Coupling Importance Sampling and Multilevel Monte Carlo using Sample Average Approximation

Ahmed Kebaier* & Jérôme Lelong†

February 19, 2016

Abstract

In this work, we propose a smart idea to couple importance sampling and Multilevel Monte Carlo. We advocate a per level approach with as many importance sampling parameters as the number of levels, which enables us to compute the different levels independently. The search for parameters is carried out using sample average approximation, which basically consists in applying deterministic optimisation techniques to Monte Carlo approximation rather than resorting to stochastic approximation. Our innovative estimator leads to a robust and efficient procedure reducing both the bias and the variance for a given computational effort. In the setting of discretized diffusions, we prove that our estimator satisfies a strong law of large numbers and a central limit theorem with optimal limiting variance. Finally, we illustrate the efficiency of our method on several numerical challenges coming from quantitative finance.

AMS 2000 Mathematics Subject Classification. 60F05, 62F12, 65C05, 60H35.
Key Words and Phrases. Sample average approximation, Multilevel Monte Carlo, variance reduction, Uniform strong large law of numbers, Central limit theorem, Importance Sampling.

1 Introduction

Many probabilistic problems boil down to the computation of expected values involving a stochastic process, which are often computed by Monte Carlo methods. For instance, computing an hedging portfolio in finance uses these tools. Generally, the asset price follows a diffusion process $(X_t)_{0\leq t\leq T} \in \mathbb{R}^d$ with a non explicit solution, the simulation of which requires a discretization scheme $(X_t^n)_{0\leq t\leq T}$ with $n \in \mathbb{N}^*$ time steps such as the Euler scheme, the Milstein scheme or some other well known higher order schemes (see Kloeden and Platen [28] for an extensive discussion). The error induced by such schemes is called the discretization error or the bias. Then, the valuation of a financial derivative using a Monte Carlo method involves the simulation of N independent samples of X_T^n. These methods are known to converge slowly. In particular, for a given discretization error of order $1/n^\alpha$, for $\alpha > 0$, the optimal

*Université Paris 13, Sorbonne Paris Cité, LAGA, CNRS (UMR 7539), kebaier@math.univ-paris13.fr. This research benefited from the support of the chair Risques Financiers, Fondation du Risque and the Laboratory of Excellence MME-DII (http://labex-mme-dii.u-cergy.fr/).

†Univ. Grenoble Alpes, Laboratoire Jean Kuntzmann, 51, rue des Mathématiques, BP 53, 38041 Grenoble, Cedex 09, France jerome.lelong@imag.fr. This project was supported by the Finance for Energy Market Research Centre, www.fime-lab.org.

choice for the number of samples is given by $N = n^{2\alpha}$. This leads to an overall complexity for the Monte Carlo method of order $n^{3\alpha}$. Nevertheless, a lot of techniques have been developed in the recent years to speed up the method. Kebaier [27] proposed the Statistical Romberg method to generate discretization schemes on two different time grids, using a coarser grid to simulate a crude approximation and a finer one to tune the bias. More recently, Giles [16] generalized the statistical Romberg method of [27] and proposed the multilevel Monte Carlo algorithm in a similar approach to Heinrich's multilevel method for parametric integration [22]. It turns out that for the Euler scheme with a given discretization error of order $1/n^{\alpha}$, $\alpha > 0$, and for a Lipschitz continuous payoff function, the optimal complexity of the Statistical Romberg and the multilevel Monte Carlo methods are respectively of order $n^{2\alpha+1/2}$ and $n^{2\alpha}(\log n)^2$, which are clearly better than a crude Monte Carlo method. We refer the reader to the extensive literature linked to Multilevel Monte Carlo for more details, see Ben Alaya and Kebaier [6], Collier, Haji-Ali, Nobile, von Schwerin and Tempone [12], Creutzig, Dereich, Muller-Gronbach and Ritter [13], Dereich [14], Giles [17], Giles, Higham and Mao [19], Giles and Szpruch [18], Heinrich [21], Heinrich and Sindambiwe [23], and Lemaire and Pagès [32].

The use of multilevel techniques clearly reduces the bias, but in many situations the high variance also brings in a significant inaccuracy, which naturally leads to trying to couple multilevel Monte Carlo with variance reduction techniques. In this work, we focus on importance sampling following the ideas of Arouna [3], who considered a parametric family of stochastic processes $(X_t(\theta))_{0\leq t\leq T}$, with $\theta \in \mathbb{R}^q$, driven by a drifted Brownian motion to build an adaptive importance sampling Monte Carlo method. His algorithm was based on the Robbins-Monro procedure to search for the drift parameter optimally reducing the main term in the variance

$$V(\theta) = \mathbb{E}\left(f^2(X_T(\theta))e^{-\theta\cdot W_T - \frac{1}{2}|\theta|^2 T}\right),$$

where f denotes the payoff function and $(W_t)_{0\leq t\leq T}$ is the q-dimensional standard Brownian motion driving the process X (see the next section for more details). In this Gaussian framework, the standard Robbins-Monro algorithm suffers from numerical instability and may even blow up . To fix this problem, a constrained version of the Robbins-Monro algorithm was proposed by Chen [10, 11] and later investigated by several authors (see, e.g. Andrieu, Moulines and Priouret [2], Lapeyre and Lelong [29] and Lelong [30]). This constrained Robbins-Monro algorithm uses random truncations on an increasing sequence of compact sets to ensure convergence. As tuning such random truncations is not easy, Lemaire and Pagès [31] proposed an alternative modification to circumvent this difficulty. The stability of these stochastic algorithms eventually depends on the choice of the gain sequence, which proves to be highly sensitive in practice. To overcome this difficulty, Jourdain and Lelong [26] proposed to apply deterministic optimization techniques to sample average estimators to search for the optimal parameter. They approximate the unique minimum of V by the unique minimum of

$$V_{n,N}(\theta) = \frac{1}{N}\sum_{k=1}^{N} f^2(X^n_{T,k}(\theta))e^{-\theta\cdot W_{T,k} - \frac{1}{2}|\theta|^2 T}.$$

where $(X^n_{T,k}(\theta))_{1\leq k\leq N}$ are i.i.d. samples of $X^n_T(\theta)$. Doing so, their approach provides a robust and fully automatic variance reduction methodology. Despite the efficiency of the sample average approximation, all attempts to couple several discretization schemes with importance sampling have relied on stochastic approximation to search for the optimal parameter. Ben Alaya, Hajji and Kebaier [8] studied a combination of the statistical Romberg method with both constrained and unconstrained versions of the Robbins-Monro algorithm. Hajji [20] investigated the coupling of multilevel Monte Carlo with the constrained Robbins-Monro algorithm.

In this work, we study how to couple importance sampling and Multilevel Monte Carlo in the framework developed in [26]. Our approach inherits its robustness from the sample average approximation to efficiently reduce the variance and at the same time the multilevel Monte Carlo feature reduces the computational time. The parameter θ is commonly optimized to minimize the asymptotic variance of the estimator, which can be implemented in many ways in the multilevel framework. We have chosen to allow for as many importance sampling parameters as the number of levels L. Hence, we minimize the variance of each level using a sample average approximation given by

$$V_{\ell,m^\ell,N_\ell}(\theta) = \frac{1}{N_\ell}\sum_{k=1}^{N_\ell}\frac{m^\ell}{(m-1)T}\left|f(X^{m^\ell}_{T,\ell,k}) - f(X^{m^{\ell-1}}_{T,\ell,k})\right|^2 e^{-\theta\cdot W_{T,\ell,k}+\frac{1}{2}|\theta|^2T}, \quad m\in\mathbb{N}\setminus\{0,1\}$$

where $(X^{m^\ell}_{T,\ell,k})_{1\le k\le N_\ell}$, $(X^{m^{\ell-1}}_{T,\ell,k})_{1\le k\le N_\ell}$ and $(W_{T,\ell,k})_{1\le k\le N}$ denote the independent copies of respectively the Euler schemes $X^{m^\ell}_T$ and $X^{m^{\ell-1}}_T$ and the Brownian motion $W_{T,\ell}$ used in the $\ell-th$ level of the method (see Section 5 for more details). This approach has many advantages. First, the computations within the different levels remain independent. Second, we actually minimize the real variance of the estimator and not its asymptotic value and more importantly it can be implemented without knowing ∇f, which however appears in the central limit theorem for multilevel Monte Carlo. Yet, our approach attains the optimal limiting variance.

In Section 2, we present our general framework and some preliminary results. In section 3, we study the convergence of the optimal parameter minimizing the map $\theta\mapsto V_{n,N}(\theta)$ when the number of time steps n of the Euler scheme and the sample size N of the Monte Carlo method both tend to infinity. Section 4 addresses the asymptotic properties of the adaptive Monte Carlo method using the estimators developed in Section 3. Theorems 4.1 and 4.2 represent a kind of refinements of the results of [26] as we let both parameters n and N tend to infinity. In section 5, we introduce our multilevel sample average approximation method. First, we study the asymptotic behavior of the optimal parameter minimizing the function $\theta\mapsto V_{\ell,m^\ell,N_\ell}(\theta)$ (see Theorem 5.1). Then, we prove a strong large of law numbers and a central limit theorem for our adaptive multilevel algorithm (see Theorem 5.3 and Theorem 5.4). The main difficulty in proving these results is the uniform control of the triangular arrays involved in the adaptive multilevel estimator. To overcome this issue, we prove in Section 6 new limit theorems for doubly indexed sequences of random variables in a general setting (see Propositions 6.1 and 6.3). In section 7, we illustrate the efficiency of our approach on challenging problems coming from quantitative finance.

2 General framework

Let $(X_t)_{0\le t\le T}$ be the solution of

$$dX_t = b(X_t)dt + \sum_{j=1}^{q}\sigma_j(X_t)dW^j_t, \quad X_0 = x\in\mathbb{R}^d \tag{2.1}$$

where $W = (W^1,\dots,W^q)$ is a q-dimensional Brownian motion on some given probability space $(\Omega,(\mathcal{F}_t)_{0\le t\le T},\mathbb{P})$ with finite time horizon $T>0$. We assume that $(\mathcal{F}_t)_{0\le t\le T}$ is the augmented natural filtration of W. The functions $b : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ and $\sigma_j : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, $1 \leq j \leq q$, satisfy the following condition

$$\forall x, y \in \mathbb{R}^d \quad |b(x) - b(y)| + \sum_{j=1}^{q} |\sigma_j(x) - \sigma_j(y)| \leq C_{b,\sigma}|x - y|, \quad \text{with } C_{b,\sigma} > 0, \tag{$\mathcal{H}_{b,\sigma}$}$$

where $|\cdot|$ denotes the Euclidean norm. This property ensures the strong existence and uniqueness of a solution to (2.1). In many applications, in particular for the pricing of financial securities, we are interested in the effective computation by Monte Carlo methods of the quantity $\mathbb{E}[\psi(X_T)]$ for a given function ψ. From a practical point of view, we have to discretize the process X. Let us consider the continuous time Euler approximation X^n with time step $\delta = T/n$ given by

$$dX_t^n = b(X_{\eta_n(t)}^n)dt + \sum_{j=1}^{q} \sigma_j(X_{\eta_n(t)}^n)dW_t, \quad \eta_n(t) = \lfloor t/\delta \rfloor \delta. \tag{2.2}$$

It is well known that, under Condition ($\mathcal{H}_{b,\sigma}$), X^n converges to X (see e.g. Bouleau and Lépingle [9])

$$\forall p \geq 1, \quad X, X^n \in L^p \quad \text{and} \quad \mathbb{E}\left[\sup_{0\leq t\leq T} |X_t - X_t^n|^p\right] \leq \frac{K_p(T)}{n^{p/2}}, \quad \text{with } K_p(T) > 0. \tag{$\mathcal{P}$}$$

The weak error was first studied by Talay and Tubaro [36] and it is now well known that if ψ, b and $(\sigma_j)_{1\leq j\leq q}$ are in $\mathscr{C}_P^4$ — they are four times differentiable and together with their derivatives and have at most polynomial growth — then we have (see Theorem 14.5.1 by Kloeden and Platen in [28])

$$\varepsilon_n \triangleq \mathbb{E}[\psi(X_T^n)] - \mathbb{E}[\psi(X_T)] = O(1/n).$$

The same result was later extended by Bally and Talay [5] for a measurable function ψ but with a non degeneracy condition of Hörmander's type on the diffusion. In the context of possibly degenerate diffusions, when ψ satisfies $|\psi(x) - \psi(y)| \leq C(1 + |x|^p + |y|^p)|x - y|$ for $C > 0$, $p \geq 0$, the estimate $|\mathbb{E}[\psi(X_T^n)] - \mathbb{E}[\psi(X_T)]| \leq \frac{c}{\sqrt{n}}$ follows easily from ($\mathcal{P}$). Moreover, Kebaier [27] proved that if in addition b and $(\sigma_j)_{1\leq j\leq q}$ are $\mathscr{C}^1$ and ψ satisfies the following condition

$$\mathbb{P}(X_T \notin \mathcal{D}_\psi) = 0, \text{ where } \mathcal{D}_\psi = \{x \in \mathbb{R}^d \mid \psi \text{ is differentiable at } x\}$$

then, $\lim_{n\to\infty} \sqrt{n}\,\varepsilon_n = 0$. Conversely, under the same assumptions, he showed that the rate of convergence can be $1/n^\gamma$, for any $\gamma \in [1/2, 1]$. So, it is worth introducing the following assumption

$$\text{for } \gamma \in [1/2, 1] \quad n^\gamma(\mathbb{E}\psi(X_T^n) - \mathbb{E}\psi(X_T)) \to C_\psi(T, \gamma), \quad C_\psi(T, \gamma) \in \mathbb{R}. \tag{2.3}$$

In order to use importance sampling based on the Girsanov Theorem, we define the family $(\mathbb{P}_\theta)_{\theta\in\mathbb{R}^q}$ of equivalent probability measures such that for all $t > 0$

$$L_t^\theta = \frac{d\mathbb{P}_\theta}{d\mathbb{P}}_{|\mathcal{F}_t} = \exp\left(\theta \cdot W_t - \frac{1}{2}|\theta|^2 t\right).$$

Hence, $(B^\theta_t \triangleq W_t - \theta t)_{t\le T}$ is a Brownian motion under $\mathbb{P}_\theta$, which yields

$$\mathbb{E}[\psi(X_T)] = \mathbb{E}_{\mathbb{P}_\theta}\left[\psi(X_T)e^{-\theta\cdot B^\theta_T-\frac{1}{2}|\theta|^2T}\right].$$

From now on, we assume that

$$\mathbb{P}(\psi(X_T)\neq 0)>0 \quad \text{and} \quad \forall\theta\in\mathbb{R}^q,\ \mathbb{E}[\psi(X_T)^2\mathrm{e}^{-\theta\cdot W_T}]<+\infty. \tag{2.4}$$

For $\alpha>0$, we introduce the set of functions

$$\begin{aligned}\mathcal{H}_\alpha = \Big\{\psi:\mathbb{R}^d\to\mathbb{R}\text{ s.t. }\exists c>0,\beta\ge 1,\ \forall x\in\mathbb{R}^d,\ |\psi(x)|\le c(1+|x|^\beta)\\ \text{and }\forall x,y\in\mathbb{R}^d,\ |\psi(x)-\psi(y)|\le c(1+(|x|^\beta\wedge|y|^\beta))|x-y|^\alpha\Big\}\end{aligned} \tag{2.5}$$

Remark 2.1. *By Hölder's inequality, for any function $\psi\in\mathcal{H}_\alpha$ Equation (2.4) implies that $\sup_n\mathbb{E}[\psi(X^n_T)^2\mathrm{e}^{-\theta\cdot W_T}]<+\infty$.*

Let us introduce the process $X(\theta)$ solution to

$$dX_t(\theta) = \left(b(X_t(\theta)) + \sum_{j=1}^q\theta_j\sigma_j(X_t(\theta))\right)dt + \sum_{j=1}^q\sigma_j(X_t(\theta))dW^j_t, \tag{2.6}$$

so that the pair of processes $(B^\theta_t, X_t)_{0\le t\le T}$ has the same distribution under $\mathbb{P}_\theta$ as the pair $(W_t, X_t(\theta))_{0\le t\le T}$ under $\mathbb{P}$. Henceforth, we get

$$\mathbb{E}[\psi(X_T)] = \mathbb{E}\left[\psi(X_T(\theta))\mathrm{e}^{-\theta\cdot W_T-\frac{1}{2}|\theta|^2T}\right],\quad\forall\theta\in\mathbb{R}^q. \tag{2.7}$$

We also introduce the continuous time Euler approximation $X^n(\theta)$ of the process $X(\theta)$

$$dX^n_t(\theta) = \left(b(X^n_{\eta_n(t)}(\theta)) + \sum_{j=1}^q\theta_j\sigma_j(X^n_{\eta_n(t)}(\theta))\right)dt + \sum_{j=1}^q\sigma_j(X^n_{\eta_n(t)}(\theta))dW^j_t.$$

It is natural to choose the value of θ minimizing $\operatorname{Var}\left(\psi(X_T(\theta))\mathrm{e}^{-\theta\cdot W_T-\frac{1}{2}|\theta|^2T}\right)$, we set

$$\theta^*\triangleq\operatorname*{argmin}_{\theta\in\mathbb{R}^q}v(\theta)\quad\text{with}\quad v(\theta)\triangleq\mathbb{E}\left[\psi(X_T)^2e^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T}\right] \tag{2.8}$$

From a practical point of view, the quantity $v(\theta)$ is not explicit so we use the Euler scheme to discretize $X(\theta)$ and approximate θ^* by

$$\theta_n\triangleq\operatorname*{argmin}_{\theta\in\mathbb{R}^q}v_n(\theta)\quad\text{with}\quad v_n(\theta)\triangleq\mathbb{E}\left[\psi(X^n_T)^2e^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T}\right] \tag{2.9}$$

Since the expectation is usually not tractable, we aim at using a sample average approximation procedure to approximate θ_n

$$\theta_{n,N}\triangleq\operatorname*{argmin}_{\theta\in\mathbb{R}^q}v_{n,N}(\theta)\quad\text{with}\quad v_{n,N}(\theta)\triangleq\frac{1}{N}\sum_{i=1}^N\left(\psi(X^n_{T,i})^2e^{-\theta\cdot W_{T,i}+\frac{1}{2}|\theta|^2T}\right), \tag{2.10}$$

where $(X^n_{T,i}, W_{T,i})_{1\le i\le N}$ are i.i.d. samples according to the law of (X^n_T, W_T). The existence and uniqueness of θ^*, θ_n and $\theta_{n,N}$ are ensured by the following Lemma.

