Multilevel Monte-Carlo methods and lower-upper bounds in Initial Margin computations

F Bourgey, S de Marco, Emmanuel Gobet, Alexandre Zhou

Multilevel Monte-Carlo methods and lower–upper bounds in Initial Margin computations

F. Bourgey[*], S. De Marco[†], E. Gobet[‡] and A. Zhou[§] [¶]

January 7, 2020

Abstract

The Multilevel Monte-Carlo (MLMC) method developed by Giles [Gil08] has a natural application to the evaluation of *nested* expectation of the form $\mathbb{E}\left[g(\mathbb{E}\left[f(X,Y)|X\right])\right]$, where f, g are functions and (X, Y) a couple of independent random variables. Apart from the pricing of American-type derivatives, such computations arise in a large variety of risk valuations (VaR or CVaR of a portfolio, CVA), and in the assessment of margin costs for centrally cleared portfolios. In this work, we focus on the computation of Initial Margin. We analyze the properties of corresponding MLMC estimators, for which we provide results of asymptotical optimality; at the technical level, we have to deal with limited regularity of the outer function g (which might fail to be everywhere differentiable). Parallel to this, we investigate upper and lower bounds for nested expectations as above, in the spirit of primal/dual algorithms for stochastic control problems.

1 Introduction

Nested expectations – to be understood as expectations of functionals of conditional expectations – are ubiquitous in the field of financial risk assessment: from standard risk management computations, where a market factor is simulated up to a certain time horizon and then the performance of a portfolio is evaluated conditionally on the value of the factor, to dynamic hedging of derivatives products, in particular nowadays after the widespread establishment of the so-called *valuation adjustments* to derivative trades.

The paradigm of linear risk-neutral pricing of financial contracts has indeed changed in the last years, under the influence of market regulators: for several type of trades, banks and financial institutions have to post collateral to a central counterparty (CCP, also called clearing house) in order to secure their positions. Every day, the CCP requires each market member to deposit a certain capital according to the risk exposure of their contracts. From the modeling point of view, taking into account this type of regulatory capitals in the valuation of a derivative trade gives rise to non-linear (backward stochastic– or partial–) differential equations.

One of these protection capitals is the Initial Margin (IM) deposit: in case of default of one of the CCP members, the aim of this capital is to cover potential losses experienced by the hedging

[*]Centre de Mathématiques Appliquées (CMAP), CNRS, Ecole Polytechnique, Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau Cedex, France. Email: florian.bourgey@polytechnique.edu

[†]CMAP, Ecole Polytechnique. Email: stefano.de-marco@polytechnique.edu

[‡]CMAP, Ecole Polytechnique. Email: emmanuel.gobet@polytechnique.edu

[§]Université Paris-Est, CERMICS (ENPC), F-77455, Marne-la-Vallée, France. Email: alexandre.zhou@enpc.fr

[¶]This research has benefited from the support of the *Chaire Risques Financiers* of the Risk Foundation, the *Finance for Energy Market Research Centre* (FiME lab, from Institut Europlace de Finance), and the *Chaire Stress Test, RISK Management and Financial Steering* of the Ecole Polytechnique Foundation.

portfolio during the liquidation period of the defaulted member – concretely, the IM is materialized by the Value-at-risk or Conditional value-at-risk (CVaR) of the member's portfolio over a time period Δ. Since the time window Δ is small in year units (typically, one week), the usual approach in view of the computations is to apply an asymptotic expansion for the solution of the involved stochastic equation as Δ becomes small, see Henry-Labordère [HL17, Section 4.2] and Agarwal et al. [ADG+19]. To be more precise, let us borrow the mathematical setting from [ADG+19]: assuming for the underlying asset price the stochastic model $\mathrm{d}S_t = \mu_t S_t \mathrm{d}t + \sigma_t S_t \mathrm{d}W_t^{\mathrm{hist}}$, the value V of the hedging portfolio replicating a claim ξ solves the (eventually) linear Backward Stochastic Differential Equation (BSDE)

$$V_t = \xi + \int_t^T \left(-r_s V_s + Z_s \sigma_s^{-1} (r_s - \mu_s) + R\, C_\alpha \sqrt{(s+\Delta) \wedge T - s}\, \left| Z_s^{\mathrm{ref}} \right| \right) \mathrm{d}s - \int_t^T Z_s \mathrm{d}W_s^{\mathrm{hist}},$$

where r_t is a risk-free rate. Recall from BSDE theory that the solution to the equation above is the couple of processes (V_t, Z_t); using standard notation in derivative pricing theory, the wealth $\delta_t S_t$ invested in the asset at time t is related to the Z component via $\delta_t S_t = Z_t \sigma_t^{-1}$, δ_t being the number of shares of the asset S contained in the portfolio at time t (a.k.a. the portfolio's delta). The second term inside the time integral appears due to the aforementioned asymptotic expansion as Δ becomes small (note that the argument of the square root is equal to Δ for every $s \leq T - \Delta$), and corresponds to the additional IM cost, computed using a reference value δ_t^{ref} of the delta via $Z_t^{\mathrm{ref}} = \delta_t^{\mathrm{ref}} S_t \sigma_t$. The constant C_α is related to the CVaR of a standard normal distribution (we refer to Section 3 for more details), and R denotes the net interest rate of the account used to fund the IM cost. The reference value δ^{ref} is fixed by some external source, and often corresponds to the delta of a standard risk-neutral portfolio (in the simplest case, the Black-Scholes portfolio - but other choices are possible). When $\xi = \Phi(S_T)$, r_t is deterministic, σ_t and δ_t^{ref} are both functions of time and the underlying price S_t, we have the following classical Feynman-Kac representation for the initial price V_0

$$V_0 = \mathbb{E}\left[e^{-\int_0^T r_s \mathrm{d}s} \Phi(S_T) + R\, C_\alpha \int_0^T e^{-\int_0^t r_s \mathrm{d}s} \sqrt{(t+\Delta) \wedge T - t}\, \left| Z^{\mathrm{ref}}(t, S_t) \right| \mathrm{d}t \right], \tag{1.1}$$

where the expectation is computed under a probability measure such that $\mathrm{d}S_t = r_t S_t \mathrm{d}t + \sigma_t S_t \mathrm{d}W_t$. The function $Z^{\mathrm{ref}}(\cdot)$ being defined via a conditional expectation, the problem of computing the price V_0 requires, in general, to evaluate a nested expectation. This is our motivation for tackling expectations of the form

$$I := \mathbb{E}\left[g\left(\mathbb{E}\left[f\left(X, Y\right) | X \right] \right) \right], \tag{1.2}$$

by Monte-Carlo simulation, notably in situations where the function g has the regularity of the absolute value function $z \mapsto |z|$, as in (1.1). Precisely, we consider the formulation in this setting of the Multi-level Monte-Carlo method, introduced by Giles in [Gil08]. For a detailed survey of the MLMC method and various recent extensions and generalizations, see also [Gil15]. Our main result in this direction (Theorem 2.6) is to prove that the so-called antithetic Multi-level nested estimator [Gil15, Section 9] asymptotically achieves the performance of an unbiased estimator (that is: complexity $\mathcal{O}(\epsilon^{-2})$ for a tolerance ϵ^2 on the mean-squared error), even in this setting of limited regularity for the function g. On the other hand, in the (practically interesting) case where g is convex, we introduce and analyse a representation of (1.2) as the solution of a (primal/dual) optimization problem. In the spirit of Americain option pricing, this representation leads to upper and lower biased non-nested estimators, that we tackle with a regression-based procedure.

In the general setting, X, Y are two independent random variables with values respectively in $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, and $f : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are two measurable functions. By independence, the conditional expectation in (1.2) can be rewritten as:

$$\mathbb{E}\left[f\left(X, Y\right) | X \right] = E_f(X), \tag{1.3}$$

where for every $x \in \mathbb{R}^d, E_f(x) := \mathbb{E}\left[f\left(x, Y\right)\right]$. A natural estimator of the conditional expectation in (1.3) is, for $n \geq 1$,

$$\hat{E}_{f,n}(X) := \frac{1}{n}\sum_{i=1}^{n} f(X, Y_i)$$

where the $(Y_i)_{1\leq i\leq n}$ are i.i.d. samples of Y independent of X. Conditional on the value of X, the estimator $\hat{E}_{f,n}(X)$ is of course unbiased; the regularity of the outer function g will determine the bias of the random variable $g\big(\hat{E}_{f,n}(X)\big)$ and therefore of related nested estimators – see Section 2 for precise definitions. In order to estimate the bias and overall the performance of such estimators, there is a tradeoff between the smoothness of the function g and that of the probability distribution of the underlying random variables – the less regularity on g, the stronger the requirements on the underlying distribution. As mentioned above, the precise regularity conditions we are going to impose on the function g, motivated by the applicability to (1.1), together with the assumptions on the law of the underlying random variables, represent the core of our study and the main difference with other works (as we discuss more in details below).

In the existing literature, efforts have been made to evaluate the nested expectation I in (1.2) in the case of limited regularity, where g is a step function $g = 1_{[a,\infty)}$. This setting appears when evaluating tail distributions and quantiles; in the financial context, $\mathbb{E}\left[f\left(X, Y\right)|X\right]$ will typically represent the loss of a portfolio, conditional on the value of some market factor X. To our knowledge, in early papers dealing with this problem using nested Monte-Carlo methods, such as the seminal paper of Gordy and Juneja [GJ10] or Broadie et al. [BDM15], it is assumed that for $n \geq 1$, the couple of random variables $\big(E_f(X), \sqrt{n}(\hat{E}_{f,n}(X) - E_f(X))\big)$ has a joint density g_n w.r.t. the two-dimensional Lebesgue measure, and that the partial derivatives $\partial_y g_n(y, \cdot)$ and $\partial_y^2 g_n(y, \cdot)$ admit finite moments (up to order 4), uniformly in n. Under these conditions, the authors obtain an expansion of the bias of the nested estimator at order 1. Assumptions of this kind may look rather strong[1] and overall do not seem obvious to check: notably, a control on the moments of the joint law that is uniform over n does not seem easy to achieve. The analysis of the bias is pushed forward in the nice paper by Giorgi, Lemaire and Pagès [GLP18], where a higher order expansion is derived. These authors work, on the one hand, with a setting where g is smooth, and on the other hand with an opposite setting where g is a step function: in the latter case, it is assumed that both couples of random variables $(E_f(X), E_f(X) - \hat{E}_{f,n}(X))$ and $(E_f(X), X)$ admit smooth densities on $\mathbb{R}^2$ for $n \geq 1$, and again that a certain number of their partial derivatives with respect to the first variable exist and are continuous[2]. As a comparison with our setting, our Assumption 2.3 do not require existence of densities and their derivatives.

Let us mention some works where other types of functions g have been considered. In Bujok et al. [BHR15], the authors consider the pricing of CDO tranches and face the problem of estimating $\mathbb{E}\left[g(\mathbb{E}\left[Z|X\right])\right]$ in the precise case where g is piecewise linear and Z a Bernoulli random variable depending on some economic factor $X \in [0, 1]$, while in the survey paper [Gil15], the case where g is twice differentiable is discussed, thus allowing for more general assumptions on X. In Giles and Goda [GG19], the aim is to estimate $\mathbb{E}\left[\max_{d\in D} \mathbb{E}\left[f_d\left(X, Y\right)|X\right]\right]$ where D is a finite set; similar assumptions to ours, though more restrictive, are made (see further below for more details). In both papers, the tolerance ϵ^2 on the mean-squared error is achieved with complexity $\mathcal{O}(\epsilon^{-2})$, using an antithetic Multi-level estimator.

In our setting, the function g would typically be the absolute value – and more generally, a continuous and piecewise C_b^1 function (see our Assumption 2.1). With respect to the works cited

[1] As an example, one can check that in the case where X is Gaussian, Y is a Bernoulli random variable with $\mathbb{P}(Y = 1) = 1 - \mathbb{P}(Y = -1) = p$ with $p > 0.5$, and $f(X, Y) = XY$, we have $E_f(X) = (2p - 1)X$, $\hat{E}_{f,1}(X) - E_f(X) = (Y_1 - 2p + 1)X$, so that the couple $((2p - 1)X, (Y - 2p + 1)X)$ does not have a density with respect to the two-dimensional Lesbesgue measure.

