo estimator. Comparing the columns "Var MC", "Var ξ^2" and "Var $\hat{\xi}^2$" points out that when the convergence of the estimator of $\theta^\star$ is too slow the first iterates of the adaptive Monte Carlo estimators use wrong values of θ and therefore cannot reach $v(\theta^\star)$ whereas if a sequential algorithm is used all the iterates of the Monte Carlo estimator use the same and better approximation of θ. We can see in Table 3 that the variance of "$\hat{\xi}^2$+MC" is half the one of ξ^2 or $\hat{\xi}^2$ but its CPU time is twice the one of ξ^2 as noted in Table 4.

The reduced algorithms are a little faster than the non-reduced ones but their real advantage is to converge much more stably and to achieve the same variance as "$\hat{\xi}^2$+MC" but in far less computational time. As in the previous examples, we still observe that the estimator "ξ^2 reduced" is faster than the others and has a variance very close to the best method which is "$\hat{\theta}^2$+MC".

6 Conclusion

In this work, we have explained how one could devise an adaptive variance reduction method for computing an expectation with a free parameter. Different algorithms have been studied both from a theoretical point of view and in practice. Although all the adaptive algorithms satisfy the same central limit theorem, they may behave very differently in practice, in particular adaptive algorithms using a non-averaging stochastic approximation of the optimal variance parameter can be implemented in a clever way which makes them as fast as a crude Monte Carlo approach. Nevertheless, the numerical convergence of these stochastic is very sensitive to the tuning of their gain sequence and one way to smooth this behaviour is to plug an averaging procedure on top of the stochastic approximation but then the computational time significantly increases and yet the dependency with respect to the gain sequence is still a serious drawback. To encounter the fine tuning of the algorithm, Jourdain and Lelong [9] have recently suggested to use deterministic optimisation techniques coupled with sample approximation, but their technique can not be implemented in an adaptive manner which increases its computational cost.

References

[1] B. Arouna. Adaptative Monte Carlo method, a variance reduction technique. *Monte Carlo Methods Appl.*, 10(1):1–24, 2004.

[2] B. Arouna. Robbins-Monro algorithms and variance reduction in finance. *The Journal of Computational Finance*, 7(2), Winter 2003/2004.

[3] H.F. Chen and Y.M. Zhu. *Stochastic Approximation Procedure with randomly varying truncations*. Scientia Sinica Series, 1986.

[4] Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 2004. , Stochastic Modelling and Applied Probability.

[5] Paul Glasserman, Philip Heidelberger, and Perwez Shahabuddin. Asymptotically optimal importance sampling and stratification for pricing path-dependent options. *Math. Finance*, 9(2):117–152, 1999.

[6] P. Hall and C. C. Heyde. *Martingale limit theory and its application*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1980. Probability and Mathematical Statistics.

[7] J. Jacod and A.N. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer-Verlag Berlin, 1987.

[8] B. Jourdain. *Advanced Financial Modelling*, chapter Adaptive variance reduction techniques in finance, pages 205–222. Radon Series Comp. Appl. Math 8. Walter de Gruyter, 2009.

[9] B. Jourdain and J. Lelong. Robust adaptive importance sampling for normal random vectors. *Annals of Applied Probability*, 19(5):1687–1718, 2009.

[10] Reiichiro Kawai. Adaptive monte carlo variance reducion for Lévy processes with two-time-scale stochastic approximation. *Methodology and Computing in Applied Probability*, 10(2):199–223, 2008.

[11] Reiichiro Kawai. Optimal importance sampling parameter search for Lévy processes via stochastic approximation. *SIAM Journal on Numerical Analysis*, 47(1):293–307, 2010.

[12] S. Kim and S. G. Henderson. Adaptive control variates. In *Proceedings of the 2004 Winter Simulation Conference*, 2004.

[13] Sujin Kim and Shane G. Henderson. Adaptive control variates for finite-horizon simulation. *Math. Oper. Res.*, 32(3):508–527, 2007.

[14] Harold J. Kushner and Jichuan Yang. Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes. *SIAM Journal on Control and Optimization*, 31(4):1045–1062, 1993.

[15] J. Lelong. *Etude asymptotique des algorithmes stochastiques et calcul du prix des options Parisiennes*. PhD thesis, Ecole Nationale des Ponts et Chasusées, http://tel.archives-ouvertes.fr/tel-00201373/fr/, 2007.

[16] J. Lelong. Almost sure convergence of randomly truncated stochastic algorithms under verifiable conditions. *Statistics & Probability Letters*, 78(16), 2008.

[17] V. Lemaire and G. Pagès. Unconstrained recursive importance sampling. *Annals of Applied Probability (to appear)*, 2009.

[18] D. Lépingle. Sur le comportement asymptotique des martingales locales. In *Séminaire de Probabilités, XII (Univ. Strasbourg, Strasbourg, 1976/1977)*, volume 649 of *Lecture Notes in Math.*, pages 148–161. Springer, Berlin, 1978.

[19] Mariane Pelletier. Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM J. Control Optim.*, 39(1):49–72 (electronic), 2000.

[20] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

[21] Rolando Rebolledo. Central limit theorems for local martingales. *Z. Wahrsch. Verw. Gebiete*, 51(3):269–286, 1980.

[22] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

[23] Yi Su and Michael C. Fu. Optimal importance sampling in securities pricing. *Journal of Computational Finance*, 5(4):27–50, 2002.

[24] Felisa J. Vázquez-Abad and Daniel Dufresne. Accelerated simulation for pricing asian options. In *WSC '98: Proceedings of the 30th conference on Winter simulation*, pages 1493–1500, Los Alamitos, CA, USA, 1998. IEEE Computer Society Press.

[25] Ward Whitt. Proofs of the martingale FCLT. *Probab. Surv.*, 4:268–302, 2007.