Lemma 2.2. *Under Condition (2.4), the functions v, v_n and $v_{n,N}$ are infinitely continuously differentiable for all $n, N \geq 1$ and for all multi-index $r \in \mathbb{N}^q$, we have*

$$\partial_\theta^r v(\theta) = \mathbb{E}\left[\partial_\theta^r \left(\psi(X_T)^2 e^{-\theta \cdot W_T + \frac{1}{2}|\theta|^2 T}\right)\right],$$
$$\partial_\theta^r v_n(\theta) = \mathbb{E}\left[\partial_\theta^r \left(\psi(X_T^n)^2 e^{-\theta \cdot W_T + \frac{1}{2}|\theta|^2 T}\right)\right].$$

Moreover, under Condition (2.4), the functions v, v_n and $v_{n,N}$ are strongly convex for any n and N large enough.

The proof of this Lemma can be easily adapted from [26, Lemma 1.1].

3 Convergence of the optimal importance sampling parameter

Theorem 3.1. *Suppose σ and b satisfy ($\mathcal{H}_{b,\sigma}$). Let ψ satisfy Condition (2.4) and belongs to $\mathcal{H}_\alpha$ for some $\alpha > 0$. Then,*

$$\theta_n \xrightarrow[n \to +\infty]{} \theta^*.$$

Using remark 2.1, the proof of this result ensues from [8, Theorem 2.2].

In the following, we let N depend on n so that $N \triangleq N_n$ tends to infinity with n.

Proposition 3.2. *Assume that Assumption ($\mathcal{H}_{b,\sigma}$) holds and that $\psi \in \mathcal{H}_\alpha$ for some $\alpha > 0$. Then, for all $K > 0$*

$$\sup_{|\theta| \leq K} |v_{n,N_n}(\theta) - v(\theta)| \xrightarrow[n \to +\infty]{} 0 \quad a.s.$$
$$\sup_{|\theta| \leq K} |\nabla v_{n,N_n}(\theta) - \nabla(\theta)| \xrightarrow[n \to +\infty]{} 0 \quad a.s.$$

Proof. The proof of the two results are very similar, we dare omit the second one and concentrate on the uniform convergence for v_{n,N_n}. To do so, we will apply Proposition 6.3. Now, we check Assumptions ln2-fm]($\mathcal{H}2$), ln2-u]($\mathcal{H}3$), ln2-sup-u]($\mathcal{H}4$). At first, note that under Assumption ($\mathcal{H}_{b,\sigma}$), we have the almost sure convergence of X_T^n towards X_T. As $\psi \in \mathcal{H}_\alpha$, it follows from Property ($\mathcal{P}$) that for all $a > 1$, $\sup_{n \in \mathbb{N}} \mathbb{E}\left[\left|\psi(X_T^n)^2 e^{-\theta \cdot W_T + \frac{1}{2}|\theta|^2 T}\right|^a\right] < \infty$. Note that for every fixed n, the sequence $\left(\psi(X_{T,i}^n)^2 e^{-\theta \cdot W_{T,i} + \frac{1}{2}|\theta|^2 T}\right)_i$ is i.i.d. Then, we deduce that for all $m \in \mathbb{N}^*$

$$\lim_{n \to \infty} \mathbb{E}\left[v_{n,m}(\theta)\right] = \mathbb{E}\left[\psi(X_T)^2 e^{-\theta \cdot W_T + \frac{1}{2}|\theta|^2 T}\right].$$

This yields ln2-fm]($\mathcal{H}2$). Let $K > 0$. As $\psi \in \mathcal{H}_\alpha$ we obtain using the Cauchy Schwarz inequality and Property ($\mathcal{P}$) that

$$\sup_n \sup_m m \mathrm{Var}\left(\sup_{|\theta| \leq K} v_{n,m}(\theta)\right) \leq \sup_n \mathbb{E}^{1/2}\left[\psi(X_T^n)^8\right] \mathbb{E}^{1/2}\left[\sup_{|\theta| \leq K} e^{-4\theta \cdot W_T + 2|\theta|^2 T}\right] < \infty.$$

Using the same arguments, we also get

$$\sup_n \sup_m \mathrm{Var}\left(\psi(X_{T,m}^n)^2 \sup_{|\theta| \leq K} e^{-\theta \cdot W_{T,m} + \frac{1}{2}|\theta|^2 T}\right) < \infty.$$

This yields ln2-u](𝓗3). Concerning the last assumption, if we fix $\delta > 0$, $\theta \in \mathbb{R}^d$ and set $B(\theta,\delta)$ — the open ball with center θ and radius δ — then we have by Cauchy Schwarz inequality

$$\sup_n \mathbb{E}\left[\psi(X_T^n)^2 \sup_{\theta'\in B(\theta,\delta)} \left|e^{-\theta'\cdot W_T+\frac{1}{2}|\theta'|^2T} - e^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T}\right|\right]^2 \le \sup_n \mathbb{E}\left[\psi(X_T^n)^4\right] \mathbb{E}\left[\sup_{\theta'\in B(\theta,\delta)} \left|e^{-\theta'\cdot W_T+\frac{1}{2}|\theta'|^2T} - e^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T}\right|^2\right].$$

Using the elementary algebric inequality $|e^x - e^y| \le |x-y|\,(e^x+e^y)$, we easily deduce that the quantity $\mathbb{E}\left[\sup_{\theta'\in B(\theta,\delta)} \left|e^{-\theta'\cdot W_T+\frac{1}{2}|\theta'|^2T} - e^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T}\right|^2\right]$ can be made arbitrarily small. Finally, Assumption ln2-sup-u]($\mathcal{H}4$) is satisfied using Remark 6.4. □

Theorem 3.3. *Assume that Assumption* ($\mathcal{H}_{b,\sigma}$) *holds and that* $\psi \in \mathcal{H}_\alpha$ *for some* $\alpha > 0$. *Then,* $\theta_{n,N_n} \xrightarrow[n\to+\infty]{a.s.} \theta^*$ *and* $\sqrt{N_n}(\theta_{n,N_n} - \theta^*) \xRightarrow[n\to+\infty]{} N(0,\Gamma)$ *where*

$$\Gamma = [\nabla^2 v(\theta^*)]^{-1}\mathrm{Var}\left[(T\theta^* - W_t)\psi(X_T)^2 e^{-\theta^*\cdot W_T+\frac{1}{2}|\theta^*|^2T}\right][\nabla^2 v(\theta^*)]^{-1}.$$

Proof. We already know from Proposition 3.2 that a.s. v_{n,N_n} converges locally uniformly to v. Let $\varepsilon > 0$. By the strict convexity of v, $\delta \triangleq \inf_{|\theta-\theta^*|\ge\varepsilon} v(\theta) - v(\theta^*) > 0$.
The local uniform convergence of v_{n,N_n} to v ensures that

$$\exists n_\delta > 0, \forall n \ge n_\delta, \forall\theta \in \mathbb{R}^q \text{ s.t. } |\theta - \theta^*| \le \varepsilon,\ |v_{n,N_n}(\theta) - v(\theta)| \le \frac{\delta}{3}. \tag{3.1}$$

For $n \ge n_\delta$ and θ such that $|\theta-\theta^*| \ge \varepsilon$, we can deduce from the convexity of v_{n,N_n} that

$$\begin{aligned}
v_{n,N_n}(\theta) - v_{n,N_n}(\theta^*) &\ge \frac{|\theta-\theta^*|}{\varepsilon}\left[v_{n,N_n}\left(\theta^* + \varepsilon\frac{\theta-\theta^*}{|\theta-\theta^*|}\right) - v_{n,N_n}(\theta^*)\right] \\
&\ge \frac{|\theta-\theta^*|}{\varepsilon}\left[v\left(\theta^* + \varepsilon\frac{\theta-\theta^*}{|\theta-\theta^*|}\right) - v(\theta^*) - \frac{2\delta}{3}\right] \ge \frac{\delta}{3}
\end{aligned}$$

where the last two inequalities come from (3.1). If we apply this inequality for $\theta = \theta_{n,N_n}$, we obtain a contradiction since $v_{n,N_n}(\theta_{n,N_n}) - v_{n,N_n}(\theta^*) \le 0$. Hence, we deduce that for all $n \ge n_\delta$, $|\theta_{n,N_n} - \theta^*| < \varepsilon$. Therefore, θ_{n,N_n} converges a.s. to θ^*. If we combine this result with the local uniform convergence of v_{n,N_n} to the continuous function v, we deduce that $v_{n,N_n}(\theta_{n,N_n})$ converges a.s. to $v(\theta^*)$.

Moreover, we get by Equation (5.9) that for all $K > 0$

$$\begin{aligned}
&\sup_{|\theta|\le K}\left|\partial_{\theta(j)}\psi(X_T)^2 e^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T}\right| \\
&\qquad \le e^{K^2T/2}\psi(X_T)^2\left(K + (e^{KW_t^{(j)}} + e^{-KW_t^{(j)}})\right)\prod_{i=1}^q (e^{KW_t^{(i)}} + e^{-KW_t^{(i)}}).
\end{aligned}$$

The r.h.s is integrable by Condition (2.4). Hence, $\mathbb{E}\left[\sup_{|\theta|\le K}\left|\nabla_\theta\psi(X_T)^2 e^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T}\right|\right] < +\infty$. Similarly, one can prove that $\mathbb{E}\left[\sup_{|\theta|\le K}\left|\nabla_\theta^2\psi(X_T)^2 e^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T}\right|\right] < +\infty$. Then, to prove the central limit theorem governing the convergence of θ_{n,N_n} to θ^*, we reproduce the proof of [35, Theorem A2, pp. 74], which is mainly based on the a.s. locally uniform convergence of $\nabla v_{n,N_n}$ and on its asymptotic normality ensuing from Theorem A.1. □

4 A second stage Monte Carlo approach

In this section, we aim at building adaptive Monte Carlo estimators in the setting of discretized diffusion processes following the spirit of [26]. Our setting differs mainly because we want to let both the number of time steps and the number of samples go to infinity. Asymptotic results rely on a uniform controls of the triangular arrays involved in the adaptive importance sampling Monte Carlo estimator. The technical results from Section 6 will be tremendously useful to provide such controls.

Using the estimators of θ^* studied in the previous section, we define a Monte Carlo estimator of $\mathbb{E}[\psi(X_T)]$ based on Equation (2.7). We introduce the σ-algebra $\mathcal{G}$ generated by the samples $(W_i)_{i\geq 1}$ used to compute θ_n and θ_{n,N_n}.

Let $(\tilde{W}_i)$ be i.i.d. samples according to the law of (W) but independent of $\mathcal{G}$. Conditionally on $\mathcal{G}$, we introduce i.i.d. samples $(\tilde{X}_i(\theta_{n,N_n}))_i$ following the law of $X(\theta_{n,N_n})$ such that for each i, $\tilde{X}_i(\theta_{n,N_n})$ is the solution of the SDE driven by $\tilde{W}_i$. We introduce $(\tilde{\mathcal{G}}_k)_{k>0}$ the filtration defined by $\tilde{\mathcal{G}}_k = \sigma(\tilde{W}_i, 1 \leq i \leq k)$ and $\mathcal{G}_k^\sharp = \mathcal{G} \vee \tilde{\mathcal{G}}_k$. For each $i > 0$, we also consider $\tilde{X}_i^n(\theta_{n,N_n})$ defined as the Euler discretization of $\tilde{X}_i(\theta_{n,N_n})$. Based on these new sets of samples, we define

$$M_{n,N_n} = \frac{1}{N_n} \sum_{i=1}^{N_n} g(\theta_{n,N_n}, \tilde{X}_{T,i}^n(\theta_{n,N_n}), \tilde{W}_{T,i}),$$

where the function $g : \mathbb{R}^q \times \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}$ is defined by

$$g(\theta, x, y) \triangleq \psi(x) \mathrm{e}^{-\theta \cdot y - \frac{1}{2}|\theta|^2 T}. \tag{4.1}$$

For the clearness of the coming proofs, it is convenient to introduce the following notation

$$M_{n,N_n}(\theta) = \frac{1}{N_n} \sum_{i=1}^{N_n} g(\theta, \tilde{X}_{T,i}^n(\theta), \tilde{W}_{T,i}).$$

Note that $M_{n,N_n} = M_{n,N_n}(\theta_{n,N_n})$.

Theorem 4.1. *Assume that Assumption ($\mathcal{H}_{b,\sigma}$) holds and that $\psi \in \mathcal{H}_\alpha$ for some $\alpha > 0$. Then, $M_{n,N_n} \longrightarrow \mathbb{E}[f(X_T)]$ a.s. when $n \to +\infty$.*

Proof. Using the conditional independence of the samples $(\tilde{X}_i^n(\theta_{n,N_n}), \tilde{W}_i)_i$, we have

$$\mathbb{E}[g(\theta_{n,N_n}, \tilde{X}_{T,i}^n(\theta_{n,N_n}), \tilde{W}_{T,i}) | \mathcal{G}] = \mathbb{E}[\psi(X_T^n)] \triangleq e_n \quad \text{for all } i > 0.$$

Let $\mathcal{V} \subset \mathbb{R}^q$ be a compact neighbourhood of θ^*. We define the sequence

$$Y_{i,n} = \left(g(\theta_{n,N_n}, \tilde{X}_{T,i}^n(\theta_{n,N_n}), \tilde{W}_{T,i}) - e_n \right) 1_{\{\theta_{n,N_n} \in \mathcal{V}\}}$$

and its empirical average $\overline{Y}_{m,n} = \frac{1}{m} \sum_{i=1}^m Y_{i,n}$ for all $m > 0$. It is obvious that $\mathbb{E}[Y_{i,n}] = 0$ and using the conditional independence $\mathbb{E}[\left|\overline{Y}_{m,n}\right|^2] = \frac{1}{m}\mathbb{E}[|Y_{1,n}|^2]$.

$$\begin{aligned} \mathbb{E}[|Y_{1,n}|^2] &\leq \mathbb{E}\left[\mathbb{E}\left[|g(\theta_{n,N_n}, \tilde{X}_{T,i}^n(\theta_{n,N_n}), \tilde{W}_{T,i}) - e_n|^2 \middle| \mathcal{G} \right] 1_{\{\theta_{n,N_n} \in \mathcal{V}\}} \right] \\ &\leq \mathbb{E}\left[v_n(\theta_{n,N_n}) 1_{\{\theta_{n,N_n} \in \mathcal{V}\}} \right] \leq \sup_{\theta \in \mathcal{V}} v_n(\theta). \end{aligned}$$

We know that v_n is convex and converges point-wise to v, which is also convex and continuous. Hence, v_n converges locally uniformly to v, which implies that for all compact sets $K \subset \mathbb{R}^q$, $\lim_{n\to+\infty} \sup_{\theta\in K} v_n(\theta) = \sup_{\theta\in K} v(\theta)$. Hence, $\sup_n \sup_{\theta\in\mathcal{V}} v_n(\theta) < +\infty$. Applying Proposition 6.1 proves that $\overline{Y}_{N_n,n} \xrightarrow[n\to+\infty]{a.s.} 0$. As θ_{n,N_n} converges a.s. to $\theta^* \in K$, this also implies that $\lim_{n\to+\infty} M_{n,N_n} = \mathbb{E}[\psi(X_T)]$ a.s. □

Theorem 4.2. *Under the assumptions of Theorem 4.1 and if Condition (2.3) holds, we have*

$$\sqrt{N_n}(M_{n,N_n} - \mathbb{E}[f(X_T)]) \Rightarrow \mathcal{N}(C_\psi(T,\alpha), \sigma^2) \text{ when } n \to +\infty.$$

where $\sigma^2 = \mathbb{E}\left[\psi(X_T)^2 \mathrm{e}^{-\theta^*\cdot W_T + \frac{1}{2}|\theta^*|^2 T}\right] - [\mathbb{E}\psi(X_T)]^2$.