[2] More precisely, denoting respectively p_1 and p_2 the two densities, in order to derive a bias expansion at order $R \in \mathbb{N}$, [GLP18] assume existence of the partial derivatives $\partial_x^{(l)} p_1(x, y)$ for $l = 0, ..., 2R+1$, $\partial_x^{(l)} p_2(x, y)$ for $l = 0, ..., 2R$, and that $\partial_x^{(2R+1)} p_1(x, y)$ is continuous.

above, we are in a situation of intermediate regularity of the function g – less than C^2, but more regular than a step function – thus allowing to drop restrictive assumptions on the distribution of the underlying random variables. It will be sufficient for us to only exploit some (mild) regularity of the law of $E_f(X)$ in the neighbourhood of the singularities of g, instead of considering the joint law of $E_f(X)$ and its estimator. The arguments used in this work are close to those in Giles and Haji-Ali [GH19], where g is a step function and the authors assume that in a neighbourhood of the discontinuity point of the step function, the random variable $\frac{E_f(X)}{\sqrt{\mathbf{Var}(f(X,Y)|X)}}$ has a bounded density; the resulting Multi-level estimator achieves mean-squared error ϵ^2 with complexity $\mathcal{O}(\epsilon^{-2}|\log\epsilon|^2)$. Once again, we would like to point out that such conditions on the underlying distributions are stronger that ours. In fact, our setting allows to treat the case of a butterfly option payoff (see Section 3.2), for which the assumption of existence of a bounded density seems out of reach, while our milder Assumption 2.3 proves to be checkable in financial meaningful cases.

The paper is organized as follows. In Section 2, we study the generic problem of estimating I. Here we compare theoretically and numerically different nested estimators. In particular, we show that using the MLMC method, the tolerance ϵ^2 on the mean-squared error can be achieved with complexity $\mathcal{O}(\epsilon^{-2})$ for a function g satisfying Assumption 2.1, provided that the probability that the conditional expectation $\mathbb{E}\left[f\left(X,Y\right)|X\right]$ is close to the singular points of g can be controlled polynomially. In Section 2.3 we introduce an algorithm computing lower and upper bounds using non nested Monte Carlo simulations. In Section 3 we show that the problem of evaluating the option price V_0 (1.1) with initial margin correction fits in the framework of our generic problem, and we numerically compare the results obtained using the different methods.

Notations. We denote $\|f\|_\infty$ the sup-norm for bounded functions, and $||A||_p := \mathbb{E}\left[|A|^p\right]^{\frac{1}{p}}$, $p > 0$, the L^p norm for vector-valued random variables. For a function $f : \mathbb{R} \mapsto \mathbb{R}$, its Total-Variation semi-norm is defined by $\|f\|_{\mathtt{TV}} := \sup\sum_{i=1}^{N}|f(x_i) - f(x_{i-1})|$, where the supremum is taken over finite sequences of increasing points $(x_i)_i$.

2 Theoretical methodology

As mentioned in the introduction, we want to design optimal Multilevel estimators in situations where the function g is of the form $g(x) = |x|$ or $g(x) = (x-a)^+$ for some $a \in \mathbb{R}$. Such examples enter in the class of functions covered by the following assumption: essentially, we are asking that the function g be continuous and piecewise C^1_b.

Assumption 2.1. *The function g is continuous, and there exists a finite set of points $-\infty = d_0 < d_1 < \ldots < d_\theta < d_{\theta+1} = \infty$ such that on each open interval (d_i, d_{i+1}), $0 \le i \le \theta$, g is of class C^1, g' is bounded and Hölder continuous for some exponent $\eta \in (0,1]$.*

We will also need to be able to control the probability that the random variable $E_f(X) = \mathbb{E}\left[f\left(X,Y\right)|X\right]$ takes values in a neighbourhood of the singularities of g; Assumption 2.3 below precisely takes care of this. Finally, our remaining Assumption 2.2 requires finiteness of some moment of order $p > 2$ for the random variable $f(X,Y)$, a condition which is met in most of the examples of practical interest.

Assumption 2.2. *There exists $p > 2$ such that $\mathbb{E}\left[|f\left(X,Y\right)|^p\right] < \infty$.*

Assumption 2.3 (Small ball estimate around the singularities)**.** *There exist some positive constants ν, K_ν and z_0 such that*

$$\mathbb{P}\left(\operatorname{dist}(E_f(X), D) \le z\right) \le K_\nu z^\nu, \qquad \forall z < z_0\,, \tag{2.1}$$

where $\operatorname{dist}(y, D) := \min_{1\le i\le\theta}|y - d_i|$.

Note that, if the random variable $E_f(X)$ has a bounded density p, then Assumption 2.3 trivially holds true with $\nu = 1$, since in this case $\mathbb{P}\left(\mathrm{dist}(E_f(X), D) \leq z\right) \leq \sum_{i=1}^{\theta} \mathbb{P}\left(|E_f(X) - d_i| \leq z\right) \leq 2\theta ||p||_\infty z$. On the other hand, Assumption 2.3 is more general: it is stated in terms of the distribution of $E_f(X)$ and does not require existence or regularity of a density for $E_f(X)$.

The basic estimator of the nested expectation (1.2) approximates the inner conditional expectation and the outer expectation with independent Monte-Carlo samples; we obtain the plain Nested Monte-Carlo (NMC) estimator

$$\hat{I}_{M,N} = \frac{1}{M} \sum_{m=1}^{M} g\left(\frac{1}{N} \sum_{j=1}^{N} f\left(X_m, Y_j^m\right)\right), \tag{2.2}$$

where $M, N \in \mathbb{N}^*$, and $(X_m)_{m\in\mathbb{N}^*}$, $\left(Y_j^m\right)_{j,m\in\mathbb{N}^*}$ are independent i.i.d. families having the same distribution as X and Y respectively. Multilevel (ML) estimators are obtained by combining estimators of the form of $\hat{I}_{M,N}$. In the following Section we consider the so called *antithetic* ML estimator, for which we show that it achieves the asymptotically unbiased setting; another "standard" (and sub-optimal, then) ML estimator is also analysed in Section 2.2, see (2.8).

2.1 Multilevel Monte-Carlo estimator

Let us consider two independent families of i.i.d. random variables $\left(X_m^l\right)_{l,m\in\mathbb{N}^*}$ and $\left(Y_j^{l,m}\right)_{j,l,m\in\mathbb{N}^*}$, distributed according to X and Y respectively. We denote $L \in \mathbb{N}^*$ the number of levels in the estimator. Let $\mathbf{M} = (M_0, ..., M_L) \in (\mathbb{N}^*)^{L+1}$, resp. $\mathbf{n} = (n_0, ..., n_L) \in (\mathbb{N}^*)^{L+1}$, be multi-indexes representing the number of samples used to approximate the outer expectation, respectively the inner conditional expectation in (1.2) at the different levels. We assume $n_l > n_{l-1}$ for every $l = 1, \ldots, L$. The antithetic Multilevel estimator of I is defined by

$$\begin{aligned}\hat{I}^{\mathrm{ML}}_{\mathbf{M},\mathbf{n}} =& \frac{1}{M_0} \sum_{m=1}^{M_0} g\left(\frac{1}{n_0} \sum_{j=1}^{n_0} f\left(X_m^0, Y_j^{0,m}\right)\right) + \sum_{l=1}^{L} \frac{1}{M_l} \sum_{m=1}^{M_l} \left\{ g\left(\frac{1}{n_l} \sum_{j=1}^{n_l} f\left(X_m^l, Y_j^{l,m}\right)\right)\right. \\ & \left. - \frac{1}{2}\left(g\left(\frac{1}{n_{l-1}} \sum_{j=1}^{n_{l-1}} f\left(X_m^l, Y_j^{l,m}\right)\right) + g\left(\frac{1}{n_{l-1}} \sum_{j=n_{l-1}+1}^{n_l} f\left(X_m^l, Y_j^{l,m}\right)\right)\right)\right\}.\end{aligned} \tag{2.3}$$

Note that, in the expression inside the curly brackets, the function g is evaluated more than once on different empirical means, constructed using the samples X_m^l and $Y_j^{l,m}$ at level l only. From now on, we set

$$n_l = n_0 2^l, \qquad 0 \leq l \leq L\,;$$

the remaining degrees of freedom of the estimator lie in the choices of n_0, L, and of the sample size M_l for every l.

The mean-squared error of $\hat{I}^{\mathrm{ML}}_{\mathbf{M},\mathbf{n}}$ can (as usual) be decomposed into bias and variance

$$\mathbf{MSE} := \mathbb{E}\left[\left(\hat{I}^{\mathrm{ML}}_{\mathbf{M},\mathbf{n}} - I\right)^2\right] = \left(\mathbb{E}\left[\hat{I}^{\mathrm{ML}}_{\mathbf{M},\mathbf{n}}\right] - I\right)^2 + \mathbf{Var}\left(\hat{I}^{\mathrm{ML}}_{\mathbf{M},\mathbf{n}}\right). \tag{2.4}$$

Under Assumptions 2.1, 2.2 and 2.3, we separately estimate the bias and variance terms in (2.4).

Proposition 2.4 (Bias estimate)**.** *Let $(Y_j)_{j\in\mathbb{N}^*}$ be an i.i.d. family of random variables distributed according to Y and independent of X. Under Assumptions 2.1, 2.2 and 2.3, there exists a positive constant κ such that*

$$\forall n \geq 1, \qquad \left| \mathbb{E}\left[g\left(\frac{1}{n} \sum_{j=1}^{n} f(X, Y_j)\right) - g\left(E_f(X)\right)\right]\right| \leq \frac{\kappa}{n^{\frac{1}{2}\left(1 + \frac{(p-1)\nu}{p+\nu} \wedge \eta\right)}}. \tag{2.5}$$

Note that the smooth case with no singularities (that is, $g \in C_b^1$ with Lipschitz continuous derivative) corresponds to $\eta = 1$ and $\nu = \infty$. In this case, the exponent for n at the denominator of (2.5) becomes $\frac{1}{2}(1+(p-1)\wedge 1)$. If $p \geq 2$, this number is worth $\frac{1}{2}\,2 = 1$. Therefore, in this case Proposition 2.4 provides the standard $\mathcal{O}\left(\frac{1}{n}\right)$ estimate for the bias in the presence of a smooth function.

We now turn to an estimate of the variance term. Proposition 2.5 estimates the contribution to the global variance of the ML estimator coming from level l. To do so, let us introduce the notation

$$V_l := \mathbf{Var}\left(g\left(\frac{1}{n_l}\sum_{j=1}^{n_l} f(X,Y_j)\right) - \frac{1}{2}g\left(\frac{1}{n_{l-1}}\sum_{j=1}^{n_{l-1}} f(X,Y_j)\right) - \frac{1}{2}g\left(\frac{1}{n_{l-1}}\sum_{j=n_{l-1}+1}^{n_l} f(X,Y_j)\right)\right).$$

Proposition 2.5 (Variance estimate)**.** *Under Assumptions 2.1, 2.2 and 2.3, there exists a positive constant $\tilde{\kappa}$ such that*

$$V_l \leq \frac{\tilde{\kappa}}{n_l^{1+\frac{(p-2)\nu}{2(p+\nu)}\wedge\eta}} \tag{2.6}$$

holds for every $l \geq 1$.

Once again, note that in the smooth case (that is: $g \in C_b^1$ with Lipschitz derivative, no singularities), we can take $\eta = 1$ and $\nu = \infty$. The exponent for n_l in (2.6) then becomes $1 + \frac{p-2}{2} \wedge 1$. If we are allowed to take $p \to \infty$, this number tends to $1 + 1 = 2$, which corresponds to a $\mathcal{O}\left(\frac{1}{n_l^2}\right)$ estimate of the variance at level l.

We are now in position to apply the Multilevel Theorem of Giles [Gil15] and conclude on global convergence rates. Since we wish to keep track of the constants arising from our assumptions (for numerical purposes), we also provide a brief proof of the theorem.