Remark 4.3. *Assume the number of time steps used in the Euler scheme is fixed to $n = 1$ and consider the estimator $M_{1,N}(\theta_{1,N})$. Then, we know from [4, Theorem 3.4] that*

$$M_{1,N}(\theta_{1,N}) \xrightarrow[N\to+\infty]{} \mathbb{E}[g(\theta_1, X_T^1(\theta_1), W_T)] \quad a.s.$$

$$\sqrt{N}(M_{1,N}(\theta_{1,N}) - \mathbb{E}[g(\theta_1, X_T^1(\theta_1), W_T)]) \xRightarrow[N\to+\infty]{} \mathcal{N}(0, \sigma_1^2)$$

with $\sigma_1^2 = \mathbb{E}\left[\psi(X_T^1)^2 \mathrm{e}^{-\theta_1\cdot W_T + \frac{1}{2}|\theta_1|^2 T}\right] - [\mathbb{E}\psi(X_T^1)]^2$.

Proof. We can write the left hand side of the convergence result by introducing $M_{n,N_n}(\theta^*)$

$$\sqrt{N_n}(M_{n,N_n} - \mathbb{E}[f(X_T)]) = \sqrt{N_n}(M_{n,N_n}(\theta_{n,N_n}) - M_n(\theta^*)) + \sqrt{N_n}(M_{n,N_n}(\theta^*) - \mathbb{E}[f(X_T)])$$

The convergence of the last term on the r.h.s $\sqrt{N_n}(M_{n,N_n}(\theta^*) - \mathbb{E}[f(X_T)])$ is governed by the central limit theorem for Euler Monte Carlo, which yields the announced limit (see [15]). It remains to prove that $\sqrt{N_n}(M_{n,N_n}(\theta_{n,N_n}) - M_{n,N_n}(\theta^*))$ converges to zero in probability.

Let $\varepsilon > 0$ and $\alpha < \frac{1}{2}$,

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{N_n}|M_{n,N_n}(\theta_{n,N_n}) - M_{n,N_n}(\theta^*)| > \varepsilon\right) \\
&= \mathbb{P}\left(\sqrt{N_n}|M_{n,N_n}(\theta_{n,N_n}) - M_{n,N_n}(\theta^*)| > \varepsilon\, ;\ N_n^\alpha|\theta_{n,N_n} - \theta^*| > 1\right) \\
&\quad + \mathbb{P}\left(\sqrt{N_n}|M_{n,N_n}(\theta_{n,N_n}) - M_{n,N_n}(\theta^*)| > \varepsilon\, ;\ N_n^\alpha|\theta_{n,N_n} - \theta^*| \le 1\right) \\
&= \mathbb{P}(\ N_n^\alpha|\theta_{n,N_n} - \theta^*| > 1) \\
&\quad + \mathbb{P}\left(\sqrt{N_n}|M_{n,N_n}(\theta_{n,N_n}) - M_{n,N_n}(\theta^*)| 1_{\{N_n^\alpha|\theta_{n,N_n} - \theta^*| \le 1\}} > \varepsilon\right).
\end{aligned}$$

By Theorem 3.3, $\mathbb{P}(\ N_n^\alpha|\theta_{n,N_n} - \theta^*| > 1)$ tends to zero when n goes to infinity. Let $K > 0$ s.t. for all n large enough $\{\theta \in \mathbb{R}^q : |\theta - \theta^*| \le N_n^{-\alpha}\} \subset B(0, K)$. We can bound the second term on the r.h.s. by using Markov's inequality

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{N_n}|M_{n,N_n}(\theta_{n,N_n}) - M_{n,N_n}(\theta^*)| 1_{\{N_n^\alpha|\theta_{n,N_n} - \theta^*| \le 1\}} > \varepsilon\right) \\
&\le \frac{N_n}{\varepsilon^2} \mathbb{E}\left[|M_{n,N_n}(\theta_{n,N_n}) - M_{n,N_n}(\theta^*)|^2 1_{\{\theta_{n,N_n} \in B(0,K)\}}\right] \\
&\le \frac{1}{\varepsilon^2} \mathbb{E}\left[|g(\theta_{n,N_n}, \tilde{X}_T^n(\theta_{n,N_n}), \tilde{W}_T) - g(\theta^*, \tilde{X}_T^n(\theta^*), \tilde{W}_T)|^2 1_{\{\theta_{n,N_n} \in B(0,K)\}}\right] \\
&\le \frac{1}{\varepsilon^2} \mathbb{E}\left[|g(\theta_{n,N_n}, \tilde{X}_T^n(\theta_{n,N_n}), \tilde{W}_T) - g(\theta_{n,N_n}, \tilde{X}_T(\theta_{n,N_n}), \tilde{W}_T)|^2 1_{\{\theta_{n,N_n} \in B(0,K)\}}\right] \\
&\quad + \frac{1}{\varepsilon^2} \mathbb{E}\left[|g(\theta_{n,N_n}, \tilde{X}_T(\theta_{n,N_n}) \tilde{W}_T) - g(\theta^*, \tilde{X}_T^n(\theta^*), \tilde{W}_T)|^2 1_{\{\theta_{n,N_n} \in B(0,K)\}}\right].
\end{aligned}$$

We treat each of the two terms separately.

▶ **First term**

From the independence between θ_{n,N_n} and $\tilde{W}$, we can write

$$
\begin{aligned}
&\mathbb{E}\left[|g(\theta_{n,N_n}, \tilde{X}_T^n(\theta_{n,N_n}), \tilde{W}_T) - g(\theta_{n,N_n}, \tilde{X}_T(\theta_{n,N_n}), \tilde{W}_T)|^2 1_{\{\theta_{n,N_n}\in B(0,K)\}}\right] \\
&\quad = \mathbb{E}\left[|\psi(X_T^n) - \psi(X_T)|^2 \exp(-\theta_{n,N_n}\cdot \tilde{W}_T + \frac{1}{2}|\theta_{n,N_n}|^2 T) 1_{\{\theta_{n,N_n}\in B(0,K)\}}\right] \\
&\quad \le \mathbb{E}\left[|\psi(X_T^n) - \psi(X_T)|^{2(1+\eta)}\right]^{\frac{1}{1+\eta}} \mathrm{e}^{\frac{1+2\eta}{2\eta}K^2 T}, \quad \text{for some } \eta > 0.
\end{aligned}
$$

Relying on the uniform integrability ensured by property ($\mathcal{P}$) and since $\psi \in \mathcal{H}_\alpha$, we can let n go to infinity inside the expectation to obtain that

$$
\lim_{n\to+\infty} \mathbb{E}\left[|g(\theta_{n,N_n}, \tilde{X}_T^n(\theta_{n,N_n}), \tilde{W}_T) - g(\theta_{n,N_n}, \tilde{X}_T(\theta_{n,N_n}), \tilde{W}_T)|^2 1_{\{\theta_{n,N_n}\in B(0,K)\}}\right] = 0.
$$

▶ **Second term**

Since the function g is continuous w.r.t its first two parameters and X_T^θ is continuous w.r.t the parameter θ, $\lim_{n\to+\infty} g(\theta_{n,N_n}, \tilde{X}_T(\theta_{n,N_n}), \tilde{W}_T) - g(\theta^*, \tilde{X}_T^n(\theta^*), \tilde{W}_T) = 0$ a.s. To conclude the proof, we need to show that the family of r.v.

$$
\left(|g(\theta_{n,N_n}, \tilde{X}_T(\theta_{n,N_n}), \tilde{W}_T) - g(\theta^*, \tilde{X}_T^n(\theta^*), \tilde{W}_T)|^2 1_{\{\theta_{n,N_n}\in B(0,K)\}}\right)_n
$$

is uniformly integrable.

First, for any $\theta \in \mathbb{R}^q$ and $2(1+\eta) > a > 2$

$$
\begin{aligned}
\mathbb{E}\left[|g(\theta, \tilde{X}_T(\theta), \tilde{W}_T)|^a\right] &= \mathbb{E}\left[|\psi(\tilde{X}_T)|^a \mathrm{e}^{-(a-1)\theta\cdot\tilde{W}_T + \frac{(a-1)|\theta|^2 T}{2}}\right] \\
&\le \mathbb{E}\left[|\psi(\tilde{X}_T)|^{2(1+\eta)}\right]^{\frac{2(1+\eta)}{a}} \mathrm{e}^{C|\theta|^2} \qquad (4.2)
\end{aligned}
$$

where C is a constant only depending on a and T. This yields that for some $\delta > 0$ and some constant $C > 0$ independent of θ

$$
\mathbb{E}\left[|g(\theta, \tilde{X}_T(\theta), \tilde{W}_T)|^{2+\delta}\right] < C\mathrm{e}^{C|\theta|^2}.
$$

Then, we get

$$
\begin{aligned}
&\sup_n \mathbb{E}\left[|g(\theta_{n,N_n}, \tilde{X}_T(\theta_{n,N_n}), \tilde{W}_T)|^{2+\delta} 1_{\{\theta_{n,N_n}\in B(0,K)\}}\right] \\
&= \sup_n \mathbb{E}\left[\mathbb{E}\left[|g(\theta_{n,N_n}, \tilde{X}_T(\theta_{n,N_n}), \tilde{W}_T)|^{2+\delta}|\theta_{n,N_n}\right] 1_{\{\theta_{n,N_n}\in B(0,K)\}}\right] \\
&\le \sup_n C\mathbb{E}\left[\mathrm{e}^{C|\theta_{n,N_n}|^2} 1_{\{\theta_{n,N_n}\in B(0,K)\}}\right] \le C\mathrm{e}^{CK}.
\end{aligned}
$$

We can similarly prove that

$$
\sup_n \mathbb{E}\left[|g(\theta^*, \tilde{X}_T^n(\theta^*), \tilde{W}_T)|^{2+\delta}\right] \le \sup_n \mathbb{E}\left[|\psi(X_T^n)|^{2(1+\eta)}\right]^{\frac{2(1+\eta)}{2+\delta}} \mathrm{e}^{C|\theta^*|^2}.
$$

This prove that the family of r.v.

$$
\left(|g(\theta_{n,N_n}, \tilde{X}_T(\theta_{n,N_n}), \tilde{W}_T) - g(\theta^*, \tilde{X}_T^n(\theta^*), \tilde{W}_T)|^2 1_{\{\theta_{n,N_n}\in B(0,K)\}}\right)_n
$$

is uniformly integrable, which ends the proof. □

5 Multilevel Importance sampling Monte Carlo

5.1 The *sample average approximation* setting

We aim at approximating the quantity $\mathbb{E}[\psi(X_T)]$ by a multilevel approach combined with some importance sampling, while allowing for one importance sampling parameter per level. Let $m \in \mathbb{N}$ such that $m \geq 2$. For $L \in \mathbb{N}^*$, our estimator is defined by

$$
\begin{aligned}
Q_L(\lambda_0, \ldots, \lambda_L) = & \frac{1}{N_0} \sum_{k=1}^{N_0} \psi(\tilde{X}^{m^0}_{T,0,k}(\lambda_0)) \mathcal{E}^-(\tilde{W}_{0,k}, \lambda_0) \\
& + \sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{k=1}^{N_\ell} \left(\psi(\tilde{X}^{m^\ell}_{T,\ell,k}(\lambda_\ell)) - \psi(\tilde{X}^{m^{\ell-1}}_{T,\ell,k}(\lambda_\ell)) \right) \mathcal{E}^-(\tilde{W}_{\ell,k}, \lambda_\ell)
\end{aligned}
$$

for any $\Lambda_L = (\lambda_0, \ldots, \lambda_L) \in (\mathbb{R}^q)^L$ with

$$
\mathcal{E}^-(\tilde{W}_{\ell,k}, \lambda) \triangleq \mathrm{e}^{-\lambda \cdot \tilde{W}_{T,\ell,k} - \frac{1}{2}|\lambda|^2 T}.
$$

For any fixed $\ell \in \{1, \cdots, L\}$, the random variables $(\tilde{W}_{\ell,k})_{1 \leq k \leq N_\ell}$ are independent and are distributed according to the Brownian law. We assume that for $\ell, \ell' \in \{1, \cdots, L\}$, with $\ell \neq \ell'$, the blocks $(\tilde{W}_{\ell,k})_{1 \leq k \leq N_\ell}$ and $(\tilde{W}_{\ell',k})_{1 \leq k \leq N_{\ell'}}$ are independent. For any fixed $\ell \in \{1, \cdots, L\}$ and $k \in \{1, \ldots, N_\ell\}$, the variables $\tilde{X}^{m^\ell}_{T,\ell,k}(\lambda_\ell)$ (resp. $\tilde{X}^{m^{\ell-1}}_{T,\ell,k}(\lambda_\ell)$) are the terminal values of the Euler schemes of $X(\lambda_\ell)$ with m^ℓ (resp. $m^{\ell-1}$) time steps built using the same Brownian path $\tilde{W}_{\ell,k}$. The key of the multilevel approach is to use the same Brownian path to compute $\tilde{X}^{m^\ell}_{T,\ell,k}(\lambda_\ell)$ and $\tilde{X}^{m^{\ell-1}}_{T,\ell,k}(\lambda_\ell)$. The blocks of random variables used in two different levels are independent. From these assumptions, one can compute the variance of the multilevel estimator given by

$$
\mathrm{Var}[Q_L] = N_0^{-1} \mathrm{Var}[\psi(X_T^{m^0}(\lambda_0)) \mathcal{E}_0^-(\lambda_0)] + \sum_{\ell=1}^{L} N_\ell^{-1} \frac{(m-1)T}{m^\ell} \sigma_\ell^2(\lambda_\ell)
$$

where

$$
\sigma_\ell^2(\lambda) \triangleq \frac{m^\ell}{(m-1)T} \mathrm{Var}\left[\left\{ \psi(X_T^{m^\ell}(\lambda)) - \psi(X_T^{m^{\ell-1}}(\lambda)) \right\} \mathcal{E}^-(W, \lambda) \right].
$$

The variance can be rewritten as $\sigma_\ell^2(\lambda) = v_\ell(\lambda) - \Xi_\ell^2$ with

$$
v_\ell(\lambda) \triangleq \frac{m^\ell}{(m-1)T} \mathbb{E}\left[\left| \psi(X_T^{m^\ell}) - \psi(X_T^{m^{\ell-1}}) \right|^2 \mathcal{E}^+(W, \lambda) \right], \tag{5.1}
$$

$$
\Xi_\ell \triangleq \sqrt{\frac{m^\ell}{(m-1)T}} \mathbb{E}\left[\psi(X_T^{m^\ell}) - \psi(X_T^{m^{\ell-1}}) \right] \tag{5.2}
$$

and

$$
\mathcal{E}^+(W_{\ell,k}, \lambda) \triangleq \mathrm{e}^{-\lambda \cdot W_{T,\ell,k} + \frac{1}{2}|\lambda|^2 T}.
$$

Thus, we can rewrite the global variance as

$$
\mathrm{Var}[Q_L] = N_0^{-1} \mathrm{Var}[\psi(X_T^{m^0}) \mathcal{E}^+(W, \lambda_0)] + \sum_{\ell=1}^{L} N_\ell^{-1} \frac{(m-1)T}{m^\ell} \left(v_\ell(\lambda_\ell) - \Xi_\ell^2 \right).
$$

We also define the Monte Carlo approximations of Ξ_ℓ and v_ℓ

$$\Xi_{\ell,N'_\ell} \triangleq \frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\sqrt{\frac{m^\ell}{(m-1)T}}(\psi(X^{m^\ell}_{T,\ell,k}) - \psi(X^{m^{\ell-1}}_{T,\ell,k})) \tag{5.3}$$

$$v_{\ell,N'_\ell}(\lambda) \triangleq \frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}\left|\psi(X^{m^\ell}_{T,\ell,k}) - \psi(X^{m^{\ell-1}}_{T,\ell,k})\right|^2\mathcal{E}^+(W_{\ell,k},\lambda) \tag{5.4}$$

where the variables W_k are i.i.d. according to the Brownian law on $[0,T]$ and are independent of the $\tilde{W}_k$. Based on these new Brownian paths, we introduce the random variables $X^{m^\ell}_{T,k}$, defined in the same way as the tilde quantities but independent of them. Hence, the estimators v_{ℓ,N'_ℓ} for $\ell = 1,\dots,L$ are independent of $Q_L(\lambda_0,\dots,\lambda_L)$. Note that the number N'_ℓ of samples used to build a Monte Carlo approximation of v_ℓ may differ from the number N_ℓ of samples used in the computation of the level ℓ of Q_L. This point will be discussed in details in the numerical section (see the end of Section 7.1). For the moment, we just require that N'_ℓ goes to infinity with ℓ.

By applying Lemma 2.2, it is clear that the functions v_ℓ and v_{ℓ,N'_ℓ} are strongly convex and infinitely differentiable. Hence, we can define

$$\widehat{\lambda}_\ell = \arg\min_{\lambda\in\mathbb{R}^q} v_{\ell,N'_\ell}(\lambda).$$

To study the convergence of $\widehat{\lambda}_\ell$, we need to introduce the process U defined by

$$dU_t = \dot{b}(X_t)U_t dt + \sum_{j=1}^{q}\dot{\sigma}_j(X_t)U_t dW^j_t - \frac{1}{\sqrt{2}}\sum_{ij,=1}^{q}\dot{\sigma}_j(X_t)\sigma_i(X_t)d\check{W}^{i,j}_t \tag{5.5}$$

where $\check{W}$ is a q^2-dimensional standard Brownian motion independent of W.