Theorem 2.6. *Let Assumptions 2.1, 2.2 and 2.3 hold true, and consider an error tolerance $\epsilon > 0$. There exist n_0, M_0, L such that the bound* $\mathbf{MSE} = \mathcal{O}(\epsilon^2)$ *as $\epsilon \to 0$ is achieved with a computational complexity $\mathcal{O}\left(\epsilon^{-2}\right)$ with the choice :*

$$n_l = n_0 2^l, \quad M_l = M_0\, 2^{-\left(1+\frac{(p-2)\nu}{4(p+\nu)}\wedge\frac{\eta}{2}\right)l}, \quad 0 \leq l \leq L.$$

As a comment on Theorem 2.6, note that it states an existence result for an optimal set of ML parameters L, n_l and M_l achieving the performance fo the unbiased setting. Unfortunately, the question of how to set numerically the constants n_0, M_0, L remains open – this is a usual situation in MLMC literature: we know how to optimize convergence rates, but some constants remain implicitly defined. However, supposing that we knew the constant κ (resp. $\tilde{\kappa}$) in Proposition 2.4 (resp. Propositon 2.5), then we could choose the following optimal values for n_0, M_0, L (see the proof of Theorem 2.6 for further details): for any $n_0 \in \mathbb{N}^*$, set

$$L = \left\lceil \frac{\log\left(\frac{\sqrt{2}\kappa}{n_0^\alpha \epsilon^2}\right)}{\alpha \log 2} \right\rceil, \quad M_0 = \left\lceil \frac{2}{\epsilon^2}\,\tilde{\kappa}\, n_0^{-\beta} \frac{1 - 2^{-\frac{L+1}{2}(\beta-\gamma)}}{1 - 2^{\frac{1}{2}(\beta-\gamma)}} \right\rceil. \tag{2.7}$$

Remark 2.7. Note also that the optimal number of Monte-Carlo samples at level l is of the form $M_l = M_0 2^{-(1+a)l}$ for a positive constant a, so that the computational cost at level l is proportional to $n_l\, M_l = n_0\, M_0\, 2^{-al} = \mathcal{O}(\epsilon^{-2})2^{-al}$: as usual in the ML framework, the most expensive levels are the coarsest ones. As we have done for Propositions 2.4 and 2.5, it is instructive to consider the case corresponding to $\eta = 1$ in Assumption 2.1. If we can take p to be large in Assumption 2.2, M_l approaches $M_0\, 2^{-\left(1+\frac{\nu}{4}\wedge\frac{1}{2}\right)}$. When there are no singularities, then we can take $\nu \to \infty$ in the small ball estimate (2.1) and obtain $M_l = M_0\, 2^{-\frac{3}{2}l}$. When there is at least one singularity but we can still take $\nu = 1$, then we obtain $M_l = M_0\, 2^{-\frac{3}{4}l}$. Overall, Theorem 2.6 tells that, as soon as $\nu > 0$ and $p > 2$, we can make the choice $M_l = M_0 2^{-(1+a)l}$ for some positive a, which is still enough to achieve an overall cost of order $\mathcal{O}\left(\epsilon^{-2}\right)$ for the ML estimator.

Proof of Theorem 2.6. We start from the decomposition (2.4). Due to the telescopic sum property of the ML estimator (2.3), it is clear that we have that $\mathbb{E}\left[\hat{I}^{\mathrm{ML}}_{\mathbf{M},\mathbf{n}}\right] = \mathbb{E}\left[g\left(\frac{1}{n_L}\sum_{j=1}^{n_L} f\left(X_1^L, Y_j^{L,1}\right)\right)\right]$. By Proposition 2.4, there exists a positive constant κ, independent of $\mathbf{n}$, such that

$$\left|\mathbb{E}\left[\hat{I}^{\mathrm{ML}}_{\mathbf{M},\mathbf{n}} - I\right]\right| = \left|\mathbb{E}\left[g\left(\frac{1}{n_L}\sum_{j=1}^{n_L} f\left(X, Y_j\right)\right)\right] - I\right| \leq \frac{\kappa}{n_L^{\alpha}} = \kappa\, n_0^{-\alpha}\, 2^{-\alpha L},$$
$$\text{with } \alpha := \frac{1}{2}\left(1 + \frac{(p-1)\nu}{p+\nu} \wedge \eta\right).$$

This estimate drives our choice of the number of levels; we want the squared bias to be smaller than $\epsilon^2/2$, which gives :

$$\kappa^2\, n_0^{-2\alpha}\, 2^{-2\alpha L} \leq \frac{\epsilon^2}{2} \iff L \geq \frac{\log\left(\frac{\sqrt{2}\kappa}{n_0^{\alpha}\epsilon^2}\right)}{\alpha \log 2}.$$

We therefore choose L as in (2.7).

On the other hand, Proposition 2.5 states that

$$V_l \leq \tilde{\kappa}\, n_0^{-\beta} 2^{-\beta l}, \qquad \forall\, 1 \leq l \leq L,$$

where $\beta = 1 + \frac{(p-2)\nu}{2(p+\nu)} \wedge \eta$. The computational cost C_l for one single sample in the layer l is proportional to n_l, hence bounded by a constant times $2^{\gamma l}$ with $\gamma = 1$. The hypotheses of [Gil15, Theorem 1] are therefore satisfied: in particular, it is easy to check that $\alpha \geq \frac{1}{2}\min(\beta, \gamma)$ and that $p > 2$ implies $\beta > \gamma$. Hence, we are in the first case of [Gil15, Theorem 1], which states that the overall complexity of the ML estimator behaves as $\mathcal{O}\left(\epsilon^{-2}\right)$.

The optimal number of Monte-Carlo samples M_l can be set by minimizing the computational cost of the ML estimator under the constraint that the global variance is smaller than $\epsilon^2/2$. The solution is of the form $M_l = M_0\, 2^{\frac{-(\beta+\gamma)l}{2}}$ where M_0 is a constant chosen so that the variance is smaller than $\epsilon^2/2$. More precisely, by independence $\mathbf{Var}\left(\hat{I}^{\mathrm{ML}}_{\mathbf{M},\mathbf{n}}\right) = \sum_{l=0}^{L} \frac{V_l}{M_l}$, and the following estimate holds :

$$\sum_{l=0}^{L} \frac{V_l}{M_l} \leq \sum_{l=0}^{L} \frac{\tilde{\kappa} n_0^{-\beta} 2^{-\beta l}}{M_0 2^{-\frac{l}{2}(\beta+\gamma)}} = \frac{\tilde{\kappa} n_0^{-\beta}}{M_0} \frac{1 - 2^{-\frac{L+1}{2}(\beta-\gamma)}}{1 - 2^{-\frac{1}{2}(\beta-\gamma)}},$$

where we recall that $\beta > \gamma$. Hence, setting $M_0 = \left\lceil \frac{2}{\epsilon^2}\tilde{\kappa} n_0^{-\beta} \frac{1-2^{-\frac{L+1}{2}(\beta-\gamma)}}{1-2^{-\frac{1}{2}(\beta-\gamma)}} \right\rceil$ yields $\mathbf{Var}\left(\hat{I}^{\mathrm{ML}}_{\mathbf{M},\mathbf{n}}\right) \leq \epsilon^2/2$ and by Equation 2.4, overall we obtain $\mathbf{MSE} \leq \epsilon^2/2 + \epsilon^2/2 = \epsilon^2$. It is sufficient to note that $\frac{\beta+\gamma}{2} = 1 + \frac{(p-2)\nu}{4(p+\nu)} \wedge \frac{\eta}{2}$ to conclude the proof. □

One can notice that, once the quantitative bias and variance estimates (2.5) and (2.6) have been established, the proof of Theorem 2.6 is relatively straightforward. The most difficult part lies indeed in proving Propositions 2.4 and 2.5; their proofs are provided in Sections 2.4 and 2.5.

2.2 Other nested estimators

We analyse in this Section alternative nested or MLMC estimators that will be tested in our numerical experiments. Some simulation-based lower and upper bounds on I that do not require nested simulations will be presented in the following Section.

The first is the plain Nested Monte-Carlo estimator (2.2). We denote $\mathbf{MSE}^{\mathrm{NMC}} := \mathbb{E}\left[\left(\hat{I}_{M,N} - I\right)^2\right]$ the associated mean-squared error.

Proposition 2.8 (Complexity of the Nested Monte-Carlo estimator (2.2))**.** *Let Assumptions 2.1, 2.2 and 2.3 hold true, and consider an error tolerance $\epsilon > 0$. As $\epsilon \to 0$, the bound $\mathbf{MSE}^{\mathrm{NMC}} = \mathcal{O}(\epsilon^2)$ is achieved with a computational complexity $\mathcal{O}\left(\epsilon^{-2\left(1+\frac{1}{1+\frac{(p-1)\nu}{p+\nu}\wedge\eta}\right)}\right)$ with the choice $M \sim \epsilon^{-2}$ and $N \sim \epsilon^{-\frac{2}{1+\frac{(p-1)\nu}{p+\nu}\wedge\eta}}$.*

Note that in the smooth case where we can take $\eta = 1$ and $\nu = \infty$, the exponent in the definition of N becomes $\frac{(p-1)\nu}{p+\nu} \wedge \eta = (p-1) \wedge \eta$. If $p \geq 2$, we retrieve the standard computational complexity $\mathcal{O}\big(\epsilon^{-3}\big)$ for nested estimators involving smooth functions.

In addition to the antithetic ML estimator $\hat{I}^{\mathrm{ML}}_{\mathbf{M},\mathbf{n}}$ considered in Section 2.1, we introduce a non-antithetic "standard" ML estimator $\hat{I}^{\mathrm{ML2}}_{\mathbf{M},\mathbf{n}}$, ML2 estimator in short, defined by

$$\hat{I}^{\mathrm{ML2}}_{\mathbf{M},\mathbf{n}} = \frac{1}{M_0}\sum_{m=1}^{M_0} g\left(\frac{1}{n_0}\sum_{j=1}^{n_0} f\left(X^0_m, Y^{0,m}_j\right)\right) + \sum_{l=1}^{L}\frac{1}{M_l}\sum_{m=1}^{M_l}\left\{ g\left(\frac{1}{n_l}\sum_{j=1}^{n_l} f\left(X^l_m, Y^{l,m}_j\right)\right) - g\left(\frac{1}{n_{l-1}}\sum_{j=1}^{n_{l-1}} f\left(X^l_m, Y^{l,m}_j\right)\right)\right\}, \tag{2.8}$$

where the notation is the same as in Section 2.1. We denote $\mathbf{MSE}^{\mathrm{ML2}} = \mathbb{E}\left[\left(\hat{I}^{\mathrm{ML2}}_{\mathbf{M},\mathbf{n}} - I\right)^2\right]$ the associated mean-squared error .

Proposition 2.9 (Complexity of the non-antithetic ML estimator (2.8))**.** *Let Assumptions 2.1, 2.2 and 2.3 hold true, and consider an error tolerance $\epsilon > 0$. There exist n_0, M_0, L such that the bound $\mathbf{MSE}^{\mathrm{ML2}} = \mathcal{O}(\epsilon^2)$ as $\epsilon \to 0$ is achieved with a computational complexity $\mathcal{O}\left((\log\epsilon)^2\epsilon^{-2}\right)$ with the choice :*

$$n_l = n_0 2^l, \quad M_l = M_0\, 2^{-l}, \quad 0 \leq l \leq L.$$

As our main Theorem 2.6, Propositions 2.8 and 2.9 are also based on the bias and variance estimates derived in Section 2.1; their proofs are postponed to the Appendix.

Observe that in terms of complexity for a given error tolerance, the two estimators considered in this Section are theoretically less efficient that the antithetic MLMC estimator (2.3). They will all be compared in our numerical tests.

2.3 Non nested upper and lower bounds

Following common practice in American option pricing (see [HK04]), another way of estimating nested expectations is to approach their value through lower and upper bounds which are computed using non nested Monte Carlo algorithms. Using Legendre-Fenchel transforms, a similar result has been proved (though their statements, proof and applications differ from ours) in the recent paper [GL18, Theorem 2.1] where the authors provide non-nested bounds for $\mathbb{E}\left[g\left(\mathbb{E}\left[H|\mathcal{F}_t\right]\right)\right]$ where $(\mathcal{F}_t)_{t\in[0,T]}$ denotes a filtration, H an $\mathcal{F}_T$-measurable random variable and g is convex. Our upper bound in Theorem 2.11 looks similar to [GL18, Theorem 2.1]), while our lower bound has a different representation from theirs. Furthermore, here we complement our lower/upper bounds with theoretical error estimates (see Theorem 2.11).

Lemma 2.10. *Let $K \in \mathbb{R}$ and $\mathcal{R}$ a real integrable random variable. For any measurable function $\varphi : \mathbb{R} \to \mathbb{R}$ and random variable $\mathcal{O}$, we have*

$$\mathbb{E}\left[(\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right] - K)_+\right] \geq \mathbb{E}\left[(\mathcal{R} - K)\mathbf{1}_{\varphi(\mathcal{O})\geq K}\right] =: C_\varphi(K),$$
$$\mathbb{E}\left[(K - \mathbb{E}\left[\mathcal{R}|\mathcal{O}\right])_+\right] \geq \mathbb{E}\left[(K - \mathcal{R})\mathbf{1}_{\varphi(\mathcal{O})\leq K}\right] =: P_\varphi(K).$$

Proof. Let φ and $\mathcal{O}$ as above. Since

$$\mathbb{E}\left[(\mathcal{R}-K)\mathbf{1}_{\varphi(\mathcal{O})\geq K}\right] = \mathbb{E}\left[(\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]-K)\mathbf{1}_{\varphi(\mathcal{O})\geq K}\right]$$

and $(\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]-K)_+ = (\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]-K)\mathbf{1}_{\mathbb{E}[\mathcal{R}|\mathcal{O}]\geq K}$, we deduce that

$$\mathbb{E}\left[(\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]-K)_+\right] - \mathbb{E}\left[(\mathcal{R}-K)\mathbf{1}_{\varphi(\mathcal{O})\geq K}\right] = \mathbb{E}\left[(\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]-K)(\mathbf{1}_{\mathbb{E}[\mathcal{R}|\mathcal{O}]\geq K} - \mathbf{1}_{\varphi(\mathcal{O})\geq K})\right].$$

Note that the random variable inside the expectation in the right hand side above is always non-negative, from which we obtain the first inequality. The second inequality is obtained with a similar argument. $\square$

Given a convex function $g:\mathbb{R}\to\mathbb{R}$, g is differentiable except on a countable set of points. For a differentiability point $z\in\mathbb{R}$ of g, we have

$$\forall x\in\mathbb{R},\; g(x) = g(z) + g'(z)(x-z) + \int_z^{\infty}(x-u)^+\mu(\mathrm{d}u) + \int_{-\infty}^{z}(u-x)^+\mu(\mathrm{d}u), \tag{2.9}$$

where $\mu(\mathrm{d}u)$ corresponds to the second derivative of g in the sense of distributions. Let us define, for a measurable function $\varphi:\mathbb{R}\to\mathbb{R}$, the functional

$$J_\varphi = g(z) + g'(z)(\mathbb{E}\left[\mathcal{R}\right]-z) + \int_z^{\infty} C_\varphi(u)\mu(\mathrm{d}u) + \int_{-\infty}^{z} P_\varphi(u)\mu(\mathrm{d}u). \tag{2.10}$$

Although J_φ also depends on z, we omit this dependence in our notation, for the sake of simplicity.

Theorem 2.11. *Let $\mathcal{R}$ be a real integrable random variable and $g:\mathbb{R}\to\mathbb{R}$ a convex function such that $g(\mathcal{R})$ is integrable. The following identity holds*

$$\sup_\varphi J_\varphi = \mathbb{E}\left[g(\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right])\right] = \inf_\varepsilon \mathbb{E}\left[g(\mathcal{R}-\varepsilon)\right], \tag{2.11}$$

where the inf *is taken over the set of random variables $\{\varepsilon$ is integrable and such that $\mathbb{E}\left[\varepsilon|\mathcal{O}\right]=0\}$. Equality is attained for $\varphi^\star(\mathcal{O}) = \mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]$ and for $\varepsilon^\star = \mathcal{R}-\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]$. Moreover, the following error estimate holds:*

$$0 \leq J_{\varphi^\star} - J_\varphi \leq 2\,\mathbb{E}\left[\left|\varphi^\star(\mathcal{O})-\varphi(\mathcal{O})\right|\mu([\varphi^\star(\mathcal{O}),\varphi(\mathcal{O})])\right]. \tag{2.12}$$

If g is Lipschitz with Lipschitz constant L_g, the error estimate

$$0 \leq \mathbb{E}\left[g(\mathcal{R}-\varepsilon)\right] - \mathbb{E}\left[g(\mathcal{R}-\varepsilon^\star)\right] \leq L_g\left\|\varepsilon^\star-\varepsilon\right\|_1. \tag{2.13}$$

is straightforward. Finally, if g is improved to $\mathcal{C}^2$ with bounded second derivative, the upper bound in (2.13) can be replaced by $\frac{1}{2}\|g''\|_\infty\left\|\varepsilon^\star-\varepsilon\right\|_2^2$.

In (2.12), $\mu([a,b])$ stands for the measure of the interval $[a,b]$ if $a\leq b$ or $[b,a]$ if $b<a$. When $g(x)=x_+$ (μ is the Dirac measure at 0) as in our numerical experiments, the bounds (2.12) and (2.13) are replaced by $2\mathbb{E}\left[|\varphi^\star(\mathcal{O})-\varphi(\mathcal{O})|\right]$ and $\mathbb{E}\left[|\varepsilon^\star-\varepsilon|\right]$, with no hope to significantly improve them. If g were C^2 with bounded second derivative, we would get the better estimate $2\|g''\|_\infty\mathbb{E}\left[|\varphi^\star(\mathcal{O})-\varphi(\mathcal{O})|^2\right]$ for (2.12).

Proof. To obtain the inequality $\mathbb{E}\left[g(\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right])\right]\geq J_\varphi$ for every φ, it is sufficient to replace x by $\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]$ in (2.9) and then apply Lemma 2.10 and Fubini's theorem. Optimality of $\varphi^\star(\mathcal{O})=\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]$ is also implied by Lemma 2.10, and the equality on the left hand side of (2.11) follows.

The identity on the right hand side of (2.11) is also straightforward to prove: for any integrable random variable ε such that $\mathbb{E}\left[\varepsilon|\mathcal{O}\right]=0$, we have $\mathbb{E}\left[g(\mathcal{R}-\varepsilon)\right]\geq\mathbb{E}\left[g([\mathbb{E}\left[\mathcal{R}-\varepsilon|\mathcal{O}\right]))\right]=\mathbb{E}\left[g(\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right])\right]$ by Jensen's inequality. It is immediate to check that equality is reached for $\varepsilon^\star=\mathcal{R}-\mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]$, which concludes this part of the proof.

We now move to the error estimates. Using $C_\varphi(u) = \mathbb{E}\left[(\mathcal{R}-u)\, 1_{\varphi(O)\geq u}\right] = \mathbb{E}\left[(\varphi^\star(\mathcal{O})-u)\, 1_{\varphi(O)\geq u}\right]$ and $P_\varphi(u) = \mathbb{E}\left[(u-\varphi^\star(\mathcal{O}))\, 1_{\varphi(O)\leq u}\right]$, we get

$$J_{\varphi^\star} - J_\varphi = \int_z^\infty \mathbb{E}\left[(\varphi^\star(\mathcal{O})-u)\left(\mathbf{1}_{\varphi^\star(\mathcal{O})\geq u} - \mathbf{1}_{\varphi(\mathcal{O})\geq u}\right)\right]\mu(\mathrm{d}u) \tag{2.14}$$

$$+ \int_{-\infty}^z \mathbb{E}\left[(u-\varphi^\star(\mathcal{O}))\left(\mathbf{1}_{\varphi^\star(\mathcal{O})\leq u} - \mathbf{1}_{\varphi(\mathcal{O})\leq u}\right)\right]\mu(\mathrm{d}u). \tag{2.15}$$

Now, if $\varphi^\star(\mathcal{O}) \geq u > \varphi(\mathcal{O})$, we have

$$0 \leq (\varphi^\star(\mathcal{O})-u)\left(\mathbf{1}_{\varphi^\star(\mathcal{O})\geq u} - \mathbf{1}_{\varphi(\mathcal{O})\geq u}\right) = (\varphi^\star(\mathcal{O})-u)\mathbf{1}_{\varphi^\star(\mathcal{O})\geq u>\varphi(\mathcal{O})} \leq (\varphi^\star(\mathcal{O})-\varphi(\mathcal{O}))\mathbf{1}_{\varphi^\star(\mathcal{O})\geq u>\varphi(\mathcal{O})}.$$

Similarly, if $\varphi(\mathcal{O}) \geq u > \varphi^\star(\mathcal{O})$,

$$0 \leq (\varphi^\star(\mathcal{O})-u)\left(\mathbf{1}_{\varphi^\star(\mathcal{O})\geq u} - \mathbf{1}_{\varphi(\mathcal{O})\geq u}\right) \leq (\varphi(\mathcal{O})-\varphi^\star(\mathcal{O}))\mathbf{1}_{\varphi(\mathcal{O})\geq u>\varphi^\star(\mathcal{O})}.$$

Otherwise, we have $(\varphi^\star(\mathcal{O})-u)\left(\mathbf{1}_{\varphi^\star(\mathcal{O})\geq u} - \mathbf{1}_{\varphi(\mathcal{O})\geq u}\right) = 0$ and overall we get (applying the positivity of the measure μ)

$$\left|\int_z^\infty \mathbb{E}\left[(\varphi^\star(\mathcal{O})-u)\left(\mathbf{1}_{\varphi^\star(\mathcal{O})\geq u} - \mathbf{1}_{\varphi(\mathcal{O})\geq u}\right)\right]\mu(\mathrm{d}u)\right| \leq \mathbb{E}\left[|\varphi^\star(\mathcal{O})-\varphi(\mathcal{O})|\mu([\varphi^\star(\mathcal{O}),\varphi(\mathcal{O})])\right].$$

Using analogous arguments, the second integral in (2.14) can be bounded by the same upper bound as above; estimate (2.12) follows.

Finally, assuming that g'' exists and is bounded, we can apply Taylor's theorem with Lagrange remainder and get

$$g(\mathcal{R}-\varepsilon) - g(\mathcal{R}-\varepsilon^\star) \leq g'(\mathcal{R}-\varepsilon^\star)(\varepsilon^\star-\varepsilon) + \frac{\|g''\|_\infty}{2}(\varepsilon^\star-\varepsilon)^2.$$

Taking expectations and using that $\mathbb{E}\left[\mathbb{E}\left[g'(\mathcal{R}-\varepsilon^\star)(\varepsilon^\star-\varepsilon) \mid \mathcal{O}\right]\right] = 0$ since $g'(\mathcal{R}-\varepsilon^\star)$ is $\mathcal{O}$-measurable and $\varepsilon^\star, \varepsilon$ are centered conditionally on $\mathcal{O}$, we obtain the announced improvement of estimate (2.13). □

Upper and lower biased estimators

We set $\mathcal{R} = f(X,Y)$ and $\mathcal{O} = X$ to restore our main setting. Note that *any* choice of admissible φ and ε in (2.11) leads to a lower bound J_φ and an upper bound $\mathbb{E}\big[g(f(X,Y)-\varepsilon)\big]$ for the nested expectation $\mathbb{E}\left[g(\mathbb{E}\left[f(X,Y)|X\right])\right]$. The optimal choices $\varphi^\star(\mathcal{O}) = \mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]$ and $\varepsilon^\star = \mathcal{R} - \mathbb{E}\left[\mathcal{R}|\mathcal{O}\right]$ would require to exactly evaluate the conditional expectation itself; we can actually *approach* these optimal choices by approximating conditional expectations, as we do below, based on a (linear) regression algorithm.