Theorem 5.1. *Assume b and σ are $\mathscr{C}^1$ with bounded derivatives, $\psi \in \mathcal{H}_\alpha$ for some $\alpha \geq 1$, ψ is C^1 and $\nabla\psi$ has polynomial growth. Then, the sequence of random functions $(v_{\ell,N'_\ell} : \lambda \in \mathbb{R}^q \to v_{\ell,N'_\ell}(\lambda))_\ell$ converges a.s. locally uniformly to the strongly convex function $v : \mathbb{R}^q \to \mathbb{R}$ defined by*

$$v(\lambda) \triangleq \mathbb{E}\left[(\nabla\psi(X_T)\cdot U_T)^2\,\mathcal{E}^+(W,\lambda)\right]. \tag{5.6}$$

Moreover, $\widehat{\lambda}_\ell$ converges a.s. to $\lambda^* \triangleq \arg\min_\lambda v(\lambda)$, when $\ell \to +\infty$.

The proofs of this result and many subsequent ones heavily rely on the following L^p control of the difference between two levels

Proposition 5.2. *Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a C^1 function such that $\psi \in \mathcal{H}_\alpha$, for some $\alpha \geq 1$ and $\nabla\psi$ has at most polynomial growth. For any real valued random variable Y defined on $(\Omega,\mathcal{F})$ such that $\mathbb{E}[|Y|^{1+\eta}]$, for some $\eta > 0$, we have, for any $\delta > 0$*

$$\mathbb{E}\left[\left(\frac{m^\ell}{(m-1)T}\right)^{\delta/2}\left(\psi(X^{m^\ell}_T) - \psi(X^{m^{\ell-1}}_T)\right)^\delta Y\right] \xrightarrow[\ell\to+\infty]{} \mathbb{E}\left[(\nabla\psi(X_T)\cdot U_T)^\delta\, Y\right].$$

Proof of Theorem 5.1. Let us define the doubly indexed sequence

$$Y_{k,\ell}(\lambda) = \frac{m^\ell}{(m-1)T} \left|\psi(X^{m^\ell}_{T,k}) - \psi(X^{m^{\ell-1}}_{T,k})\right|^2 \mathcal{E}^+(W_k, \lambda).$$

For any fixed ℓ, the sequence $(Y_{k,\ell}(\lambda))_k$ is i.i.d. so that for any k, $\mathbb{E}[Y_{k,\ell}(\lambda)] = y_\ell(\lambda)$ with

$$y_\ell(\lambda) = \mathbb{E}\left[\frac{m^\ell}{(m-1)T} \left|\psi(X^{m^\ell}_T) - \psi(X^{m^{\ell-1}}_T)\right|^2 \mathcal{E}^+(W, \lambda)\right].$$

We deduce from Proposition 5.2 that the sequence $(y_\ell)_\ell$ converges pointwise to the continuous function $\mathbb{E}\left[(\nabla\psi(X_T)\cdot U_T)^2\, \mathcal{E}^+(W,\lambda)\right]$, thus satisfying Assumption m-pointwise]($\mathcal{H}2$)-*i*. The i.i.d. property of the sequence $(Y_{k,\ell}(\lambda))_k$ also implies that

$$\mathbb{E}\left[\sup_{|\lambda|\le K} \frac{1}{N}\left(\sum_{k=1}^N Y_{k,\ell}(\lambda)\right)^2\right] \le \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^N \sup_{|\lambda|\le K} Y_{k,\ell}(\lambda)^2\right] \le \frac{1}{N}\mathbb{E}\left[\sup_{|\lambda|\le K} Y_{1,\ell}(\lambda)^2\right]. \tag{5.7}$$

$$\mathbb{E}\left[\sup_{|\lambda|\le K} Y_{1,\ell}(\lambda)^2\right]^2 \le \mathbb{E}\left[\left(\frac{m^\ell}{(m-1)T}\left|\psi(X^{m^\ell}_T) - \psi(X^{m^{\ell-1}}_T)\right|^2\right)^4\right] \mathbb{E}\left[\sup_{|\lambda|\le K} \mathcal{E}^+(W,\lambda)^4\right]. \tag{5.8}$$

Using the following upper bound

$$\sup_{|\lambda|\le K} e^{-\lambda\cdot W_T + \frac{1}{2}|\lambda|^2 T} \le \mathrm{e}^{\frac{1}{2}K^2 T}\prod_{l=1}^q (\mathrm{e}^{K W^{(l)}_T} + \mathrm{e}^{-K W^{(l)}_T}), \tag{5.9}$$

$\mathbb{E}\left[\sup_{|\lambda|\le K} \mathcal{E}^+(W,\lambda)^4\right] < +\infty$. Let us have a closer look at the first term in (5.8). From Condition (2.5), we can write

$$\mathbb{E}\left[\left(m^\ell\left|\psi(X^{m^\ell}_T) - \psi(X^{m^{\ell-1}}_T)\right|^2\right)^4\right] \le C\mathbb{E}\left[m^{4\ell}\left|X^{m^\ell}_T - X^{m^{\ell-1}}_T\right|^{8\alpha}\left(1 + \left|X^{m^\ell}_T\right|^{8\beta} + \left|X^{m^{\ell-1}}_T\right|^{8\beta}\right)\right].$$

By using the strong rate of convergence of the Euler scheme, we notice that for any $p > 1$,

$$\mathbb{E}\left[m^{4\ell p}\left|X^{m^\ell}_T - X^{m^{\ell-1}}_T\right|^{8\alpha p}\right] \le m^{4\ell p} C\left(m^{-4\alpha p\ell} + m^{-4\alpha p(\ell-1)}\right) \le C m^{4\alpha p - 4\ell p(\alpha-1)}.$$

Hence, since $\alpha \ge 1$, by using the Cauchy Schwartz inequality we easily check that

$$\sup_\ell \mathbb{E}\left[\left(\frac{m^\ell}{(m-1)T}\left|\psi(X^{m^\ell}_T) - \psi(X^{m^{\ell-1}}_T)\right|^2\right)^4\right] < +\infty.$$

By combining all these results into (5.8), we obtain that $\sup_\ell \mathbb{E}\left[\sup_{|\lambda|\le K} Y^2_{1,\ell}(\lambda)\right] < +\infty$. Then, we deduce along with (5.7) that the sequence $(Y_{k,\ell})_{k,\ell}$ satisfies Assumption ln2-u]($\mathcal{H}3$) of Proposition 6.3.

Let $\delta > 0$ and $\lambda \in \mathbb{R}^d$.

$$\mathbb{E}\left[\sup_{|\mu-\lambda|\leq\delta} |Y_{1,\ell}(\lambda) - Y_{1,\ell}(\mu)|\right]^2 \leq$$
$$\mathbb{E}\left[\left(\frac{m^\ell}{(m-1)T}\left|\psi(X_T^{m^\ell}) - \psi(X_T^{m^{\ell-1}})\right|^2\right)^2\right] \mathbb{E}\left[\sup_{|\mu-\lambda|\leq\delta} \left|\mathcal{E}^+(W,\lambda) - \mathcal{E}^+(W,\mu)\right|^2\right].$$

We have just proved that the first expectation on the r.h.s is bounded uniformly in ℓ. Since the exponential weights are a.s. continuous with respect to λ, it is clear that $\lim_{\delta\to 0}\sup_{|\mu-\lambda|\leq\delta} \left|\mathcal{E}^+(W,\lambda) - \mathcal{E}^+(W,\mu)\right|^2 = 0$ a.s. Moreover, we can apply Lebesgue's theorem with the upper–bound given by (5.9) to deduce that

$$\lim_{\delta\to 0}\sup_{\ell} \mathbb{E}\left[\sup_{|\mu-\lambda|\leq\delta} |Y_{1,\ell}(\lambda) - Y_{k,\ell}(\mu)|\right] = 0.$$

Thus, Assumption ln2-sup-u]$(\mathcal{H}4)$ of Proposition 6.3 is satisfied. Finally, we can apply Proposition 6.3 to prove that the sequence $\frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell} Y_{k,\ell}$ converges a.s locally uniformly to 0.

The convergence of $\widehat{\lambda_\ell}$ to λ^* can be deduced by closely mimicking the proof of Theorem 3.3. □

Proof of Proposition 5.2. The Taylor-Young expansion applied to the real valued function ψ yields

$$\begin{aligned}\psi(X_T^{m^\ell}) - \psi(X_T^{m^{\ell-1}}) =& \nabla\psi(X_T)\cdot(X_T^{m^\ell} - X_T^{m^{\ell-1}}) \\ &+ (X_T^{m^\ell} - X_T)\cdot\varepsilon(X_T^{m^\ell} - X_T) - (X_T^{m^{\ell-1}} - X_T)\cdot\varepsilon(X_T^{m^{\ell-1}} - X_T)\end{aligned}$$

with $\varepsilon : \mathbb{R}^d \to \mathbb{R}^d$ satisfying $\lim_{|x|\to 0}\varepsilon(x) = 0$. From Property $(\mathcal{P})$, we easily get

$$\sqrt{\frac{m^\ell}{(m-1)T}}\left((X_T^{m^\ell} - X_T)\cdot\varepsilon(X_T^{m^\ell} - X_T) - (X_T^{m^{\ell-1}} - X_T)\cdot\varepsilon(X_T^{m^{\ell-1}} - X_T)\right) \xrightarrow[\ell\to\infty]{\mathbb{P}} 0.$$

So, we conclude from Lemma A.2 and Theorem A.3 that

$$\sqrt{\frac{m^\ell}{(m-1)T}}\left(\psi(X_T^{m^\ell}) - \psi(X_T^{m^{\ell-1}})\right) \Rightarrow^{stably} \nabla\psi(X_T).U_T, \quad \text{as} \quad \ell\to\infty.$$

Let $\eta > \kappa > 0$. From the assumptions on ψ together with Property $(\mathcal{P})$, we get

$$\sup_{\ell\geq 0}\mathbb{E}\left[\left|\left(\frac{m^\ell}{(m-1)T}\right)^{\delta/2}\left(\psi(X_T^{m^\ell}) - \psi(X_T^{m^{\ell-1}})\right)^\delta Y\right|^{1+\kappa}\right] < \infty,$$

which yields the uniform integrability of the family $\left(\left(\frac{m^\ell}{(m-1)T}\right)^{\delta/2}\left(\psi(X_T^{m^\ell}) - \psi(X_T^{m^{\ell-1}})\right)^\delta Y\right)_\ell$. The conclusion easily follows. □

5.2 Strong law of large numbers and central limit theorem

Let us introduce a sequence $(a_\ell)_{\ell\in\mathbb{N}}$ of positive real numbers such that $\lim_{L\to\infty}\sum_{\ell=1}^{L} a_\ell = \infty$. We assume that the sample size N_ℓ has the following form

$$N^{\rho}_{\ell,L} = \frac{\rho(L)}{m^\ell a_\ell}\sum_{k=1}^{L} a_k, \quad \ell \in \{0,\cdots,L\} \tag{5.10}$$

for some increasing function $\rho : \mathbb{N} \to \mathbb{R}$.

We choose this form for N_ℓ because it is a generic form allowing us a straightforward use of the Toeplitz Lemma, which is a key tool to prove the central limit theorem. Since $\lim_{L\to\infty}\sum_{\ell=1}^{L} a_\ell = \infty$, for any sequence $(x_\ell)_{\ell\geq 1}$ converging to some limit $x \in \mathbb{R}$,

$$\lim_{L\to+\infty} \frac{\sum_{\ell=1}^{L} a_\ell x_\ell}{\sum_{\ell=1}^{L} a_\ell} = x.$$

We define the σ-algebra $\mathcal{G}$ generated by the samples $(W_{\ell,k})_{\ell,k\geq 1}$ used to compute $\widehat{\lambda}_L$. In the above framework, the variables $(\tilde{W}_{\ell,k})_{\ell,k}$ are independent of $\mathcal{G}$. We also introduce the filtration $(\tilde{\mathcal{G}}_\ell)_{\ell>0}$ generated by $(\tilde{W}_{\ell,k}, k \geq 1)_\ell$ and the filtration $(\mathcal{G}^\sharp_\ell)_{\ell>0}$ defined as $\mathcal{G}^\sharp_l = \mathcal{G} \vee \tilde{\mathcal{G}}_\ell$.

Theorem 5.3. *Assume that* $\sup_L \sup_\ell \frac{L^2 a_\ell}{\rho(L)\sum_{k=1}^{L} a_k} < +\infty$. *Then, under the assumptions of Theorem 5.1,* $Q_L(\widehat{\lambda}_0,\dots,\widehat{\lambda}_L) \longrightarrow \mathbb{E}[\psi(X_T)]$ *a.s. when* $L \to +\infty$.

For the choice $a_\ell = 1$ for all ℓ, the condition on ρ reduces to $\sup_L \frac{L}{\rho(L)} < +\infty$.

Proof. As $\mathbb{E}[\psi(X^L_T)]$ converges to $\mathbb{E}[\psi(X_T)]$ as L goes to infinity, it is enough to show that $Q_L(\widehat{\lambda}_0,\dots,\widehat{\lambda}_L) - \mathbb{E}[\psi(X^L_T)]$ tends to 0.

$$\begin{aligned} Q_L(\widehat{\lambda}_0,\dots,\widehat{\lambda}_L) - \mathbb{E}[\psi(X^L_T)] &= \frac{1}{N^\rho_{0,L}}\sum_{k=1}^{N^\rho_{0,L}} \psi(\tilde{X}^{m^0}_{T,0,k}(\widehat{\lambda}_0))\mathcal{E}^-(\tilde{W}_{0,k},\widehat{\lambda}_0) - \mathbb{E}[\psi(X^{m^0}_{T,0})] \\ &+ \sum_{\ell=1}^{L}\frac{1}{N^\rho_{\ell,L}}\Bigg(\sum_{k=1}^{N^\rho_{\ell,L}}\Big(\psi(\tilde{X}^{m^\ell}_{T,\ell,k}(\widehat{\lambda}_\ell)) - \psi(\tilde{X}^{m^{\ell-1}}_{T,\ell,k}(\widehat{\lambda}_\ell))\Big)\mathcal{E}^-(\tilde{W}_{\ell,k},\widehat{\lambda}_\ell) \\ &\qquad - \mathbb{E}\left[\psi(\tilde{X}^{m^\ell}_{T,\ell}) - \psi(\tilde{X}^{m^{\ell-1}}_{T,\ell})\right]\Bigg). \end{aligned} \tag{5.11}$$

From Theorem 4.1 and Remark 4.3, we know that

$$\frac{1}{N^\rho_{0,L}}\sum_{k=1}^{N^\rho_{0,L}} \psi(\tilde{X}^{m^0}_{T,0,k}(\widehat{\lambda}_0))\mathcal{E}^-(\tilde{W}_{0,k},\widehat{\lambda}_0) - \mathbb{E}[\psi(X^{m^0}_{T,0})] \xrightarrow[L\to+\infty]{a.s.} 0.$$

Then, it suffices to prove that the remaining terms in (5.11) tend to 0 with L. Let $\mathcal{V}$ be a compact neighbourhood of λ^*.

$$\sum_{\ell=1}^{L}\frac{1}{N^\rho_{\ell,L}}\left(\sum_{k=1}^{N^\rho_{\ell,L}}\left(\psi(\tilde X^{m^\ell}_{T,\ell,k}(\widehat\lambda_\ell))-\psi(\tilde X^{m^{\ell-1}}_{T,\ell,k}(\widehat\lambda_\ell))\right)\mathcal{E}^-(\tilde W_{\ell,k},\widehat\lambda_\ell)-\mathbb{E}\left[\psi(\tilde X^{m^\ell}_{T,\ell})-\psi(\tilde X^{m^{\ell-1}}_{T,\ell})\right]\right)=$$
$$\sum_{\ell=1}^{L}\frac{1}{N^\rho_{\ell,L}}\left(\sum_{k=1}^{N^\rho_{\ell,L}}\left(\psi(\tilde X^{m^\ell}_{T,\ell,k}(\widehat\lambda_\ell))-\psi(\tilde X^{m^{\ell-1}}_{T,\ell,k}(\widehat\lambda_\ell))\right)\mathcal{E}^-(\tilde W_{\ell,k},\widehat\lambda_\ell)-\mathbb{E}\left[\psi(\tilde X^{m^\ell}_{T,\ell})-\psi(\tilde X^{m^{\ell-1}}_{T,\ell})\right]\right)1_{\{\widehat\lambda_\ell\in\mathcal{V}\}}$$
$$+\sum_{\ell=1}^{L}\frac{1}{N^\rho_{\ell,L}}\left(\sum_{k=1}^{N^\rho_{\ell,L}}\left(\psi(\tilde X^{m^\ell}_{T,\ell,k}(\widehat\lambda_\ell))-\psi(\tilde X^{m^{\ell-1}}_{T,\ell,k}(\widehat\lambda_\ell))\right)\mathcal{E}^-(\tilde W_{\ell,k},\widehat\lambda_\ell)-\mathbb{E}\left[\psi(\tilde X^{m^\ell}_{T,\ell}-\psi(\tilde X^{m^{\ell-1}}_{T,\ell})\right]\right)1_{\{\widehat\lambda_\ell\notin\mathcal{V}\}}$$