Let $(p_i)_{i\geq 0}$ (resp. $(q_i)_{i\geq 0}$) be a basis of $L^2(X)$ (resp. $L^2(Y)$). Given $N \in \mathbb{N}^*$ and $k, d \in \mathbb{N}$, we solve the following Ordinary Least Squares minimization problem

$$(l_j^*)_{0\leq j\leq k} = \underset{l_j\in\mathbb{R},\, 0\leq j\leq k}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^N \left(f(\tilde{X}_i,\tilde{Y}_i) - \sum_{j=0}^k l_j p_j(\tilde{X}_i)\right)^2, \tag{2.16}$$

$$(u_{ab}^*)_{0\leq a,b\leq d} = \underset{u_{ab}\in\mathbb{R},\, 0\leq a,b\leq d}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^N \left(f(\tilde{X}_i,\tilde{Y}_i) - \sum_{a,b=0}^d u_{ab}\, p_a(\tilde{X}_i)\Big(q_b(\tilde{Y}_i) - \mathbb{E}\big[q_b(\tilde{Y}_i)\big]\Big)\right)^2, \tag{2.17}$$

where for $1 \leq i \leq N$, each $(\tilde{X}_i, \tilde{Y}_i)$ is a independent copy of (X,Y). We then define:

$$\varphi_k^*(X) := \sum_{j=0}^k l_j^* p_j(X), \qquad \varepsilon_d^* := \sum_{a,b=0}^d u_{ab}^*\, p_a(X)(q_b(Y) - \mathbb{E}\left[q_b(Y)\right]). \tag{2.18}$$

Notice that the coefficients $(l_j^*)_{0\le j\le k}$ and $(u_{ab}^*)_{0\le a,b\le d}$ are random and independent of (X,Y), being functions of the sample $(\tilde{X}_i,\tilde{Y}_i)_{1\le i\le N}$. This property allows to obtain the following

Proposition 2.12. *The following inequalities hold*

$$\mathbb{E}\left[J_{\varphi_k^*}\right] \le \mathbb{E}\left[g(\mathbb{E}\left[f(X,Y)|X\right])\right] \le \mathbb{E}\left[g(f(X,Y)-\varepsilon_d^*)\right]. \tag{2.19}$$

Proof. In order to obtain the inequality on the left hand side, it is sufficient to take conditional expectation with respect to $(\tilde{X}_i,\tilde{Y}_i)_{i\le N}$, then exploit the independence with respect to (X,Y) and proceed as in the first part of the proof of Theorem 2.11. Similarly, since the coefficients l^* and u^* are independent of (X,Y), it is easy to check that $\mathbb{E}\left[\varepsilon_d^*|X\right]=0$, so that ε_d^* is admissible for (2.11), and the inequality on the right hand side follows. □

When $k,d,N\to+\infty$ (see [GKKW02, Chapter XI] for precise conditions), φ_k^* and ε_d^* converge in L_2 to the optimal φ^* and ε^*. Owing to the error estimates in Theorem 2.11, the lower-upper estimators in Proposition 2.12 converge to the true value of the nested expectation $\mathbb{E}\left[g(\mathbb{E}\left[f(X,Y)|X\right])\right]$. Even if we do not derive here precise convergence rates (though this is theoretically possible, building on the reference above), the practical interest of the estimators in Proposition 2.12 is clear: we can replace the expectations on the left and right-hand side of (2.19) with standard (non-nested) Monte-Carlo estimates based on i.i.d. samples of (X,Y) (see Section 3.3 for details about the final simulation algorithm). This provides an upper and a lower bound for the nested expectation; by observing the size of the interval between the two bounds, we decide whether the precision of this estimation procedure is acceptable for our purposes.

2.4 Proof of the bias estimate in Proposition 2.4

For $x\in\mathbb{R}^d$ and $n\ge1$, let us define

$$\hat{E}_{f,n}(x)=\frac{1}{n}\sum_{j=1}^{n} f(x,Y_j) \quad\text{and}\quad \delta E_n(x)=\hat{E}_{f,n}(x)-E_f(x).$$

We first control the moments of $\delta E_n(X)$.

Lemma 2.13. *Assume that there exists $p\ge2$ such that $\mathbb{E}\left[|f(X,Y)|^p\right]<\infty$. Then, there exists a positive constant C_p independent of n such that*

$$\forall n\ge1,\ \forall q\in(0,p],\qquad ||\delta E_n(X)||_q\le\frac{C_p}{\sqrt{n}}.$$

Proof. Since $||\delta E_n(X)||_q\le||\delta E_n(X)||_p$ for $q\in(0,p]$, it is sufficient to deal with the case $q=p$. Let us define $Z_j=f(X,Y_j)-E_f(X)$ and, for $n\ge1$, $S_n=\sum_{i=1}^n Z_i$. Conditionally with respect to X, S is a martingale, since the variables Y_j are i.i.d. and independent from X. By Burkhölder's inequality [HH80, Theorem 2.10], there exists a constant c_p only depending on p such that a.s.,

$$\mathbb{E}\left[|S_n|^p|X\right]\le c_p\,\mathbb{E}\left[\left|\sum_{i=1}^n Z_i^2\right|^{p/2}|X\right]\le c_p\,n^{p/2}\,\mathbb{E}\left[|Z_1|^p|X\right],$$

where in the last inequality we have used the fact that $\left|\sum_{i=1}^n Z_i^2\right|^{p/2}\le n^{\frac{p}{2}-1}\sum_{i=1}^n|Z_i|^p$. Since $\delta E_n(X)=\frac{S_n}{n}$, we obtain $||\delta E_n(X)||_p=\frac{1}{n}\mathbb{E}\left[|S_n|^p\right]^{\frac{1}{p}}\le\frac{C_p}{\sqrt{n}}$, where $C_p=c_p^{1/p}||Z_1||_p$. □

In order to prove estimate (2.5), we expand g around the point $E_f(X)$, using a Taylor expansion with integral remainder. To cope with the non smooth character of g, let us introduce the generalized derivative of g which is equal to 0 on the set $\{d_1,...,d_\theta\}$, and to g' elsewhere. By a slight abuse of notation, we will also denote this function by g'. The following lemma is standard: we provide a short proof for completeness.

Lemma 2.14. *Under Assumption 2.1, for any $x \notin \{d_1, ..., d_\theta\}$ and any $y \in \mathbb{R}$, we have*

$$g(y) - g(x) = g'(x)(y-x) + (y-x)\int_0^1 \left[g'(x+\lambda(y-x)) - g'(x)\right]\mathrm{d}\lambda\,. \tag{2.20}$$

Moreover, for every $x, y \in (d_k, d_{k+1})$,

$$|g(y) - g(x) - g'(x)(y-x)| \le \frac{[g']_\eta}{1+\eta}|y-x|^{1+\eta}, \tag{2.21}$$

where $[g']_\eta = \max_{0\le k\le\theta}\sup_{z_1,z_2\in(d_k,d_{k+1}),z_1\neq z_2}\frac{|g'(z_1)-g'(z_2)|}{|z_1-z_2|^\eta}$.

Proof. Without loss of generality, we assume that $x < y$ and consider the finite set $(e_1, ...e_k) = [x, y] \cap \{d_1, ..., d_\theta\}$, ordered in ascending order. Let us also denote $e_0 = x$ and $e_{k+1} = y$ if $e_k < y$. Since g is continuous on $\mathbb{R}$ and C^1 with Hölder continuous derivative on each interval (e_i, e_{i+1}), we have $g(e_{i+1}) - g(e_i) = \int_{e_i}^{e_{i+1}} g'(s)\mathrm{d}s$. Summing over the index i, we obtain $g(y) - g(x) = \int_x^y g'(s)\mathrm{d}s$. We then obtain (2.20) adding and subtracting $g'(x)$ inside the integral and then applying the change of variable $s = x + \lambda(y - x)$. To prove (2.21), it is sufficient to use (2.20) and the Hölder property of g' on the interval (d_k, d_{k+1}). □

We are now ready to prove Proposition 2.4. By Lemma 2.14, applying (2.20), we have that the identity

$$\begin{aligned} g\left(\hat{E}_{f,n}(X)\right) - g\left(E_f(X)\right) &= \delta E_n(X) g'\left(E_f(X)\right) \\ &\quad + \delta E_n(X)\int_0^1 \left(g'\left(E_f(X) + \lambda\delta E_n(X)\right) - g'\left(E_f(X)\right)\right)\mathrm{d}\lambda =: T_1 + T_2 \end{aligned} \tag{2.22}$$

holds on the set $E_f(X) \notin \{d_1, \dots, d_\theta\}$. By Assumption 2.3, this set has probability equal to one. Using the tower property of the conditional expectation, we have

$$\mathbb{E}\left[T_1\right] = \mathbb{E}\left[\delta E_n(X) g'\left(E_f(X)\right)\right] = 0,$$

since $\mathbb{E}\left[\delta E_n(X)|X\right] = \mathbb{E}\left[\hat{E}_{f,n}(X)\middle|X\right] - E_f(X) = 0$ a.s., and g' is bounded.

We now control the term $\mathbb{E}\left[T_2\right]$. We have to consider separately the cases where $E_f(X)$ is close to, respectively far from, the singularities of g. To do so, let us set $h = \frac{1}{2}\min_{1\le i\neq j\le\theta}\left(|d_i - d_j|\right)$ and introduce a parameter $z \in (0, h)$ to be fixed later. We consider the following partition:

$$\Omega_1 = \left\{\min_{1\le j\le\theta}|E_f(X) - d_j| \le z\right\}, \quad \Omega_2 = \Omega_1^c \cap \left\{|\delta E_n(X)| > z\right\}, \quad \Omega_3 = \Omega_1^c \cap \left\{|\delta E_n(X)| \le z\right\}.$$

We upper bound the values of the three expectations $\mathbb{E}\left[T_2\mathbf{1}_{\Omega_1}\right], \mathbb{E}\left[T_2\mathbf{1}_{\Omega_2}\right]$ and $\mathbb{E}\left[T_2\mathbf{1}_{\Omega_3}\right]$. For the first term, note that since g' is bounded, we have

$$\left|\mathbb{E}\left[T_2\mathbf{1}_{\Omega_1}\right]\right| \le 2\left\|g'\right\|_\infty \mathbb{E}\left[\left|\delta E_n(X)\right|\mathbf{1}_{\Omega_1}\right] \le 2\left\|g'\right\|_\infty \mathbb{P}\left(\Omega_1\right)^{\frac{1}{q}}\left\|\delta E_n(X)\right\|_{q^*}, \tag{2.23}$$

where we have applied Hölder's inequality with $q, q^* \ge 1$ such that $\frac{1}{q} + \frac{1}{q^*} = 1$ and $q^* \le p$, where p is the exponent appearing in Assumption 2.2. Assumption 2.3 ensures that $\mathbb{P}\left(\Omega_1\right) \le K_\nu z^\nu$. Then, it follows from Lemma 2.13 that

$$\left|\mathbb{E}\left[T_2\mathbf{1}_{\Omega_1}\right]\right| \le 2\left\|g'\right\|_\infty K_\nu^{\frac{1}{q}} z^{\frac{\nu}{q}}\frac{C_p}{\sqrt{n}}.$$

For the second term, note that $\mathbf{1}_{\Omega_2} \le \left(\frac{|\delta E_n(X)|}{z}\right)^r$ for any $r > 0$. Therefore

$$\left|\mathbb{E}\left[T_2\mathbf{1}_{\Omega_2}\right]\right| \le 2\frac{\left\|g'\right\|_\infty}{z^r}\mathbb{E}\left[\left|\delta E_n(X)\right|^{r+1}\right] = 2\frac{\left\|g'\right\|_\infty}{z^r}\left\|\delta E_n(X)\right\|_{r+1}^{r+1} \le 2\frac{\left\|g'\right\|_\infty C_p^{r+1}}{z^r n^{\frac{r+1}{2}}},$$

where we applied Lemma 2.13 in the last inequality, under the condition that $r+1 \leq p$. For the third term, note that on the set Ω_3, both $E_f(X)$ and $\hat{E}_{f,n}(X)$ belong to the same interval (d_i, d_{i+1}). Applying estimate (2.21), we obtain :

$$|\mathbb{E}\left[T_2 \mathbf{1}_{\Omega_3}\right]| \leq \frac{[g']_\eta}{1+\eta} \mathbb{E}\left[|\delta E_f(X)|^{1+\eta} \mathbf{1}_{\Omega_3}\right] \leq \frac{[g']_\eta}{1+\eta} \frac{C_p^{1+\eta}}{n^{\frac{1+\eta}{2}}},$$

where we have applied Lemma 2.13 in the last inequality. Putting things together, we have shown that the inequality

$$\left|\mathbb{E}\left[g\left(\hat{E}_{f,n}(X)\right) - g\left(E_f(X)\right)\right]\right| \leq 2\,\|g'\|_\infty K_\nu^{\frac{1}{q}} z^{\frac{\nu}{q}} \frac{C_p}{\sqrt{n}} + 2\frac{\|g'\|_\infty C_p^{r+1}}{z^r n^{\frac{r+1}{2}}} + \frac{[g']_\eta}{1+\eta} \frac{C_p^{1+\eta}}{n^{\frac{1+\eta}{2}}}$$

holds for all $z \in (0,h)$, $r \leq p-1$ and q such that $q^* \leq p$.

The minimum point of the right hand side corresponds to the minimum point of the function

$$z \in \mathbb{R} \to \Psi(z) := \Gamma_q z^{\frac{\nu}{q}} n^{-\frac{1}{2}} + \Lambda_r z^{-r} n^{-\frac{r+1}{2}},$$

where $\Gamma_q = 2\,\|g'\|_\infty K_\nu^{\frac{1}{q}} C_p$ and $\Lambda_r = 2\,\|g'\|_\infty C_p^{r+1}$. The minimum of Ψ is achieved at $z^* = \left(\frac{\Lambda_r r q}{\nu \Gamma_q}\right)^{\frac{q}{rq+\nu}} n^{-\frac{1}{2}\frac{rq}{rq+\nu}}$, which is smaller than h for n large enough. The value $\Psi(z^*)$ at the minimum point is proportional to $n^{-\frac{1}{2}\left(1+\frac{r\nu}{rq+\nu}\right)}$. In order to maximize $\frac{r\nu}{rq+\nu}$ under the constraints $r+1 \leq p$, $\frac{1}{q} + \frac{1}{q^*} = 1$ and $q^* \leq p$, we set $r = p-1$ and $q = \frac{p}{p-1}$, which finally yields the estimate

$$\left|\mathbb{E}\left[g\left(\hat{E}_{f,n}(X)\right) - g\left(E_f(X)\right)\right]\right| \leq \frac{\kappa_1}{n^{\frac{1}{2}\left(1+\frac{(p-1)\nu}{p+\nu}\wedge\eta\right)}},$$

for all $n \geq h^{-2\left(1+\frac{\nu}{rq}\right)} \left(\frac{\Lambda_r r q}{\nu\Gamma_q}\right)^{\frac{2}{r}} =: N$, where $\kappa_1 = \Gamma_q \left(\frac{\Lambda_r r q}{\nu\Gamma_q}\right)^{\frac{\nu}{rq+\nu}} + \Lambda_r \left(\frac{\Lambda_r r q}{\nu\Gamma_q}\right)^{\frac{-rq}{rq+\nu}} + \frac{[g']_\eta}{1+\eta} C_p^{1+\eta}$. We conclude the proof of estimate (2.5) for all $n \geq 1$ (and not only for $n \geq N$) by improving the constant κ_1 to $\kappa = \kappa_1 + \kappa_2\, N^{\frac{1}{2}\left(1+\frac{(p-1)\nu}{p+\nu}\wedge\eta\right)}$, where κ_2 is a uniform bound on $|\mathbb{E}\left[g(\hat{E}_{f,n}(X)) - g\left(E_f(X)\right)\right]|$ for $1 \leq n \leq N$. $\square$

2.5 Proof of the variance estimate in Proposition 2.5

For simplicity, let us define

$$\hat{Z}_1 := \frac{1}{n_{l-1}} \sum_{j=1}^{n_{l-1}} f(X, Y_j), \quad \hat{Z}_2 := \frac{1}{n_{l-1}} \sum_{j=n_{l-1}+1}^{n_l} f(X, Y_j), \quad \hat{Z} := \frac{\hat{Z}_1 + \hat{Z}_2}{2} = \frac{1}{n_l} \sum_{j=1}^{n_l} f(X, Y_j), \tag{2.24}$$

so that $V_l = \mathbf{Var}\left(g(\hat{Z}) - \frac{1}{2}g(\hat{Z}_1) - \frac{1}{2}g(\hat{Z}_2)\right)$. Let us also define

$$\delta\hat{Z}_1 = \hat{Z}_1 - E_f(X), \quad \delta\hat{Z}_2 = \hat{Z}_2 - E_f(X), \quad \delta\hat{Z} = \tfrac{1}{2}(\delta\hat{Z}_1 + \delta\hat{Z}_2) = \hat{Z} - E_f(X).$$

Recall that $h = \frac{1}{2}\min_{1\leq i\neq j\leq\theta}\left(|d_i - d_j|\right)$, and introduce a parameter $z \in (0,h)$, to be fixed later. We consider a partition of Ω into disjoint sets, analogous to the one used in the proof of Proposition 2.4:

$$\tilde{\Omega}_1 = \left\{\min_{1\leq j\leq\theta} |E_f(X) - d_j| \leq z\right\} = \Omega_1,$$
$$\tilde{\Omega}_2 = \tilde{\Omega}_1^c \cap \left\{\max\left(|\delta\hat{Z}_1|, |\delta\hat{Z}_2|\right) > z\right\},$$

$$\tilde{\Omega}_3 = \tilde{\Omega}_1^c \cap \left\{ \max\left(|\delta \hat{Z}_1|, |\delta \hat{Z}_2| \right) \leq z \right\}.$$

We bound V_l from above using the second moment $\mathbb{E}\left[\left(g(\hat{Z}) - \frac{1}{2} g(\hat{Z}_1) - \frac{1}{2} g(\hat{Z}_2) \right)^2 \right]$. For $A \in \{\hat{Z}, \hat{Z}_1, \hat{Z}_2\}$, it follows from (2.20) and Assumption 2.3 that

$$g(A) = g\left(E_f(X)\right) + \delta A \int_0^1 g'\left(E_f(X) + \lambda\, \delta A\right) \mathrm{d}\lambda \qquad \text{with } \delta A = A - E_f(X),$$

and therefore, using the identity $\delta \hat{Z} = \frac{\delta \hat{Z}_1 + \delta \hat{Z}_2}{2}$, we get :

$$\begin{aligned}
\left(g(\hat{Z}) - \frac{1}{2} g(\hat{Z}_1) - \frac{1}{2} g(\hat{Z}_2) \right)^2 &\leq 2 \left(\frac{\delta \hat{Z}_1}{2} \int_0^1 \left(g'\left(E_f(X) + \lambda \delta \hat{Z} \right) - g'\left(E_f(X) + \lambda \delta \hat{Z}_1 \right) \right) \mathrm{d}\lambda \right)^2 \\
&\quad + 2 \left(\frac{\delta \hat{Z}_2}{2} \int_0^1 \left(g'\left(E_f(X) + \lambda \delta \hat{Z} \right) - g'\left(E_f(X) + \lambda \delta \hat{Z}_2 \right) \right) \mathrm{d}\lambda \right)^2 \\
&\leq 2 ||g'||_\infty^2 \left((\delta \hat{Z}_1)^2 + (\delta \hat{Z}_2)^2 \right).
\end{aligned}$$

We first focus on the set $\tilde{\Omega}_1$. According to Lemma 2.13, there exists a positive constant C_p s.t.

$$\mathbb{E}\left[\left((\delta \hat{Z}_1)^2 + (\delta \hat{Z}_2)^2 \right) \mathbf{1}_{\tilde{\Omega}_1} \right] \leq ||\delta \hat{Z}_1||_{2q^*}^2 \mathbb{P}\left(\tilde{\Omega}_1 \right)^{\frac{1}{q}} + ||\delta \hat{Z}_2||_{2q^*}^2 \mathbb{P}\left(\tilde{\Omega}_1 \right)^{\frac{1}{q}} \leq 2 \frac{C_p^2}{n_{l-1}} K_\nu^{\frac{1}{q}} z^{\frac{\nu}{q}},$$

where we have applied Hölder's inequality twice with conjugate exponents $q, q^* \geq 1$ such that $2q^* \leq p$, and we have used $n_l - n_{l-1} = n_{l-1}$. On the set $\tilde{\Omega}_2$, since $\mathbf{1}_{\tilde{\Omega}_2} < \frac{\max\left(|\delta \hat{Z}_1|, |\delta \hat{Z}_2| \right)^r}{z^r}$ for any $r > 0$, we obtain :

$$\mathbb{E}\left[\left((\delta \hat{Z}_1)^2 + (\delta \hat{Z}_2)^2 \right) \mathbf{1}_{\tilde{\Omega}_2} \right] \leq \frac{2}{z^r} \mathbb{E}\left[\max\left(|\delta \hat{Z}_1|, |\delta \hat{Z}_2| \right)^{r+2} \right] \leq \frac{4}{z^r} \frac{C_p^{r+2}}{n_{l-1}^{\frac{r+2}{2}}},$$

for any positive r such that $r + 2 \leq p$, again owing to Lemma 2.13. Finally, in order to control the term $\mathbb{E}\left[(g(\hat{Z}) - \frac{1}{2} g(\hat{Z}_1) - \frac{1}{2} g(\hat{Z}_2))^2 \mathbf{1}_{\tilde{\Omega}_3} \right]$, we use the fact that, on the set $\tilde{\Omega}_3$, all the variables $\hat{Z}_1$, $\hat{Z}_2$, $\hat{Z}$ and $E_f(X)$ belong to the same interval (d_i, d_{i+1}). Using the fact that g' is Hölder continuous together with the identity $\delta \hat{Z} = \frac{1}{2}(\delta \hat{Z}_1 + \delta \hat{Z}_2)$, we have on the set $\tilde{\Omega}_3$:

$$\begin{aligned}
\left(g(\hat{Z}) - \frac{1}{2} g(\hat{Z}_1) - \frac{1}{2} g(\hat{Z}_2) \right)^2 &\leq 2 \frac{(\delta \hat{Z}_1)^2}{4} \int_0^1 \left| g'\left(E_f(X) + \lambda \delta \hat{Z} \right) - g'\left(E_f(X) + \lambda \delta \hat{Z}_1 \right) \right|^2 \mathrm{d}\lambda \\
&\quad + 2 \frac{(\delta \hat{Z}_2)^2}{4} \int_0^1 \left| g'\left(E_f(X) + \lambda \delta \hat{Z} \right) - g'\left(E_f(X) + \lambda \delta \hat{Z}_2 \right) \right|^2 \mathrm{d}\lambda \\
&\leq 2 \frac{[g']_\eta^2}{4^{\eta+1} (1 + 2\eta)} \left((\delta \hat{Z}_2)^2 + (\delta \hat{Z}_1)^2 \right) \left| \delta \hat{Z}_2 - \delta \hat{Z}_1 \right|^{2\eta}.
\end{aligned}$$

Now using the inequalities $\left| \delta \hat{Z}_2 - \delta \hat{Z}_1 \right|^{2\eta} \leq 2^{2\eta} \left(|\delta \hat{Z}_2|^{2\eta} + |\delta \hat{Z}_1|^{2\eta} \right)$, $|\delta \hat{Z}_1|^{2\eta} |\delta \hat{Z}_2|^2 + |\delta \hat{Z}_1|^2 |\delta \hat{Z}_2|^{2\eta} \leq |\delta \hat{Z}_1|^{2\eta+2} + |\delta \hat{Z}_2|^{2\eta+2}$ together with Lemma 2.13 one more time, we obtain :

$$\mathbb{E}\left[\left(g(\hat{Z}) - \frac{1}{2} g(\hat{Z}_1) - \frac{1}{2} g(\hat{Z}_2) \right)^2 \mathbf{1}_{\tilde{\Omega}_3} \right] \leq 2 \frac{[g']_\eta^2 C_p^{2\eta+2}}{(1 + 2\eta)\, n_{l-1}^{1+\eta}}.$$

Gathering the estimates obtained so far, we have shown that the upper bound

$$\begin{aligned} V_l &\leq \mathbb{E}\left[\left(g(\hat{Z}) - \frac{1}{2}g(\hat{Z}_1) - \frac{1}{2}g(\hat{Z}_2)\right)^2\right] \\ &\leq 8||g'||_\infty^2 K_\nu^{\frac{1}{q}} z^{\frac{\nu}{q}} C_p^2 \frac{1}{n_l} + \frac{8||g'||_\infty^2 C_p^{r+2} 2^{\frac{r+2}{2}}}{z^r} \frac{1}{n_l^{\frac{r+2}{2}}} + 2\frac{[g']_\eta^2 C_p^{2\eta+2} 2^{1+\eta}}{(1+2\eta)} \frac{1}{n_l^{1+\eta}}, \end{aligned}$$

holds for every $z < h$, $r \leq p-2$ and q such that $2q^* \leq p$. We conclude by minimizing the right hand side with respect to z, and setting $r = p-2$ and $q = \frac{p/2}{p/2-1}$. □

2.6 An alternative set of hypotheses

We can formulate a version of Theorem 2.6 under an alternative set of hypotheses, where we relax the assumption on the regularity of g, at the price of strengthening the condition on the law of the random variable $\mathbb{E}[f(X,Y)|X]$. We obtain the following result, which may present interest in its own. Note that, as a difference with Assumption 2.1, here g' may fait to be Hölder continuous.