For ℓ large enough (although random), $1_{\{\widehat\lambda_\ell\notin\mathcal{V}\}}=0$. Hence, the second term in the above equation tends to 0 a.s. when L goes to infinity. It remains to prove that the first term also converges to zero. To do so, we will apply Proposition 6.1 to the sequence

$$\begin{aligned}Y_{\ell,q}=&q\frac{1}{N^\rho_{\ell,q}}\left(\sum_{k=1}^{N^\rho_{\ell,q}}\left(\psi(\tilde X^{m^\ell}_{T,\ell,k}(\widehat\lambda_\ell))-\psi(\tilde X^{m^{\ell-1}}_{T,\ell,k}(\widehat\lambda_\ell))\right)\mathcal{E}^-(\tilde W_{\ell,k},\widehat\lambda_\ell)\right.\\&\left.-\mathbb{E}\left[\left(\psi(\tilde X^{m^\ell}_{T,\ell})-\psi(\tilde X^{m^{\ell-1}}_{T,\ell})\right)\right]\right)1_{\{\widehat\lambda_\ell\in\mathcal{V}\}}\end{aligned}$$

and $\overline{Y}_{L,q}=\frac{1}{L}\sum_{\ell=1}^{L}Y_{\ell,q}$. Note that $\mathbb{E}[Y_{\ell,q}]=0$ for all ℓ and q. Since the samples used in the different levels are independent and the $\hat\lambda_\ell$'s are independent of the filtration $\tilde{\mathcal{G}}$, we can write

$$\mathbb{E}\left[\left|\overline{Y}_{L,q}\right|^2\right]=\frac{1}{L^2}\mathbb{E}\left[\mathbb{E}\left[\left|\sum_{\ell=1}^{L}Y_{\ell,q}\right|^2\Big|\mathcal{G}\right]\right]=\frac{1}{L^2}\sum_{\ell=1}^{L}\mathbb{E}\left[|Y_{\ell,q}|^2\right]. \tag{5.12}$$

Using the same kind of arguments, we obtain

$$\begin{aligned}\mathbb{E}\left[|Y_{\ell,q}|^2\right]&\le q^2\frac{1}{N^\rho_{\ell,q}}\mathbb{E}\left[\left(\psi(\tilde X^{m^\ell}_{T,\ell})-\psi(\tilde X^{m^{\ell-1}}_{T,\ell})\right)^2\mathcal{E}^+(\tilde W_\ell,\widehat\lambda_\ell)1_{\{\widehat\lambda_\ell\in\mathcal{V}\}}\right]\\&\le\frac{q^2a_\ell}{\rho(q)\sum_{k=1}^{q}a_k}\left\{m^\ell\mathbb{E}\left[\left(\psi(\tilde X^{m^\ell}_{T,\ell})-\psi(\tilde X^{m^{\ell-1}}_{T,\ell})\right)^2\mathcal{E}^+(\tilde W_\ell,\widehat\lambda_\ell)1_{\{\widehat\lambda_\ell\in\mathcal{V}\}}\right]\right\}\end{aligned}$$

From Proposition 5.2, the term into braces converges when ℓ goes to infinity. Hence, using the assumptions on the function ρ, we get

$$\sup_q\sup_\ell\mathbb{E}\left[|Y_{\ell,q}|^2\right]<+\infty. \tag{5.13}$$

By combining Equations (5.12) and (5.13), we get that $\sup_L\sup_q L\mathbb{E}\left[\left|\overline{Y}_{L,q}\right|^2\right]<+\infty$. Hence, Proposition 6.1 yields that $\overline{Y}_{L,L}$ vanishes when L goes to infinity and this ends the proof. □

Theorem 5.4. *Suppose that the assumptions of Theorem 5.1 hold and that Condition (2.3) is satisfied. Then, for $N^\rho_{\ell,L}$ given by (5.10) with $\rho(L)=m^{2\gamma L}(m-1)T$ and the sequence $(a_\ell)_\ell$ satisfying*

$$\lim_{L\to\infty}\frac{1}{\left(\sum_{\ell=1}^{L}a_\ell\right)^{p/2}}\sum_{\ell=1}^{L}a_\ell^{p/2}=0,\ \textit{for } p>2, \tag{5.14}$$

we have

$$m^{\gamma L}(Q_L(\widehat{\lambda}_0,\ldots,\widehat{\lambda}_L)-\mathbb{E}[\psi(X_T)]) \underset{L\to+\infty}{\Longrightarrow} \mathcal{N}(C_\psi(T,\gamma),v(\lambda^*))$$

where the function v is defined by (5.6).

The convergence rate does not depend on the number of samples N'_ℓ provided that they tend to infinity with ℓ.

Proof. By assumption (2.3), we have that $\lim_{L\to+\infty} m^{\gamma L}(\mathbb{E}[\psi(X_T^{m^L})-\psi(X_T)]=C_\psi(T,\gamma)$. The convergence of the first empirical mean is governed by Theorem 4.2 (see Remark 4.3) which yields

$$\left(\frac{1}{\sqrt{N^\rho_{0,L}}}\sum_{k=1}^{N^\rho_{0,L}}\psi(\tilde{X}^{m^0}_{T,0,k}(\widehat{\lambda}_0))\mathcal{E}^-(\tilde{W}_{0,k},\widehat{\lambda}_0)-\mathbb{E}[\psi(X_T^{m^0})]\right)\underset{L\to+\infty}{\Longrightarrow}\mathcal{N}(0,\sigma_0^2).$$

with $\sigma_0^2=\mathbb{E}\Big[\psi(X_T^{m^0})^2\mathcal{E}^+(W,\lambda_0^*)\Big]-(\mathbb{E}[\psi(X_T^{m^0})])^2$. Then, we deduce from the choice of the function ρ that

$$m^{\gamma L}\left(\frac{1}{N^\rho_{0,L}}\sum_{k=1}^{N^\rho_{0,L}}\psi(\tilde{X}^{m^0}_{T,0,k}(\widehat{\lambda}_0))\mathcal{E}^-(\tilde{W}_{0,k},\widehat{\lambda}_0)-\mathbb{E}[\psi(X_T^{m^0})]\right)\xrightarrow[L\to+\infty]{\mathbb{P}}0.$$

Since all the blocks are independent, it is sufficient to prove that

$$m^{\gamma L}\left(\sum_{\ell=1}^{L}\frac{1}{N^\rho_{\ell,L}}\sum_{k=1}^{N^\rho_{\ell,L}}\left(\psi(\tilde{X}^{m^\ell}_{T,\ell,k}(\widehat{\lambda}_\ell))-\psi(\tilde{X}^{m^{\ell-1}}_{T,\ell,k}(\widehat{\lambda}_\ell))\right)\mathcal{E}^-(\tilde{W}_{\ell,k},\widehat{\lambda}_\ell)-\mathbb{E}[\psi(X_T^n)]\right)\underset{L\to+\infty}{\Longrightarrow}\mathcal{N}(0,v(\lambda^*)).$$

To do so, we introduce the $(\mathcal{G}^\sharp_l)_{l\ge1}$-martingale array $(Y^n_l)_{l\ge1}$ defined by

$$Y^n_l\triangleq\sum_{\ell=1}^{l}\frac{m^{\gamma L}}{N^\rho_{\ell,L}}\sum_{i=1}^{N^\rho_{\ell,L}}\left[\left(\psi(\tilde{X}^{m^\ell}_{T,\ell,i}(\widehat{\lambda}_\ell))-\psi(\tilde{X}^{m^{\ell-1}}_{T,\ell,i}(\widehat{\lambda}_\ell))\right)\mathcal{E}^-(\tilde{W}_{\ell,i},\widehat{\lambda}_\ell)-\mathbb{E}\left[\psi(\tilde{X}^{m^\ell}_T)-\psi(\tilde{X}^{m^{\ell-1}}_T)\right]\right],$$

so $\mathbb{E}[Y^n_l]=0$ for all l,n. According to Theorem A.1, we need to study the asymptotic behaviors of the two quantities

$$\langle Y^n\rangle_L=\sum_{\ell=1}^{L}\mathbb{E}\left[|Y^n_\ell-Y^n_{\ell-1}|^2\Big|\mathcal{G}^\sharp_{\ell-1}\right]\ \text{and}\ \sum_{\ell=1}^{L}\mathbb{E}\left[|Y^n_\ell-Y^n_{\ell-1}|^p|\mathcal{G}^\sharp_{\ell-1}\right],\quad\text{for } p>2 \text{ as } n\to\infty.$$

Note that $\widehat{\lambda}_\ell$ is $\mathcal{G}^\sharp_{\ell-1}$–measurable and for any $\lambda\in\mathbb{R}^q$ the variables $(\tilde{X}^{m^\ell}_{T,\ell,i}(\lambda),\tilde{X}^{m^{\ell-1}}_{T,\ell,i}(\lambda))_{1\le i\le N_l}$ are independent of $\mathcal{G}^\sharp_{\ell-1}$, then using (5.10) with $\rho(L)=m^{2\gamma L}(m-1)T$, we rewrite the first quantity as follows

$$\langle Y^n\rangle_L=\frac{1}{\sum_{\ell=1}^L a_\ell}\sum_{\ell=1}^{L}a_\ell\left[v_\ell(\widehat{\lambda}_\ell)-\Xi_\ell^2\right]$$

with v_ℓ defined by (5.1) and Ξ_ℓ defined by (5.2). Let $\mathcal{V}$ be a compact neighbourhood of λ^*. We can write

$$\langle Y^n \rangle_L = \frac{1}{\sum_{\ell=1}^L a_\ell} \sum_{\ell=1}^L a_\ell \left[v_\ell(\widehat{\lambda}_\ell) - \Xi_\ell^2 \right] 1_{\{\widehat{\lambda}_\ell \in \mathcal{V}\}} + \frac{1}{\sum_{\ell=1}^L a_\ell} \sum_{\ell=1}^L a_\ell \left[v_\ell(\widehat{\lambda}_\ell) - \Xi_\ell^2 \right] 1_{\{\widehat{\lambda}_\ell \notin \mathcal{V}\}} \tag{5.15}$$

From Proposition 5.2, we know that $\Xi_\ell \xrightarrow[\ell\to\infty]{} \mathbb{E}[\nabla\psi(X_T).U_T] = 0$, where the last equality is a straightforward consequence of [27, Proposition 2.1]. From Proposition 5.2, we know that the sequence of fucntions v_ℓ converges pointwise to v defined by (5.6). Moreover, we can easily prove that this convergence is locally uniform. Hence, by the convergence of $\widehat{\lambda}_\ell$ to λ^* (see Theorem 5.1), we deduce that $v_\ell(\widehat{\lambda}_\ell)1_{\{\widehat{\lambda}_\ell \in \mathcal{V}\}}$ converges to $v(\lambda^*)$ when $\ell \to +\infty$. Moreover, for ℓ large enough (although random), $1_{\{\widehat{\lambda}_\ell \notin \mathcal{V}\}} = 0$.

Thus, we deduce from the Toeplitz lemma that $\langle Y^n \rangle_L \xrightarrow[L\to\infty]{} v(\lambda^*)$ a.s. Using Burkholder's inequality and Jensen's inequalty together with the assumptions on ψ and Property ($\mathcal{P}$), we obtain that for any $p > 2$, there exists $C_p > 0$ such that

$$\sum_{\ell=1}^L \mathbb{E}\left[|Y_\ell^n - Y_{\ell-1}^n|^p | \mathcal{G}_{\ell-1}^\sharp \right] \le \frac{C_p}{\left(\sum_{\ell=1}^L a_\ell\right)^{p/2}} \sum_{\ell=1}^L a_\ell^{p/2} \xrightarrow[L\to\infty]{} 0$$

where the convergence to zero is ensured by (5.14). Consequently, we can apply Theorem A.1 to achieve the proof. $\square$

Remark 5.5. *As usual, one can rescale $m^{\gamma L}(Q_L(\widehat{\lambda}_0, \dots, \widehat{\lambda}_L) - \mathbb{E}[\psi(X_T)])$ by an estimator of $v(\lambda^*)$ to obtain a central limit theorem with variance 1. Thanks to Theorem 5.1, we know that $v_{\ell,N_\ell}(\hat{\lambda}_\ell)$ is a convergent estimator of $v(\lambda^*)$ and we can easily deduce from the proof of Theorem 5.4 that under its assumptions*

$$m^{2\gamma L} \left\{ \frac{1}{N_{0,L}^\rho} \left(\frac{1}{N_{0,L}^\rho} \sum_{k=1}^{N_{0,L}^\rho} (\psi(\tilde{X}_T^{m^0}) \mathcal{E}^+(\tilde{W}_{0,k}, \lambda_0))^2 - \left(\frac{1}{N_{0,L}^\rho} \sum_{k=1}^{N_{0,L}^\rho} \psi(\tilde{X}_T^{m^0}) \mathcal{E}^+(\tilde{W}_{0,k}, \lambda_0) \right)^2 \right) \right.$$
$$\left. + \sum_{\ell=1}^L N_\ell^{-1} \frac{(m-1)T}{m^\ell} \left(\tilde{v}_{\ell,N_\ell}(\lambda_\ell) - \tilde{\Xi}_{\ell,N_\ell}^2 \right) \right\} \xrightarrow[L\to+\infty]{} v(\lambda^*).$$

Note the quantities v_{ℓ,N_l} and Ξ_{N_ℓ} are defined as in Equations (5.3) and (5.4) but using the tilde sample paths $(\tilde{X}_{\ell,k})$ and $(\tilde{W}_{\ell,k})$. The term into braces, which can be computed online during the multilevel Monte Carlo procedure, can be used to build confidence intervals. Any convergent estimator of $v(\lambda^*)$ could of course be used, but this one has the advantage to correspond to the true variance of the multilevel Monte Carlo estimator for any finite number of levels L and not only asymptotically.

6 Strong law of large numbers for doubly indexed sequences

In this section, we prove two corner stone results used in the convergence of the multilevel approach. We tackle the convergence of empirical averages of doubly indexed random sequences when both indices tend to infinity together.

Proposition 6.1. *Let $(X_{n,m})_{n,m}$ be a doubly indexed sequence of vector valued random variables such that for all n, $\mathbb{E}[X_{n,m}] = x_m$ with $\lim_{m\to+\infty} x_m = x$. We define $\overline{X}_{n,m} = \frac{1}{n}\sum_{i=1}^n X_{i,m}$. Assume that the two following assumptions are satisfied*

($\mathcal{H}$1) i. $\sup_n \sup_m n\mathrm{Var}\left(\overline{X}_{n,m}\right) < +\infty$.

ii. $\sup_n \sup_m \mathrm{Var}\left(X_{n,m}\right) < +\infty$.

Then, for all increasing functions $\rho : \mathbb{N} \to \mathbb{N}$, $\overline{X}_{n,\rho(n)} \xrightarrow[n\to+\infty]{} x$ a.s. and in $\mathbb{L}^2$.

From this proposition, one can easily deduce the following corollary by extracting a bespoke subsequence

Corollary 6.2. *Assume that $(X_{i,m})_{i,m}$ be a doubly indexed sequence of vector valued random variables satisfying the assumptions of Proposition 6.1. Then, for any strictly increasing function $\xi : \mathbb{N} \to \mathbb{N}$, $\overline{X}_{\xi(n),n} \xrightarrow[n\to+\infty]{} x$ a.s. and in $\mathbb{L}^2$.*

Proof of Proposition 6.1. The proof of this result closely mimics the one of [34, Theorem IV.1.1]. We introduce the sequence $(Y_{i,m})_{i,m}$ defined by $Y_{i,m} = X_{i,m} - x_m$, which satisfies $\mathbb{E}[Y_{i,m}] = 0$. As $\lim_{m\to\infty} x_m = x$, it is sufficient to prove that $\overline{Y}_{n,\rho(n)} \xrightarrow[n\to+\infty]{} 0$ a.s.