Theorem 2.15. *Let us assume that g admits a derivative g' in the sense of distributions which has bounded variation. Let us assume that $E_f(X)$ has a bounded density χ_E w.r.t. the Lebesgue measure and that there exists $p > 2$ such that $\mathbb{E}[|f(X,Y)|^p] < \infty$. The, the tolerance ϵ^2 on the MSE of the antithetic ML estimator (2.3) can be achieved with complexity $\mathcal{O}(\epsilon^{-2})$.*

The proof of Theorem 2.15 mimicks the one of Theorem 2.6; once again, we need to rely on an estimate of the bias and of the variance at each level, which is the content of the two propositions below.

Under the boundedness assumption on the density of $E_f(X)$, we manage to relax the regularity hypothesis on the function g' by exploiting the arguments of [Avi09, Thm 2.4 (i)].

Proposition 2.16 (Bias estimate under bounded variation condition for g')**.** *Under the assumptions of Theorem 2.15, there exists a constant $\kappa > 0$ such that*

$$\left|\mathbb{E}\left[g\left(\hat{E}_{f,n}(X)\right)\right] - g\left(E_f(X)\right)\right| \leq \frac{\kappa}{n^{1-\frac{1}{p+1}}}.$$

Proof. Proceeding as in the beginning of the proof of Proposition 2.4 (see equation (2.22)), we have

$$\mathbb{E}\left[g\left(\hat{E}_{f,n}(X)\right) - g\left(E_f(X)\right)\right] = \mathbb{E}[T_2]$$

$$\text{with } T_2 := \delta E_n(X) \int_0^1 \left(g'\left(E_f(X) + \lambda\delta E_n(X)\right) - g'\left(E_f(X)\right)\right) \mathrm{d}\lambda.$$

Applying Hölder's inequality with q, q^* in the interval $[1,p]$ such that $\frac{1}{q} + \frac{1}{q^*} = 1$, we obtain

$$|\mathbb{E}[T_2]| \leq \int_0^1 \|\delta E_n(X)\|_{q^*} \left\|\left(g'\left(E_f(X) + \lambda\delta E_n(X)\right) - g'\left(E_f(X)\right)\right)\right\|_q \mathrm{d}\lambda.$$

By Lemma 2.13, we have $\|\delta E_n(X)\|_{q^*} \leq \frac{C_p}{\sqrt{n}}$. For the second term, it follows from [Avi09, Thm 2.4(i)] that for every $0 \leq \lambda \leq 1$ and $r \leq p$,

$$\left\|g'\left(E_f(X) + \lambda\delta E_n(X)\right) - g'\left(E_f(X)\right)\right\|_q \leq 3^{1+\frac{1}{q}} ||g'||_{\mathrm{TV}} \|\chi_E\|_\infty^{\frac{r}{q(r+1)}} \|\delta E_n(X)\|_r^{\frac{r}{q(r+1)}}.$$

Now applying again Lemma 2.13 to the last term on the right hand side and maximizing the exponent $\frac{r}{q(r+1)}$ by choosing $q = \frac{p}{p-1}$ (corresponding to $q^* = p$) and $r = p$, we obtain

$$\left|\mathbb{E}\left[g\left(\hat{E}_{f,n}(X)\right) - g\left(E_f(X)\right)\right]\right| \leq \frac{\kappa}{n^{1-\frac{1}{p+1}}},$$

where $\kappa = 3^{1+\frac{1}{q}} ||g'||_{\mathrm{TV}} \|\chi_E\|_\infty^{\frac{r}{q(r+1)}} C_p$, and this concludes the proof. □

Proposition 2.17 (Variance estimate under bounded variation condition for g')**.** *Under the assumptions of Theorem 2.15, there exists a constant $\tilde{\kappa}$ such that*

$$V_l \leq \frac{\tilde{\kappa}}{n_l^{1+\frac{p-2}{2(p+1)}}},$$

for every $l \geq 1$.

Proof. We restore the notation of Proposition 2.5. Applying the Taylor expansion (2.20) to $x = E_f(X)$ and $y \in \{\hat{Z}_1, \hat{Z}_2, \hat{Z}\}$ defined in (2.24), we have

$$\begin{aligned}\left(g(\hat{Z}) - \frac{1}{2}(g(\hat{Z}_1) + g(\hat{Z}_2))\right)^2 \leq 3 &\left(\delta\hat{Z}\int_0^1 \Big(g'(E_f(X) + \lambda\delta\hat{Z}) - g'(E_f(X))\Big)\,\mathrm{d}\lambda\right)^2 \\ &+ \frac{3}{4}\left(\delta\hat{Z}_1\int_0^1 \Big(g'(E_f(X) + \lambda\delta\hat{Z}_1) - g'(E_f(X))\Big)\,\mathrm{d}\lambda\right)^2 \\ &+ \frac{3}{4}\left(\delta\hat{Z}_2\int_0^1 \Big(g'(E_f(X) + \lambda\delta\hat{Z}_2) - g'(E_f(X))\Big)\,\mathrm{d}\lambda\right)^2. \end{aligned}\tag{2.25}$$

The expectation of the first term on the right hand side of (2.25) can be bounded by

$$\mathbb{E}\left[\left(\int_0^1 \Big(g'(E_f(X) + \lambda\delta\hat{Z}) - g'(E_f(X))\Big)\,\mathrm{d}\lambda\right)^{2q}\right]^{\frac{1}{q}} \mathbb{E}\left[(\delta\hat{Z})^{2q^*}\right]^{\frac{1}{q^*}}, \tag{2.26}$$

where we used Hölder's inequality with $q > 1$ and q^* such that $\frac{1}{q} + \frac{1}{q^*} = 1$ and $2q^* \leq p$. By Lemma 2.13, the second factor in (2.26) is bounded by $\frac{C_p}{n_l}$. Moreover, by Jensen's inequality and [Avi09, Thm 2.4(i)],

$$\begin{aligned}\mathbb{E}\left[\left(\int_0^1 \Big(g'(E_f(X) + \lambda\delta\hat{Z}) - g'(E_f(X))\Big)\,\mathrm{d}\lambda\right)^{2q}\right] &\leq \int_0^1 \mathbb{E}\left[\Big(g'(E_f(X) + \lambda\delta\hat{Z}) - g'(E_f(X))\Big)^{2q}\right]\mathrm{d}\lambda \\ &\leq 3^{2q+1}||g'||_{\mathrm{TV}}^{2q}||\chi_E||_\infty^{\frac{p}{p+1}}||\delta\hat{Z}||_p^{\frac{p}{p+1}}.\end{aligned}$$

We proceed as in the proof of Proposition 2.16: we apply again Lemma 2.13 to $||\delta\hat{Z}||_p^{\frac{p}{p+1}}$ and then optimize the resulting $n_l^{-\frac{p}{2q(p+1)}}$ over q under the constraint $q^* \leq p/2$, which leads to the choice $q = \frac{p}{p-2}$. Overall, we have just showed that the expectation of the first term on the right hand side of (2.25) is bounded by a constant times $n_l^{-1-\frac{p-2}{2(p+1)}}$. Similar computations for the two other terms in (2.25) lead to the same upper bound, which concludes the proof. □

The proof of Theorem 2.15 now boils down to the application of [Gil15, Theorem 1] with coefficients $\alpha = 1 - \frac{1}{p+1}$, $\beta = 1 + \frac{p-2}{2(p+1)}$ and $\gamma = 1$, which satisfy $\alpha > \frac{1}{2}\max(\beta, \gamma)$ and $\beta > \gamma$.

3 Application to initial margin computations

In the following, we work ona a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and consider the augmented filtration $\mathcal{F}$ of a Brownian motion W, that is $\mathcal{F}_t = \mathcal{F}_t^W := \sigma(W_s, 0 \leq s \leq t, \mathcal{N}_\mathbb{P})$ where $\mathcal{N}_\mathbb{P}$ denotes the family of $\mathbb{P}-$negligible sets of $\mathcal{A}$. In this Section, we consider examples of IM computation where Theorem 2.6 applies. More precisely, we generate expectations by means of a Black-Scholes (BS) model

$$S_t = S_0 e^{\left(r - \frac{\sigma^2}{2}\right)t + \sigma W_t}, \quad S_0 > 0, \tag{3.1}$$

where the interest rate $r > 0$ and the the volatility $\sigma > 0$ are constant. The model (3.1) provides a simple (yet meaningful) setting where we can evaluate explicit reference values (or unbiased estimates) for some of the involved nested expectations (see Proposition 3.3 and Section 3.3 for a detailed discussion of several financial examples), which we can use to test the multi-level estimators.

In [ADG⁺19], the IM correction for an option with payoff $\Phi(S_T)$ is computed according to the CVaR of the future evolution of the replicating portfolio over a small time interval $\Delta > 0$. When $\Delta = 0$ (that is : no IM correction), the price at time t is given by the classical BS price, $\mathbb{E}\left[e^{-r(T-t)}\Phi(S_T)\middle| S_t\right]$, with first derivative $\delta(t,S) = \partial_s \mathbb{E}\left[e^{-r(T-t)}\Phi(S_T)\middle| S_t = s\right]$. As shown in [ADG⁺19], for small values of Δ, a first order correction to this BS price is given precisely by the integral term in (1.1), Z^{ref} being computed according to the same BS model:

$$R\,C_\alpha\,\mathbb{E}\left[\int_0^T e^{-rt}\sqrt{(t+\Delta)\wedge T - t}\,|Z_t|\,\mathrm{d}t\right], \tag{3.2}$$

where: R is the funding cost net interest rate, $C_\alpha := \mathbf{CVaR}^\alpha\left(\mathcal{N}(0,1)\right) = \left.\frac{e^{-\frac{x^2}{2}}}{(1-\alpha)\sqrt{2\pi}}\right|_{x=\mathcal{N}^{-1}(\alpha)}$ is the CVaR of a standard Gaussian random variable, and

$$Z_t = z(t,S_t) := \sigma S_t\,\delta(t,S_t). \tag{3.3}$$

Using the likelihood ratio method of Broadie and Glasserman [BG96], we can restore an expression of Z in terms of an expectation:

$$\begin{aligned} z^{BS}(t,s) &= \sigma s\,\partial_s \mathbb{E}\left[e^{-r(T-t)}\Phi(S_t)\middle| S_t = s\right] \\ &= \mathbb{E}\left[e^{-r(T-t)}\Phi(S_t)\frac{W_T - W_t}{T-t}\middle| S_t = s\right] \\ &= \mathbb{E}\left[e^{-r(T-t)}\left(\Phi(S_t) - \Phi(S_t)\right)\frac{W_T - W_t}{T-t}\middle| S_t = s\right]. \end{aligned} \tag{3.4}$$

In the last expression, the conditionally centered term $-\Phi(S_t)\frac{W_T-W_t}{T-t}$ that we have artificially introduced allows to reduce variance in the simulation, by playing the role of a control variate.

Assuming that every hedging operation is performed before $\tilde{T} := T-\Delta$, we have $\sqrt{(t+\Delta)\wedge T - t} = \sqrt{\Delta}$ inside (3.2), which allows us to consider the slightly modified quantity:

$$R\,C_\alpha\sqrt{\Delta}\,\tilde{T}\;\mathbb{E}\left[\frac{1}{\tilde{T}}\int_0^{\tilde{T}} e^{-rt}\,|z(t,S_t)|\,\mathrm{d}t\right] =: R\,C_\alpha\sqrt{\Delta}\,\tilde{T}\times I. \tag{3.5}$$

The proposition below, whose proof is elementary, shows that (3.5) can be cast under the form of a nested expectation as (1.2). In particular, instead of discretizing the time integral over $[0,\tilde{T}]$ – which would produce a bias – we introduce an independent random variable with uniform distribution over $[0,\tilde{T}]$. Note that the product of constants $R\,C_\alpha\sqrt{\Delta}\,\tilde{T}$ is fixed once and for all; in the following, we therefore focus on the evaluation of the expectation $I = \mathbb{E}\left[\frac{1}{\tilde{T}}\int_0^{\tilde{T}} e^{-rt}\,|z(t,S_t)|\,\mathrm{d}t\right]$.