Condition rbar]($\mathcal{H}$1)-i implies the $\mathbb{L}^2$ convergence to 0. We introduce the sequence $(Z_{n,m})_n$ defined by $Z_{n,m} = \sup\{\left|\bar{Y}_{k,m}\right| \ : \ n^2 \le k < (n+1)^2\}$. Let k be such that $n^2 \le k < (n+1)^2$, then

$$
\left|\bar{Y}_{k,m}\right| \le n^{-2}\left(n^2\left|\bar{Y}_{n^2,m}\right| + \sum_{i=n^2+1}^{k} |Y_{i,m}|\right),
$$

$$
Z_{n,m} \le \left|\bar{Y}_{n^2,m}\right| + \frac{1}{n^2}\sum_{i=n^2+1}^{(n+1)^2} |Y_{i,m}|.
$$

Then,

$$
\mathbb{E}[Z_{n,m}^2] \le \mathbb{E}[\bar{Y}_{n^2,m}^2] + \sum_{i=n^2+1}^{(n+1)^2}\left(\frac{\mathbb{E}[|Y_{i,m}|^2]}{n^4} + 2\frac{\mathbb{E}[\left|\bar{Y}_{n^2,m}\right| |Y_{i,m}|]}{n^2}\right) + 2\sum_{i,j=n^2+1;i\ne j}^{(n+1)^2} \frac{\mathbb{E}[|Y_{j,m}|\,|Y_{i,m}|]}{n^4}.
$$

Let $\kappa > 0$ denote the maximum of the upper bounds involved in Assumption ln2]($\mathcal{H}$1). Using the Cauchy Schwartz inequality, we get

$$
\begin{aligned}
\mathbb{E}[Z_{n,m}^2] &\le \frac{\kappa}{n^2} + \frac{\kappa((n+1)^2 - n^2)}{n^4} + 2\frac{\kappa^2((n+1)^2 - n^2)}{n^3} + 2\frac{\kappa^2((n+1)^2 - n^2)^2}{n^4} \\
&\le \frac{\kappa}{n^2} + \frac{\kappa(2n+1)}{n^4} + 2\frac{\kappa^2(2n+1)}{n^3} + 2\frac{\kappa^2(2n+1)^2}{n^4}
\end{aligned}
$$

Hence, for any function $\rho : \mathbb{N} \to \mathbb{N}$, $\mathbb{E}[Z_{n,\rho(n)}^2] \le Cn^{-2}$ where $C > 0$ is a constant independent of ρ. Therefore, we have $\mathbb{P}(Z_{n,\rho(n)} \ge n^{-1/4}) \le Cn^{-3/2}$. This inequality implies using the Borel Cantelli Lemma that, for n large enough $Z_{n,\rho(n)} \le n^{-1/4}$ a.s. which yields the a.s. convergence to 0. □

Proposition 6.3. *Let $(F_{n,m})_{n,m}$ be a doubly indexed sequence of random variables with values in the set of continuous functions, ie. for all n, m, $F_{n,m} : \Omega \longrightarrow \mathcal{C}^0(\mathbb{R}^d, \mathbb{R}^p)$ and for all n. Moreover, we assume that there exists a sequence of functions f_m satisfying $\mathbb{E}[F_{n,m}] = f_m$ for all m. We define $\overline{F}_{n,m} = \frac{1}{n}\sum_{i=1}^n F_{i,m}$. Assume that the two following assumptions are satisfied*

($\mathcal{H}2$) *One of the following criteria holds*

- *i.* *The sequence $(f_m)_m$ converges pointwise to some continuous function f.*
- *ii.* *The sequence $(f_m)_m$ converges locally uniformly to some function f.*

($\mathcal{H}3$) *For any compact set $W \subset \mathbb{R}^d$,*

- *i.* $\sup_n \sup_m n\mathrm{Var}\left(\sup_{x\in W}\left|\overline{F}_{n,m}(x)\right|\right) < +\infty$.
- *ii.* $\sup_n \sup_m \mathrm{Var}\left(\sup_{x\in W}|F_{n,m}(x)|\right) < +\infty$.

($\mathcal{H}4$) *For all $y \in \mathbb{R}^d$, $\lim_{\delta\to 0}\sup_n \sup_m \mathbb{E}\left[\sup_{|x-y|\le\delta}|F_{n,m}(x) - F_{n,m}(y)|\right] = 0$.*

Then, for all functions $\rho : \mathbb{N} \to \mathbb{N}$, the sequence of random functions $\overline{F}_{n,\rho(n)}$ converges a.s. locally uniformly to the locally continuous function f.

Remark 6.4.

- *When for every fixed m, the sequence $(F_{n,m})_m$ is independent and identically distributed, Assumption ln2-sup-u]($\mathcal{H}4$) is ensured by*

$$\forall\, y \in \mathbb{R}^d,\ \lim_{\delta\to 0}\limsup_m \mathbb{E}\left[\sup_{|x-y|\le\delta}|F_{1,m}(x) - F_{1,m}(y)|\right] = 0$$

and Assumption r-u]($\mathcal{H}3$)-ii implies rbar-u]($\mathcal{H}3$)-i.

- *As in Corollary 6.2, for any strictly increasing function $\xi : \mathbb{N} \to \mathbb{N}$, the sequence $\overline{F}_{\xi(n),n}$ converges a.s. locally uniformly to the locally continuous function f.*

Proof. We can apply Proposition 6.1, to deduce that a.s. $\overline{F}_{n,\rho(n)}$ converges pointwise to the function f. If we do not already know that f is continuous, then thanks to r-u]($\mathcal{H}3$)-*ii*, we can apply Lebesgue's theorem to deduce that the functions f_m are continuous. The uniform convergence of the sequence f_m to f (see m-unif]($\mathcal{H}2$)-*ii*) proves that the function f is continuous.

Let W be a compact set of $\mathbb{R}^d$, we can cover W with a finite number K of open balls W_k with centers $(x_k)_k$ and radiuses $(r_k)_k$, i.e. $W_k = B(x_k, r_k)$ and $W = \cup_{k=1}^K W_k$. We want to prove that

$$\sup_{x\in W}\left|\overline{F}_{n,\rho(n)}(x) - f(x)\right| \xrightarrow[n\to+\infty]{a.s.} 0.$$

We write

$$\sup_{x\in W}\left|\overline{F}_{n,\rho(n)}(x) - f(x)\right| = \sum_{k=1}^K \sup_{x\in W_k}\left|\overline{F}_{n,\rho(n)}(x) - f(x)\right|. \tag{6.1}$$

We split each term

$$\sup_{x\in W_k}\left|\overline{F}_{n,\rho(n)}(x)-f(x)\right| = \sup_{x\in W_k}\left|\overline{F}_{n,\rho(n)}(x)-\overline{F}_{n,\rho(n)}(x_k)\right| + \sup_{x\in W_k}|f(x)-f(x_k)| + \left|\overline{F}_{n,\rho(n)}(x_k)-f(x_k)\right| \tag{6.2}$$

Let $\varepsilon > 0$. The idea is to choose the radiuses r_k small enough to ensure that each term is controlled by a function of ε. Now, we make the idea precise. For all $k = 1, \dots, K$, the last term term can be made smaller that ε/K for n larger that some N_k using the pointwise convergence. For all $n \geq \max_{k\leq K} N_k$, and all $1 \leq k \leq K$,

$$\left|\overline{F}_{n,\rho(n)}(x_k)-f(x_k)\right| \leq \varepsilon/K.$$

The function f being continuous, it is uniformly continuous on every W_k. If we choose the W_k such that their radiuses are small enough (we may need to increase K), we can ensure that for all $1 \leq k \leq K$

$$\sup_{x\in W_k}|f(x)-f(x_k)| \leq \varepsilon/K.$$

The first term on the r.h.s of (6.2) deserves more attention

$$\sup_{x\in W_k}\left|\overline{F}_{n,\rho(n)}(x)-\overline{F}_{n,\rho(n)}(x_k)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\sup_{x\in W_k}\left|F_{i,\rho(n)}(x)-F_{i,\rho(n)}(x_k)\right|. \tag{6.3}$$

Now, for every $1 \leq k \leq K$, we want to apply Proposition 6.1 to the sequence of random variables $\left(\sup_{x\in W_k}|F_{n,m}(x)-F_{n,m}(x_k)|\right)_{n,m}$. Assumption ln2]($\mathcal{H}1$) is clearly satisfied using Minkowski's inequality.

Let us define the sequence $(Y_{n,m})_{n,m}$ by

$$Y_{n,m} = \sup_{x\in W_k}|F_{n,m}(x)-F_{n,m}(x_k)| - \mathbb{E}\left[\sup_{x\in W_k}|F_{n,m}(x)-F_{n,m}(x_k)|\right],$$

satisfying $\mathbb{E}[Y_{n,m}] = 0$ and the assumptions of Proposition 6.1. Hence, it yields that

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\sup_{x\in W_k}\left|F_{i,\rho(n)}(x)-F_{i,\rho(n)}(x_k)\right| - \mathbb{E}\left[\sup_{x\in W_k}\left|F_{n,\rho(n)}(x)-F_{n,\rho(n)}(x_k)\right|\right] = 0. \tag{6.4}$$

From ln2-sup-u]($\mathcal{H}4$), we know that if the W_k are chosen small enough,

$$\sup_n \mathbb{E}\left[\sup_{x\in W_k}\left|F_{n,\rho(n)}(x)-F_{n,\rho(n)}(x_k)\right|\right] \leq \varepsilon/K. \tag{6.5}$$

Then, Combining with (6.3), (6.4) and (6.5) yields that

$$\sup_{x\in W_k}\left|\overline{F}_{n,\rho(n)}(x)-\overline{F}_{n,\rho(n)}(x_k)\right| \leq \varepsilon/K.$$

Going back to Equations (6.1) and (6.2), we deduce that for n large enough

$$\sup_{x\in W}\left|\bar{F}_{n,\rho(n)}(x)-f(x)\right| \leq 3\varepsilon,$$

which achieves the proof. □

7 Numerical experiments

7.1 Practical implementation

Our approach cleverly mixes the famous multilevel Monte Carlo technique with importance sampling to reduce the variance. A classical approach would have been to consider the multilevel approximation of $\mathbb{E}\left[\psi(X_T(\theta))\mathrm{e}^{-\theta\cdot W_T-\frac{1}{2}|\theta|^2T}\right]$ while choosing the value of θ which minimizes the variance of the central limit theorem for multilevel Monte Carlo (see [7]). This asymptotic variances involves both $\nabla\psi$ and the process U given in (5.5). Hence, a classical approach to importance sampling for multilevel Monte Carlo would require extra knowledge than the function ψ and the underlying process X, thus precluding any kind of automation.

We have chosen a completely different approach allowing for one importance sampling parameter per level, which enables us to treat each level independently of the others. In each level, we use a sample average approximation as in [26] to compute the optimal importance sampling parameter defined as the one minimizing the variance of the current level. From Theorem 5.4, we know that this approach is optimal in the sense that our multilevel estimator $Q_L(\hat{\lambda}_0,\dots,\hat{\lambda}_L)$ satisfies a central limit theorem with a limiting variance given by $\inf v$ where v defined by (5.6) is the variance of the standard multilevel Monte Carlo estimator. We managed to provide an algorithm reaching the optimal limiting variance without computing $\nabla\psi$ nor the process U, hence our approach can be made fully automatic. Our overall algorithm is described in Algorithm 1.

The minimization step (items 2 and 4 in Algorithm 1) is performed using a Newton algorithm. Unlike what happens in a classical Monte Carlo method in which a new sample is drawn at each iteration, here all the samples must be stored since the same random variables are used in all the iterations of the Newton procedure. This feature is specific to the optimisation step and may make the algorithm highly memory demanding as soon as the numbers N'_ℓ become large. As the parameter λ is not involved in the function ψ, all the quantities $\psi(X^{m^\ell}_{T,\ell,k})-\psi(X^{m^{\ell-1}}_{T,\ell,k})$ for $k=1,\dots,N_\ell$ can be precomputed before starting the minimization algorithm, which enables us to save a lot of computational time. The efficiency of Newton's algorithm very much depends on the convexity of the v_{ℓ,N_ℓ} functions. As already pointed out in [26], the smallest eigenvalue of the Hessian matrix $\nabla^2 v_{\ell,N'_\ell}$ is basically $\frac{T}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}\left|\psi(X^{m^\ell}_{T,k})-\psi(X^{m^{\ell-1}}_{T,k})\right|^2\mathcal{E}^+(W_k,\lambda)$, which can become extremely small and then conflicts with the will to have the strongest possible convexity in order to speed up Newton's algorithm. This difficulty is circumvented by noticing that $\hat{\lambda}_\ell$ can be interpreted as the root of

$$\nabla u_{\ell,N'_\ell}(\lambda)=\lambda T-\frac{\frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}W_{k,\ell,T}\left|\psi(X^{m^\ell}_{T,k})-\psi(X^{m^{\ell-1}}_{T,k})\right|^2\mathrm{e}^{-\lambda\cdot W_{k,\ell,T}}}{\frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}\left|\psi(X^{m^\ell}_{T,k})-\psi(X^{m^{\ell-1}}_{T,k})\right|^2\mathrm{e}^{-\lambda\cdot W_{k,\ell,T}}} \tag{7.1}$$

with

$$u_{\ell,N'_\ell}(\lambda)=\frac{|\lambda|^2T}{2}+\log\left(\frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}\left|\psi(X^{m^\ell}_{T,k})-\psi(X^{m^{\ell-1}}_{T,k})\right|^2\mathrm{e}^{-\lambda\cdot W_{k,\ell,T}}\right).$$

Algorithm 1 Multilevel Importance Sampling (MLIS)

1. Generate $X^{m^0}_{T,0,1},\dots,X^{m^0}_{T,0,N'_0}$ i.i.d. samples following the law of $X^{m^0}_T$ independently of the other blocks.
2. Compute the minimizer $\hat{\lambda}_0$ of u_{0,N'_0} by solving $\nabla u_{0,N'_0}(\hat{\lambda}_0)=0$.

for $\ell=1:L$ **do**

3. Generate $(X^{m^\ell}_{T,\ell,1},X^{m^{\ell-1}}_{T,\ell,1}),\dots,(X^{m^\ell}_{T,\ell,N'_\ell},X^{m^{\ell-1}}_{T,\ell,N'_\ell})$ i.i.d. samples following the law of $(X^{m^\ell}_T,X^{m^{\ell-1}}_T)$ independently of the other blocks.
4. Compute the minimizer $\hat{\lambda}_\ell$ of of u_{ℓ,N'_ℓ} by solving $\nabla u_{\ell,N'_\ell}(\hat{\lambda}_\ell)=0$.

end for

5. Conditionally on $\hat{\lambda}_0$, generate $\tilde{X}^{m^0}_{T,0,1}(\hat{\lambda}_0),\dots,\tilde{X}^{m^0}_{T,0,N_0}(\hat{\lambda}_0)$ i.i.d. samples following the law of $X^{m^0}_T(\hat{\lambda}_0)$ independently of the other blocks. The tilde and non tilde quantities are conditionally independent.

for $\ell=1:L$ **do**

6. Conditionally on $\hat{\lambda}_\ell$, generate $(\tilde{X}^{m^\ell}_{T,\ell,1}(\hat{\lambda}_\ell),\tilde{X}^{m^{\ell-1}}_{T,\ell,1}(\hat{\lambda}_\ell)),\dots,(\tilde{X}^{m^\ell}_{T,\ell,N_\ell}(\hat{\lambda}_\ell),\tilde{X}^{m^{\ell-1}}_{T,\ell,N_\ell}(\hat{\lambda}_\ell))$ i.i.d. samples following the law of $(X^{m^\ell}_T(\hat{\lambda}_\ell),X^{m^{\ell-1}}_T(\hat{\lambda}_\ell))$ independently of the other blocks. The tilde and non tilde quantities are conditionally independent.

end for

7. Compute the multilevel importance sampling estimator

$$\begin{aligned}Q_L(\hat{\lambda}_0,\dots,\hat{\lambda}_L)=&\frac{1}{N_0}\sum_{k=1}^{N_0}\psi(\tilde{X}^{m^0}_{T,0,k}(\hat{\lambda}_0))\mathcal{E}^-(\tilde{W}_{0,k},\hat{\lambda}_0)\\&+\sum_{\ell=1}^{L}\frac{1}{N_\ell}\sum_{k=1}^{N_\ell}\Big(\psi(\tilde{X}^{m^\ell}_{T,\ell,k}(\hat{\lambda}_\ell))-\psi(\tilde{X}^{m^{\ell-1}}_{T,\ell,k}(\hat{\lambda}_\ell))\Big)\mathcal{E}^-(\tilde{W}_{\ell,k},\hat{\lambda}_\ell).\end{aligned}$$

The Hessian matrix of u_{ℓ,N'_ℓ} is given by

$$\begin{aligned}\nabla^2 u_{\ell,N'_\ell}(\lambda) =& TI + \frac{\frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}W_{k,\ell,T}(W_{k,\ell,T})^*\left|\psi(X^{m^\ell}_{T,k})-\psi(X^{m^{\ell-1}}_{T,k})\right|^2 \mathrm{e}^{-\lambda\cdot W_{k,\ell,T}}}{\frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}\left|\psi(X^{m^\ell}_{T,k})-\psi(X^{m^{\ell-1}}_{T,k})\right|^2 \mathrm{e}^{-\lambda\cdot W_{k,\ell,T}}} \\ &- \frac{\left(\frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}W_{k,\ell,T}\left|\psi(X^{m^\ell}_{T,k})-\psi(X^{m^{\ell-1}}_{T,k})\right|^2 \mathrm{e}^{-\lambda\cdot W_{k,\ell,T}}\right)}{\frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}\left|\psi(X^{m^\ell}_{T,k})-\psi(X^{m^{\ell-1}}_{T,k})\right|^2 \mathrm{e}^{-\lambda\cdot W_{k,\ell,T}}} \\ &\quad \frac{\left(\frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}W_{k,\ell,T}\left|\psi(X^{m^\ell}_{T,k})-\psi(X^{m^{\ell-1}}_{T,k})\right|^2 \mathrm{e}^{-\lambda\cdot W_{k,\ell,T}}\right)^*}{\frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}\left|\psi(X^{m^\ell}_{T,k})-\psi(X^{m^{\ell-1}}_{T,k})\right|^2 \mathrm{e}^{-\lambda\cdot W_{k,\ell,T}}}. \end{aligned} \tag{7.2}$$

From the Cauchy Schwartz inequality, it is clear that $\nabla^2 u_{\ell,N'_\ell}(\lambda)$ is lower bounded by TI, where the inequality is to be understood in the sense of the order on symmetric matrices.

Complexity analysis. In this paragraph, we focus on the impact of the number of levels L on the overall computational time of our algorithm. The computational cost of the standard multilevel estimator is proportional to

$$C_{ML} = \sum_{\ell=0}^{L} N_\ell m^\ell = m^{2L+1}L^2.$$

The global cost of our algorithm writes as the sum of the cost of the computation of the $(\hat{\lambda}_\ell)_\ell$ and of the standard multilevel estimator

$$C_{MLIS} = \sum_{\ell=0}^{L} N'_\ell(m^\ell + 3K_\ell) + \sum_{\ell=0}^{L} N_\ell m^\ell$$

where K_ℓ is the number of iterations of Newton's algorithm to approximate $\hat{\lambda}_\ell$ and the factor 3 corresponds to the fact that building $\nabla u_{\ell,N'_\ell}$ and $\nabla^2 u_{\ell,N'_\ell}$ basically boils down to three Monte Carlo summations. In practice, $K_\ell \le 5$ as the problem is strongly convex. Because the same random variables are used at each iteration of the optimisation step, they must be stored, which makes the memory footprint of our algorithm proportional to N'_ℓ.

So, if we choose $N'_\ell = \frac{N_\ell m^\ell}{m^\ell+15}$, the total cost of our MLIS algorithm should be roughly twice the cost of the standard multilevel estimator. This choice of N'_ℓ reduces the number of samples used to approximate the variance of the first levels compared to using directly N_ℓ. However, when L increases, N'_ℓ can become extremely large for small values of ℓ which leads to an even larger memory footprint (see Section 7.1). Not to break the scalability of the algorithm, the values of N'_ℓ have to be kept reasonable depending on the amount of memory available on the computer. For an instance, enforcing $N'_\ell \le 500000$ is reasonable on a computer with $8Gb$ of RAM. Anyway, it is crystal clear that a fairly good approximation of the variance v_ℓ is enough and running for an ultimately accurate estimator would lead to a tremendous waste of computational time. Monitoring the convergence of v_{ℓ,N'_ℓ} would really help choosing sensible values for N'_ℓ.

7.2 Experiment settings

We compare four methods in terms of their root mean squared error (RMSE): the crude Monte Carlo method (MC), the adaptive Monte Carlo method proposed in [26] (MC+IS), the Multilevel Monte Carlo method (ML) and our Importance Sampling Multilevel Monte Carlo estimator (ML+IS). We recall that the RMSE is defined by $RMSE = \sqrt{\text{Bias}^2 + \text{Variance}}$. In the computation of the bias, the true value is replaced by its multilevel Monte Carlo estimator with $L = 9$ levels, which yields a very accurate approximation. Not to mention, the CPU times showed on the graphs take into account both the time to the search for the optimal parameter and the time for the second stage Monte Carlo, be it multilevel or not.

7.3 Multidimensional Dupire's framework

We consider a $d-$dimensional local volatility model, in which the dynamics, under the risk neutral measure, of each asset S^i is supposed to be given by

$$dS_t^i = S_t^i(r\,dt + \sigma(t, S_t^i)dW_t^i), \qquad S_0 = (S_0^1, \dots, S_0^d)$$

where $W = (W^1, \dots, W^d)$, each component W^i being a standard Brownian motion with values in $\mathbb{R}$. For the numerical experiments, the covariance structure of W will be assumed to be given by $\langle W^i, W^j\rangle_t = \rho t 1_{\{i\neq j\}} + t 1_{\{i=j\}}$. We suppose that $\rho \in (-\frac{1}{d-1}, 1)$, which ensures that the matrix $C = (\rho 1_{\{i\neq j\}} + 1_{\{i=j\}})_{1\leq i,j\leq d}$ is positive definite. Let L denote the lower triangular matrix involved in the Cholesky decomposition $C = LL^*$. To simulate W on the time-grid $0 < t_1 < t_2 < \ldots < t_N$, we need $d \times N$ independent standard normal variables and set

$$\begin{pmatrix} W_{t_1} \\ W_{t_2} \\ \vdots \\ W_{t_{N-1}} \\ W_{t_N} \end{pmatrix} = \begin{pmatrix} \sqrt{t_1}L & 0 & 0 & \ldots & 0 \\ \sqrt{t_1}L & \sqrt{t_2 - t_1}L & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sqrt{t_{N-1} - t_{N-2}}L & 0 \\ \sqrt{t_1}L & \sqrt{t_2 - t_1}L & \ldots & \sqrt{t_{N-1} - t_{N-2}}L & \sqrt{t_N - t_{N-1}}L \end{pmatrix} G,$$

where G is a normal random vector in $\mathbb{R}^{d\times N}$. The maturity time and the interest rate are respectively denoted by $T > 0$ and $r > 0$. The local volatility function σ we have chosen is of the form

$$\sigma(t, x) = 0.6(1.2 - \mathrm{e}^{-0.1t}\mathrm{e}^{-0.001(x\mathrm{e}^{rt}-s)^2})\mathrm{e}^{-0.05\sqrt{t}}, \tag{7.3}$$

with $s > 0$. We know that there exists a duality between the variables (t, x) and (T, K) in Dupire's framework. Hence for formula (7.3) to make sense, one should choose s equal to the spot price of the underlying asset so that the bottom of the smile is located at the forward money. We refer to Figure 1 to have an overview of the smile.

Basket option We consider options with payoffs of the form $(\sum_{i=1}^d \omega^i S_T^i - K)_+$ where $(\omega^1, \dots, \omega^d)$ is a vector of algebraic weights. The strike value K can be taken negative to deal with Put like options. With no surprise, we can see on Figure 2 that multilevel estimators always outperform their classical Monte Carlo counterpart. The comparison for very little accurate estimators may be meaningless as it is pretty difficult to reliably measure short execution times and the empirical variance of the estimator is in this case even less accurate than the estimator itself. Note that the points on the extreme right hand side are obtained for

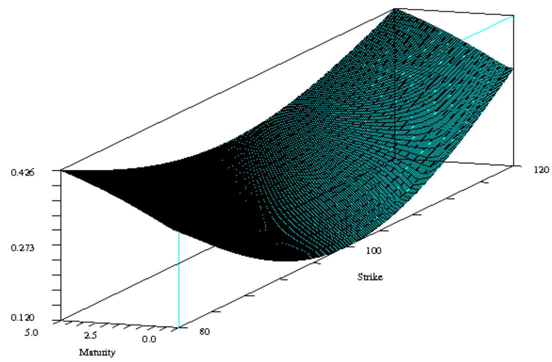

Figure 1: Local volatility function

multilevel estimators with $L = 2$, respectively for Monte Carlo estimators with 256 samples. For RMSE between 0.1 and 0.005, our MLIS estimator is 10 times faster than the standard ML estimator. When a very high accuracy is required, namely when RMSE is smaller than 0.001, the MLIS estimator remains between 3 and 4 times faster than the standard multilevel estimator, which is already a great achievement since for this level of accuracy, the ML estimator may need several dozens of minutes to yield its result.

7.4 Multidimensional Heston model

The multidimensional Heston model can be easily written by specifying on the one hand that each asset follows a 1-D Heston model and on the other hand the correlation structure between the involved Brownian motions. The asset price process $S = (S^1, \ldots, S^d)$ and the volatility process $\sigma = (\sigma^1, \ldots, \sigma^d)$ solve

$$\begin{aligned} dS_t^i &= rS_t^i dt + \sqrt{\sigma_t^i} S_t^i dB_t^i \\ d\sigma_t^i &= \kappa^i (a^i - \sigma_t^i) dt + \nu_t^i \sqrt{\sigma_t^i} (\gamma^i dB_t^i + \sqrt{1 - (\gamma^i)^2} d\tilde{B}_t^i) \end{aligned}$$

where all the components of $B = (B^1, \ldots, B^d)$ and $\tilde{B} = (\tilde{B}^1, \ldots, \tilde{B}^d)$ are real valued Brownian motions. The vectors $\kappa = (\kappa^1, \ldots, \kappa^d)$ and $a = (a^1, \ldots, a^d)$ denote respectively the reversion rate and the mean level of each volatility process, while the vector ν is the volatility of the volatility process. The vector $\bar{\gamma} = (\gamma^1, \ldots, \gamma^d)$ embodies the correlations between an asset and its volatility process, with $\gamma^i \in]-1, 1[$ for all $1 \leq i \leq d$. The vector valued processes B and $\tilde{B}$ are independent and satisfy

$$d\langle B \rangle_t = \Gamma_S \, dt \quad \text{and} \quad d\langle \tilde{B} \rangle_t = I_d \, dt$$

where we assume for our experiments that the covariance matrix Γ_S has the structure

$$\Gamma_S = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \ldots & \rho & 1 \end{pmatrix} \tag{7.4}$$

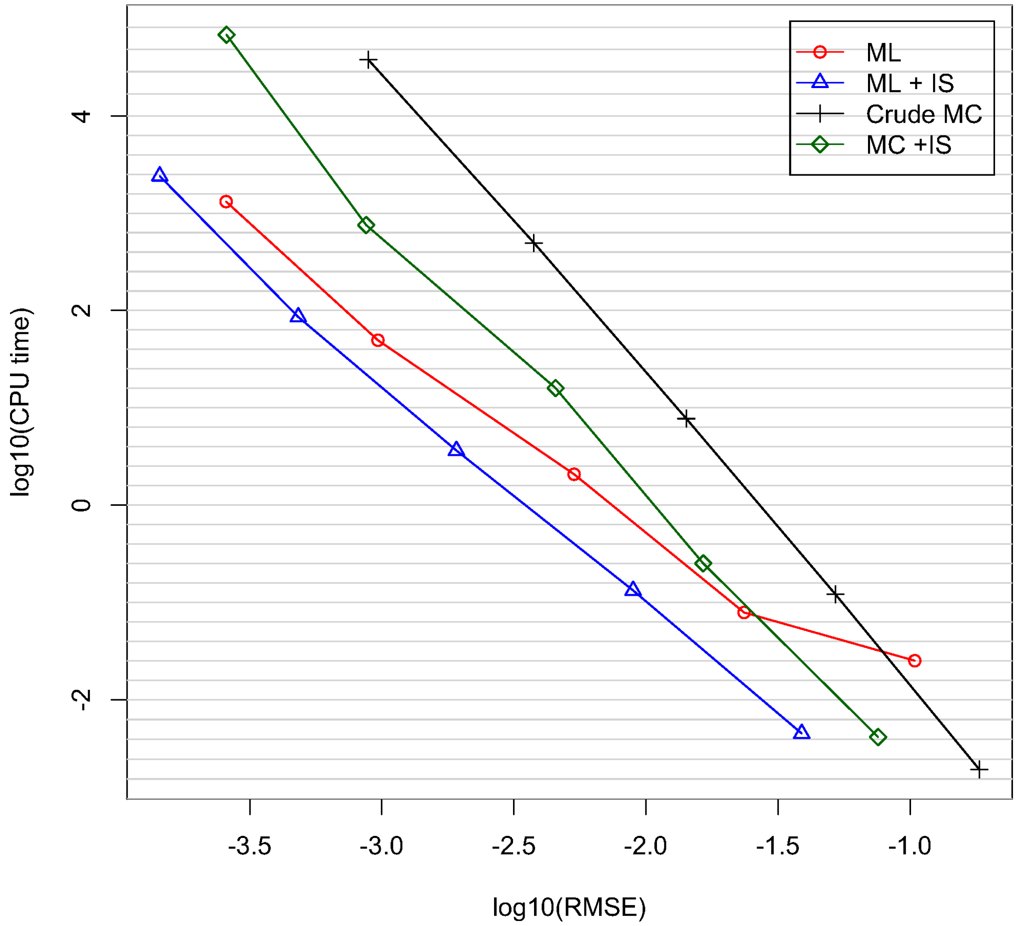

Figure 2: $\sqrt{MSE}$ vs. CPU time for a basket option in the local volatility model with $I = 5$, $r = 0.05$, $T = 1$, $S_0 = 100$, $K = 100$, $m = 4$.

with $\rho \in \left]\frac{-1}{I-1}, 1\right[$, such that the matrix Γ_S is positive definite. The processes B and $\tilde{B}$ are Wiener processes with covariance matrices given by Γ_S and I_d respectively.

For the sake of simplicity, we decided not to add any extra correlation between the components of $\tilde{B}$, hence the choice $d\langle\tilde{B}\rangle = I_d\, dt$ and we assume in the following that for all the γ^i's are equal for $1 \le i \le d$, $\gamma^i = \gamma$. The correlations between the volatilities are entirely specified by the correlations between the assets. Even though we do not aim at discussing the correlation structure of the multidimensional Heston model, we believe it is important to make precise the underlying correlation structure in the multidimensional model so that the experiments are easily reproducible.

The model can be equivalently written

$$
\begin{aligned}
dS_t^i &= rS_t^i dt + \sqrt{\sigma_t^i} S_t^i dB_t^i \\
d\sigma_t^i &= \kappa^i (a^i - \sigma_t^i) dt + \nu_t^i \sqrt{\sigma_t^i} dW_t^i
\end{aligned}
$$

where the processes W and B are Wiener processes satisfying

$$\begin{aligned} d\langle B\rangle_t &= \Gamma_S \, dt \\ d\langle B, W\rangle_t &= \gamma \Gamma_S \, dt \\ d\langle W\rangle_t &= (\gamma^2 \Gamma_S + (1-\gamma^2) I_d) \, dt. \end{aligned}$$

The process (B, W) with values in $\mathbb{R}^{2d}$ is a Wiener process with covariance matrix

$$\Gamma = \begin{pmatrix} \Gamma_S & \gamma \Gamma_S \\ \gamma \Gamma_S & \gamma^2 \Gamma_S + (1-\gamma^2) I_d \end{pmatrix}.$$

Hence, the pair of processes (B, W) can be easily simulated by applying the Cholesky factorization of Γ to a standard Brownian motion with values in $\mathbb{R}^{2d}$.

Basket Option We consider a basket option as in the local volatility model. Figure 3 looks very much the same as in the case of the local volatility model (see Figure 2). The MLIS estimator always outperforms all the ML estimator by a factor of 3 to 4. Note that for small RMSE, the computational time can go beyond several hours, hence cutting it down by two or three times represents a real improvement.

Best of option We consider options with payoffs of the form $(\max_{1 \le i \le d} S^i_T - K)_+$. The payoff of this option does obviously not satisfy the assumptions of Theorem 4.1 as the payoff of the "best of" options is not Hölder with $\alpha \ge 1$. Nonetheless, the multilevel approach beats the standard Monte Carlo technology by far (see Figure 4). Moreover, coupling importance sampling with the multilevel approach improves the accuracy. For a fixed RMSE, we can expect MLIS to be 3 faster that ML. This example shows the robustness of the method, which performs well whereas the theoretical assumptions are not satisfied.

8 Conclusion

We have presented a new estimator making the most of the recent works on multilevel Monte Carlo and on adaptive importance sampling. As expected, this new estimator outperforms the standard multilevel Monte Carlo estimator by a great deal. For a fixed accuracy measured in terms the mean squared error, the MLIS estimator is between 3 and 10 times faster that the standard multilevel Monte Carlo estimator. This efficiency of our MLIS approach could still be improved by monitoring the number of samples N'_ℓ to be used to approximate the variance v_{ℓ, N'_ℓ} in each level. Actually, we believe that there is no need to compute a too accurate approximation of this variance as a slight decrease in the accuracy of $\hat{\lambda}_\ell$ would not lead to a serious deterioration of the accuracy of the MLIS estimator but it could help to save a lot of computational time.

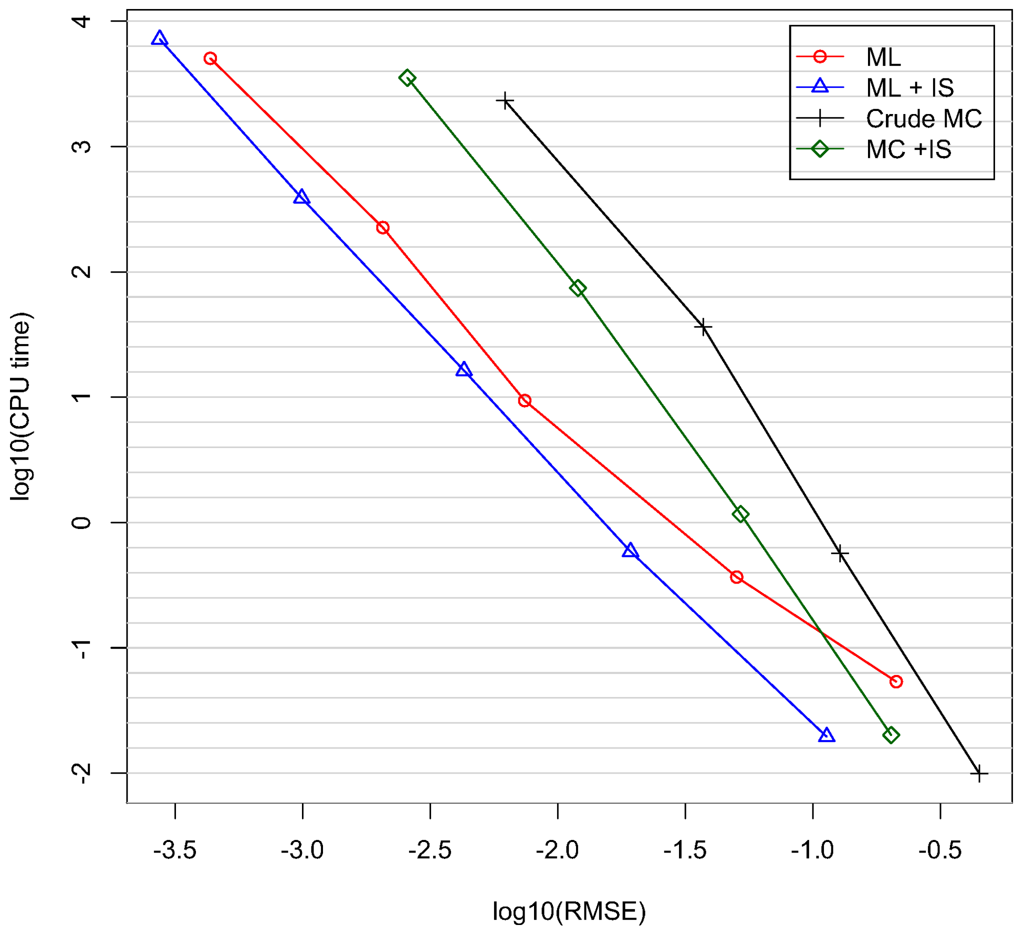

Figure 3: $\sqrt{MSE}$ vs. CPU time for a best of option in the multidimensional Heston model with $I = 10$, $r = 0.03$, $T = 1$, $S_0 = 100$, $K = 100$, $\nu = 0.01$, $\kappa = 2$, $a = 0.04$, $\gamma = -0.2$, $\rho = 0.3$ and $m = 4$.

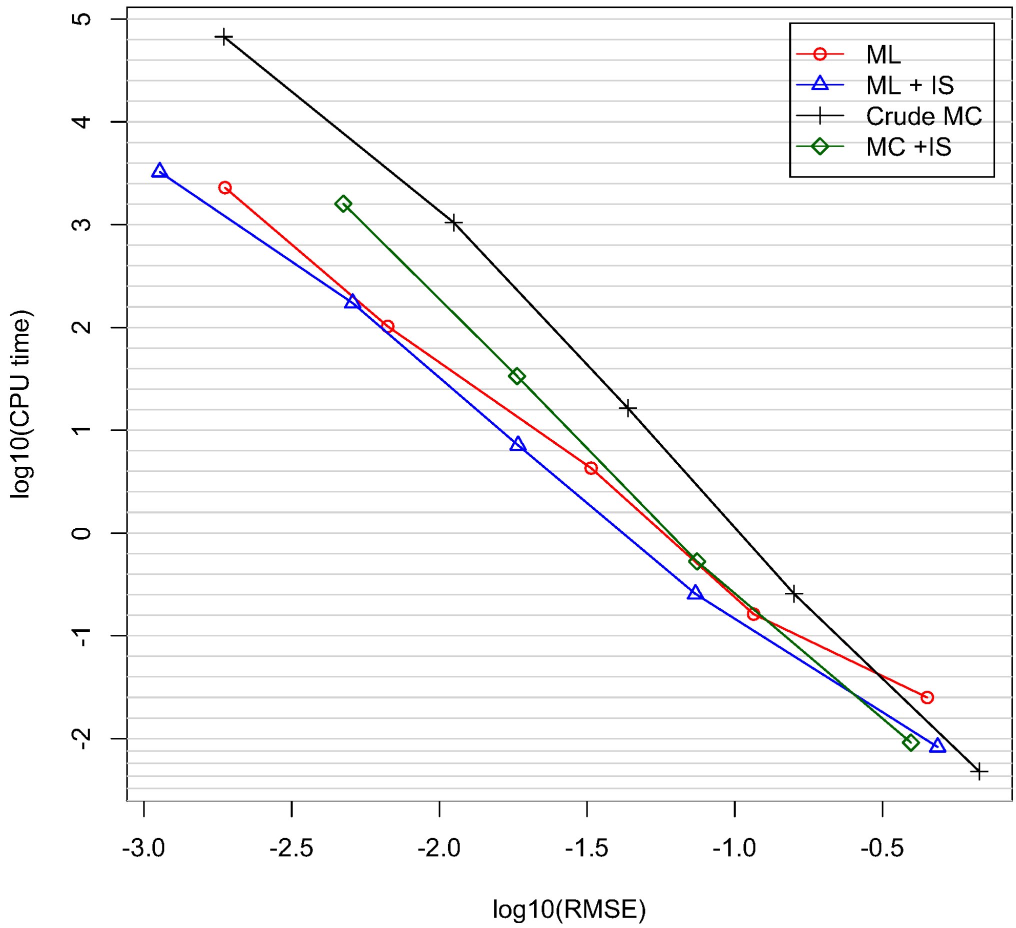

Figure 4: $\sqrt{MSE}$ vs. CPU time for a best of option in the multidimensional Heston model with $I = 5$, $r = 0.03$, $T = 1$, $S_0 = 100$, $K = 140$, $\nu = 0.25$, $\kappa = 2$, $a = 0.04$, $\gamma = 0.2$, $\rho = 0.5$ and $m = 4$.

A Auxiliary lemmas

A.1 Central limit theorems for martingale arrays

Theorem A.1 (Central limit theorem for triangular array)**.** *Suppose that* $(\Omega, \mathbb{F}, \mathbb{P})$ *is a probability space and that for each* n*, we have a filtration* $\mathbb{F}_n = (\mathcal{F}_k^n)_{k\geq 0}$*, a sequence* $k_n \longrightarrow \infty$ *as* $n \longrightarrow \infty$ *and a real vector martingale* $Y^n = (Y_k^n)_{k\geq 0}$ *adapted to* $\mathbb{F}_n$*. We make the following two assumptions.*

($\mathcal{H}$5) *i. There exists a deterministic symmetric positive semi-definite matrix* Γ*, such that*

$$\langle Y^n \rangle_{k_n} = \sum_{k=1}^{k_n} \mathbb{E}\left[|Y_k^n - Y_{k-1}^n|^2 | \mathcal{F}_{k-1}^n\right] \xrightarrow[n\to\infty]{\mathbb{P}} \Gamma.$$

ii. There exists a real number $a > 1$*, such that*

$$\sum_{k=1}^{k_n} \mathbb{E}\left[|Y_k^n - Y_{k-1}^n|^{2a} | \mathcal{F}_{k-1}^n\right] \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Then

$$Y_{k_n}^n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma) \quad \text{as } n \to \infty.$$

A.2 Asymptotic behavior of the process $\left(X^{m^\ell} - X^{m^{\ell-1}}\right)_{\ell\geq 0}$

In the following we recall some results around the stable convergence. Let Z_n be a sequence of random variables with values in a Polish space E, all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ be an extension of $(\Omega, \mathcal{F}, \mathbb{P})$, and let Z be an E-valued random variable on the extension. We say that (Z_n) converges in law to Z stably and write $Z_n \Rightarrow^{stably} Z$, if

$$\mathbb{E}(Uh(Z_n)) \to \tilde{\mathbb{E}}(Uh(Z))$$

for all $h : E \to \mathbb{R}$ bounded continuous and all bounded random variable U on $(\Omega, \mathcal{F})$. This convergence, introduced by Rényi [33] and studied by Aldous and Eagelson [1], is obviously stronger than convergence in law that we will denote here by "$\Rightarrow$". According to Section 2 of Jacod [24] and Lemma 2.1 of Jacod and Protter [25], we have the following result

Lemma A.2. *Let* V_n *and* V *be defined on* $(\Omega, \mathcal{F})$ *with values in another metric space.*

$$\text{If} \quad V_n \xrightarrow{\mathbb{P}} V, \quad Z_n \Rightarrow^{stably} Z \quad \text{then} \quad (V_n, Z_n) \Rightarrow^{stably} (V, X).$$

The following result proved by Ben Alaya and Kebaier [7, Theorem 3] is an improvement of Theorem 3.2 of Jacod and Protter [25], for the setting of Multilevel Euler scheme. More precisely, if $(X_t^{m^\ell})_{t\geq 0}$ denotes the Euler scheme with time step m^ℓ, with $m, \ell \in \mathbb{N} \setminus \{0, 1\}$ solution to (2.2), then we have the following weak convergence in the Skorohod topology.

Theorem A.3. *Assume that* b *and* σ *are* $\mathcal{C}^1$ *with linear growth then the following result holds.*

$$\text{For all } m \in \mathbb{N} \setminus \{0, 1\}, \quad \sqrt{\frac{m^\ell}{(m-1)T}} (X^{m^\ell} - X^{m^{\ell-1}}) \Rightarrow^{stably} U, \quad \text{as } \ell \to \infty,$$

with $(U_t)_{0\leq t\leq T}$ *the d-dimensional diffusion process solution to (5.5)*

References

[1] D. J. Aldous and G. K. Eagleson. On mixing and stability of limit theorems. *Ann. Probability*, 6(2):325–331, 1978. 31

[2] C. Andrieu, E. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.*, 44(1):283–312, 2005. ISSN 0363-0129. doi: 10.1137/S0363012902417267. URL http://dx.doi.org/10.1137/S0363012902417267. 2

[3] B. Arouna. Adaptative Monte Carlo method, a variance reduction technique. *Monte Carlo Methods Appl.*, 10(1):1–24, 2004. ISSN 0929-9629. doi: 10.1163/156939604323091180. URL http://dx.doi.org/10.1163/156939604323091180. 2

[4] L. Badouraly Kassim, J. Lelong, and I. Loumrhari. Importance sampling for jump processes and applications to finance. *Journal of Computational Finance (to appear)*, 00 (00), 2014. http://hal.archives-ouvertes.fr/hal-00842362. 9

[5] V. Bally and D. Talay. The law of the Euler scheme for stochastic differential equations. I. Convergence rate of the distribution function. *Probab. Theory Related Fields*, 104(1): 43–60, 1996. ISSN 0178-8051. 4

[6] M. Ben Alaya and A. Kebaier. Multilevel Monte Carlo for Asian options and limit theorems. *Monte Carlo Methods Appl.*, 20(3):181–194, 2014. ISSN 0929-9629. doi: 10.1515/mcma-2013-0025. URL http://dx.doi.org/10.1515/mcma-2013-0025. 2

[7] M. Ben Alaya and A. Kebaier. Central limit theorem for the multilevel monte carlo euler method. *Ann. Appl. Probab.*, 25(1):211–234, 02 2015. 22, 31

[8] M. Ben Alaya, K. Hajji, and A. Kebaier. Importance sampling and statistical Romberg method. *Bernoulli*, 21(4):1947–1983, 2015. ISSN 1350-7265. doi: 10.3150/14-BEJ622. URL http://dx.doi.org/10.3150/14-BEJ622. 2, 6

[9] N. Bouleau and D. Lépingle. *Numerical methods for stochastic processes*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1994. ISBN 0-471-54641-0. A Wiley-Interscience Publication. 4

[10] H. F. Chen and Y. M. Zhu. Stochastic approximation procedures with randomly varying truncations. *Sci. Sinica Ser. A*, 29(9):914–926, 1986. ISSN 0253-5831. 2

[11] H. F. Chen, G. Lei, and A. J. Gao. Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stochastic Process. Appl.*, 27(2):217–231, 1988. ISSN 0304-4149. doi: 10.1016/0304-4149(87)90039-1. URL http://dx.doi.org/10.1016/0304-4149(87)90039-1. 2

[12] N. Collier, A.-L. Haji-Ali, F. Nobile, E. von Schwerin, and R. Tempone. A continuation multilevel Monte Carlo algorithm. *BIT*, 55(2):399–432, 2015. ISSN 0006-3835. doi: 10.1007/s10543-014-0511-3. URL http://dx.doi.org/10.1007/s10543-014-0511-3. 2

[13] J. Creutzig, S. Dereich, T. Müller-Gronbach, and K. Ritter. Infinite-dimensional quadrature and approximation of distributions. *Found. Comput. Math.*, 9(4): 391–429, 2009. ISSN 1615-3375. doi: 10.1007/s10208-008-9029-x. URL http://dx.doi.org/10.1007/s10208-008-9029-x. 2

[14] S. Dereich. Multilevel Monte Carlo algorithms for Lévy-driven SDEs with Gaussian correction. *Ann. Appl. Probab.*, 21(1):283–311, 2011. ISSN 1050-5164. doi: 10.1214/10-AAP695. URL http://dx.doi.org/10.1214/10-AAP695. 2

[15] D. Duffie and P. Glynn. Efficient Monte Carlo simulation of security prices. *Ann. Appl. Probab.*, 5(4):897–905, 1995. 9

[16] M. B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56 (3):607–617, 2008. ISSN 0030-364X. doi: 10.1287/opre.1070.0496. URL http://dx.doi.org/10.1287/opre.1070.0496. 2

[17] M. B. Giles. Improved multilevel Monte Carlo convergence using the Milstein scheme. In *Monte Carlo and quasi-Monte Carlo methods 2006*, pages 343–358. Springer, Berlin, 2008. doi: 10.1007/978-3-540-74496-2_20. URL http://dx.doi.org/10.1007/978-3-540-74496-2_20. 2

[18] M. B. Giles and L. Szpruch. Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. *Ann. Appl. Probab.*, 24(4):1585–1620, 2014. ISSN 1050-5164. doi: 10.1214/13-AAP957. URL http://dx.doi.org/10.1214/13-AAP957. 2

[19] M. B. Giles, D. J. Higham, and X. Mao. Analysing multi-level Monte Carlo for options with non-globally Lipschitz payoff. *Finance Stoch.*, 13(3): 403–413, 2009. ISSN 0949-2984. doi: 10.1007/s00780-009-0092-1. URL http://dx.doi.org/10.1007/s00780-009-0092-1. 2

[20] K. Hajji. *Accélération de la méthode de Monte Carlo pour des processus de diffusions et applications en Finance*. PhD thesis, Université Paris 13, 2014. 3

[21] S. Heinrich. Monte Carlo complexity of global solution of integral equations. *J. Complexity*, 14(2):151–175, 1998. ISSN 0885-064X. doi: 10.1006/jcom.1998.0471. URL http://dx.doi.org/10.1006/jcom.1998.0471. 2

[22] S. Heinrich. Multilevel monte carlo methods. *Lecture Notes in Computer Science, Springer-Verlag*, 2179(1):58–67, 2001. 2

[23] S. Heinrich and E. Sindambiwe. Monte carlo complexity of parametric integration. *J. Complexity*, 15(3):317–341, 1999. ISSN 0885-064X. doi: 10.1006/jcom.1999.0508. URL http://dx.doi.org/10.1006/jcom.1999.0508. Dagstuhl Seminar on Algorithms and Complexity for Continuous Problems (1998). 2

[24] J. Jacod. On continuous conditional Gaussian martingales and stable convergence in law. In *Séminaire de Probabilités, XXXI*, volume 1655 of *Lecture Notes in Math.*, pages 232–246. Springer, Berlin, 1997. 31

[25] J. Jacod and P. Protter. Asymptotic error distributions for the Euler method for stochastic differential equations. *Ann. Probab.*, 26(1):267–307, 1998. ISSN 0091-1798. 31

[26] B. Jourdain and J. Lelong. Robust Adaptive Importance Sampling for Normal Random Vectors. *Ann. Appl. Probab.*, 19(5):1687–1718, 2009. doi: 10.1214/09-AAP595. URL http://arxiv.org/pdf/0811.1496v2. 2, 3, 6, 8, 22, 25

[27] A. Kebaier. Statistical Romberg extrapolation: a new variance reduction method and applications to option pricing. *Ann. Appl. Probab.*, 15(4):2681–2705, 2005. doi: 10.1214/105051605000000511. 2, 4, 18

[28] P. E. Kloeden and E. Platen. Numerical methods for stochastic differential equations. In *Nonlinear dynamics and stochastic mechanics*, CRC Math. Model. Ser., pages 437–461. CRC, 1995. 1, 4

[29] B. Lapeyre and J. Lelong. A framework for adaptive Monte Carlo procedures. *Monte Carlo Methods Appl.*, 17(1):77–98, 2011. ISSN 0929-9629. doi: 10.1515/MCMA.2011.002. URL http://dx.doi.org/10.1515/MCMA.2011.002. 2

[30] J. Lelong. Almost sure convergence for randomly truncated stochastic algorithms under verifiable conditions. *Statist. Probab. Lett.*, 78(16):2632–2636, 2008. ISSN 0167-7152. doi: 10.1016/j.spl.2008.02.034. URL http://dx.doi.org/10.1016/j.spl.2008.02.034. 2

[31] V. Lemaire and G. Pagès. Unconstrained recursive importance sampling. *Ann. Appl. Probab.*, 20(3):1029–1067, 2010. ISSN 1050-5164. doi: 10.1214/09-AAP650. URL http://dx.doi.org/10.1214/09-AAP650. 2

[32] V. Lemaire and G. Pagès. Multilevel richardson-romberg extrapolation. Technical report, Arxiv, 2014. URL http://arxiv.org/abs/1401.1177v3. 2

[33] A. Rényi. On stable sequences of events. *Sankhyā Ser. A*, 25:293 302, 1963. 31

[34] D. Revuz. *Probabilités*. Hermann, 1997. 19

[35] R. Y. Rubinstein and A. Shapiro. *Discrete event systems*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester, 1993. Sensitivity analysis and stochastic optimization by the score function method. 7

[36] D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, 8(4):483–509 (1991), 1990. ISSN 0736-2994. doi: 10.1080/07362999008809220. URL http://dx.doi.org/10.1080/07362999008809220. 4