Proposition 3.1. *Let Y, $\tilde{Z}$ and U be three independent random variables such that $Y,\tilde{Z} \sim \mathcal{N}(0,1)$ and $U \sim \mathcal{U}([0,\tilde{T}])$, and consider the function $g : x \in \mathbb{R} \to |x|$. Then $X = (U, S_0 e^{(r-\frac{\sigma^2}{2})U + \sigma\sqrt{U}\tilde{Z}}) \perp Y$, and the expectation I in (3.5) satisfies*

$$I = \mathbb{E}\left[g\left(E_f(X)\right)\right] = \mathbb{E}\big[|\mathbb{E}\left[f\left(X,Y\right)|X\right]|\big],$$

where f is defined by

$$f : ((t,s),y) \in ([0,\tilde{T}]\times(0,\infty))\times\mathbb{R} \to e^{-rT}\,\frac{\left(\Phi\left(s\,e^{(r-\frac{\sigma^2}{2})(T-t)+\sigma\sqrt{T-t}\,y}\right) - \Phi\left(s\right)\right)y}{\sqrt{T-t}}. \tag{3.6}$$

Proof. Using the fact that S_t and $W_T - W_t$ are independent, we have :

$$\begin{aligned} I &= \frac{1}{\tilde{T}} \int_0^{\tilde{T}} \mathbb{E}\left[e^{-rt} \left|z(t, S_t)\right|\right] \mathrm{d}t \\ &= \frac{1}{\tilde{T}} \int_0^{\tilde{T}} \mathbb{E}\left[\left|\mathbb{E}\left[e^{-rT}\left(\Phi\left(s\, e^{(r-\frac{\sigma^2}{2})(T-t)+\sigma(W_T-W_t)}\right) - \Phi\left(s\right)\right) \frac{W_T - W_t}{T-t}\right]_{s=S_t}\right|\right] \mathrm{d}t. \end{aligned}$$

Since $W_T - W_t \overset{d}{=} \sqrt{T-t}\, Y$ and $U \sim \mathcal{U}([0, \tilde{T}])$ is independent of Z, we can write

$$\begin{aligned} I &= \mathbb{E}\left[\left|\mathbb{E}\left[e^{-rT} \frac{\left(\Phi\left(se^{(r-\frac{\sigma^2}{2})(T-t)+\sigma\sqrt{T-t}Y}\right) - \Phi\left(s\right)\right) Y}{\sqrt{T-t}}\right]_{t=U, s=S_0 e^{\left(r-\frac{\sigma^2}{2}\right)U+\sigma\sqrt{U}\tilde{Z}}}\right|\right] \\ &= \mathbb{E}\left[g\left(\mathbb{E}\left[f\left(\left(t,s\right), Y\right)\right]_{t=U, s=S_0 e^{\left(r-\frac{\sigma^2}{2}\right)U+\sigma\sqrt{U}\tilde{Z}}}\right)\right]. \end{aligned}$$

Now, using the independence of $X = (U, S_0 e^{(r-\frac{\sigma^2}{2})U+\sigma\sqrt{U}\tilde{Z}})$ and Y, the argument of g in the last expression is precisely $\mathbb{E}\left[f\left(X,Y\right)|X\right]$. □

From now on, the function g will correspond to the absolute value, $g(x) = |x|$.

Additional notation. In the rest of this Section we will make use of the following notations, specific to the BS framework: for every $T > 0$, $K > 0$, $r > 0$, $s > 0$, every $t \in [0, T)$ and $a \in [0, K)$, we define :

$$\mathcal{N} : x \in \mathbb{R} \to \int_{-\infty}^{x} \frac{e^{-z^2/2}}{\sqrt{2\pi}} \mathrm{d}z, \qquad \text{(the c.d.f of the standard normal distribution)}, \tag{3.7}$$

$$d_1\left(t, T, s, K\right) := \frac{\log\left(\frac{s}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)\left(T-t\right)}{\sigma\sqrt{T-t}}, \tag{3.8}$$

$$A(t) := \frac{\log\left(\frac{K}{K-a}\right)}{\sigma\sqrt{T-t}}, \quad B(t) := \frac{\log\left(\frac{K+a}{K}\right)}{\sigma\sqrt{T-t}}, \tag{3.9}$$

$$\delta_t : x \in \mathbb{R} \to \mathcal{N}\left(x - B(t)\right) + \mathcal{N}\left(x + A(t)\right) - 2\mathcal{N}(x), \tag{3.10}$$

$$H_t : x \in \mathbb{R} \to e^{-\frac{A(t)^2}{2}} e^{-A(t)x} + e^{-\frac{B(t)^2}{2}} e^{B(t)x} - 2. \tag{3.11}$$

For future purposes, it is important to observe the identity

$$\sqrt{2\pi}\, \delta_t'(x)\, e^{\frac{x^2}{2}} = H_t(x). \tag{3.12}$$

Remark 3.2. Notice that $A(t) > B(t) > 0$, since $\frac{K^2}{K^2-a^2} > 1$.

We will also make use of the following well-known asymptotic expansion for the c.d.f. of the standard normal distribution (see e.g. [Gor41]) :

$$\mathcal{N}\left(z\right) \underset{z\to-\infty}{\sim} -\frac{\exp\left(-\frac{z^2}{2}\right)}{z\sqrt{2\pi}}. \tag{3.13}$$

3.1 Call and put options

The usual price at time $0 \le t \le T$ of a call option of maturity $T > 0$ and strike $K > 0$ is defined by :

$$v^{call}(t, S_t) = \mathbb{E}\left[e^{-r(T-t)} \Phi^{call}(S_t) \middle| S_t\right], \tag{3.14}$$

where $\Phi^{call}(s) = (s-K)_+$. Similarly, we define v^{put} from $\Phi^{put}(s) = (K-s)_+$ for a put option. Recall that the delta of a call option is positive : in the BS case (3.1),

$$\delta^{call}(t, S_t) = \mathcal{N}\left(d_1\left(t, T, S_t, K\right)\right) > 0, \tag{3.15}$$

while for a put option, the delta is negative:

$$\delta^{put}(t, S_t) = -\mathcal{N}\left(-d_1\left(t, T, S_t, K\right)\right) < 0. \tag{3.16}$$

Hence, we are in a situation where the conditional expectation $\mathbb{E}[f(X,Y)|X]$ has constant sign and the function g, here the absolute value function, does not induce any nonlinearity. It is also possible to derive closed formulas for the expectation I.

Proposition 3.3. *Denote* $I^{call} = \mathbb{E}\left[\left|E_f^{call}(X)\right|\right]$ *resp.* $I^{put} = \mathbb{E}\left[\left|E_f^{put}(X)\right|\right]$ *the expectation* I *in (3.5) associated to a call resp. put option, where* X *and* $E_f^{call}(X)$ *are given in Proposition 3.1. Then,*

$$I^{call} = S_0 \sigma \mathcal{N}\left(d_1\left(0, T, S_0, K\right)\right), \qquad I^{put} = S_0 \sigma \mathcal{N}\left(-d_1\left(0, T, S_0, K\right)\right). \tag{3.17}$$

As a comment, note that I^{call} and I^{put} do not depend on $\tilde{T}$.

Proof. We justify the first expression in (3.17). We follow the arguments of [GM12, Lemma 2.2]: using the martingale property of $(e^{-rt}v^{call}(t,S_t))_{t\leq \tilde{T}}$, it is possible to show (by differentiating w.r.t. S_0) that the process $e^{-rt}\delta^{call}(t,S_t)\sigma S_t = e^{-rt}Z_t^{call}$ is also a martingale. Since the sign of Z_t^{call} is positive (see (3.15)), we obtain:

$$\begin{aligned} I^{call} = \frac{1}{\tilde{T}} \int_0^{\tilde{T}} e^{-rt}\mathbb{E}\left[\left|Z_t^{call}\right|\right] \mathrm{d}t &= \frac{1}{\tilde{T}} \int_0^{\tilde{T}} \mathbb{E}\left[e^{-rt} Z_t^{call}\right] \mathrm{d}t \\ &= \frac{1}{\tilde{T}} \int_0^{\tilde{T}} \mathbb{E}\left[Z_0^{call}\right] \mathrm{d}t = \sigma S_0 \mathcal{N}\left(d_1\left(0, T, S_0, K\right)\right). \end{aligned}$$

The proof for the second expression in (3.17) is analogous. □

3.2 The butterfly option

In this Section, we focus on a butterfly option – a non-trivial case for which we will show that Theorem 2.6 applies. Note that the delta of such an option does change sign (see Figure 1) and thus, the simple arguments of Proposition 3.3 do not hold. The butterfly option payoff, price and delta are respectively defined by:

$$\Phi^{butterfly}(s) = \left(s - (K+a)\right)_+ + \left(s - (K-a)\right)_+ - 2(s-K)_+, \tag{3.18}$$

$$v_{T,K,a}^{butterfly}(t, S_t) = v_{T,K+a}^{call}(t, S_t) + v_{T,K-a}^{call}(t, S_t) - 2\, v_{T,K}^{call}(t, S_t), \tag{3.19}$$

$$\begin{aligned} \delta^{butterfly}(t, S_t) := \Delta_t(S_t) &= \left.\frac{\partial v_{T,K,a}^{butterfly}(t,s)}{\partial s}\right|_{s=S_t} \qquad (3.20) \\ &= \mathcal{N}\left(d_1\left(t,T,S_t,K+a\right)\right) + \mathcal{N}\left(d_1\left(t,T,S_t,K-a\right)\right) - 2\,\mathcal{N}\left(d_1(t,T,S_t,K)\right), \end{aligned}$$

for a given strike $K > 0$ and parameter $a > 0$. Note that we now make the maturity and strike parameters appear explicitly in the notation for the call option price $v_{T,K}^{call}$.
To fix ideas, we plot in Figure 1 the delta of a call, put and butterfly option where $S_0 = K = 100, a = \frac{K}{2}, T = 1, t = \frac{T}{5}, r = 0.1, \sigma = 0.3$. Observe that (as expected) while the call and put deltas have a constant sign, the delta of the butterfly takes values of both signs.

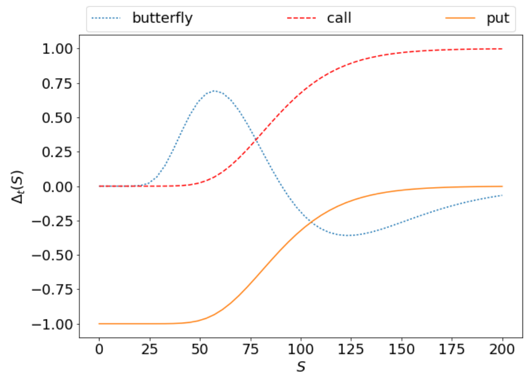


Figure 1: Derivatives of a call, put and butterfly option prices with respect to the variable price S.

The aim of this Section will be to show that Assumptions 2.1, 2.2, and 2.3 are satisfied for such a butterfly option. Since $\delta^{butterfly}$ changes its sign and the absolute value function g has a singular point at zero, we will need to study the behavior of $\delta^{butterfly}$ around zero in order to check that Assumption 2.3 is verified. Although the computations we are going to perform are specific to the butterfly payoff, we believe that the same arguments can be used to check Assumptions 2.1, 2.2 and 2.3 for more general payoffs, such as the ones we are going to consider in our numerical experiments in Section 3.3.

Theorem 3.4. *Assumptions 2.1, 2.2, 2.3 hold true in the butterfly case for* $\eta = 1$*, any* $p > 2$ *and* $\nu = \frac{1}{2}\left(1 \wedge \frac{T-\tilde{T}}{\tilde{T}(1+A)}\right)$ *for any* $A > 0$*. Therefore, Theorem 2.6 applies.*

The proof of Theorem 3.4 relies on a set of technical results, that we collect in the following Section.

3.2.1 Technical results required in the proof of Theorem 3.4

Notice that we can rewrite the butterfly delta as (see (3.10))

$$\Delta_t(S_t) = \delta_t\left(d_1(t, T, S_t, K)\right). \tag{3.21}$$

Lemma 3.5. *Let* $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^*$ *and* Z *be a standard normal r.v. Then for every* $A > 0$*, there exists* $z_0(A) > 0$ *s.t.*

$$\forall\, 0 < z < z_0(A), \qquad \mathbb{P}\left(\mathcal{N}\left(\mu + \sigma Z\right) < z\right) \le z^{\frac{1}{(1+A)\sigma^2}}.$$

The proof of Lemma 3.5 is postponed to Appendix A.3.

Lemma 3.6. *Denote* $s \in (0, \infty) \to p_t(s)$ *the density function of* S_t *defined in (3.1). Then*

$$\forall s \in (0, \infty), \quad p_t(s) := \frac{e^{-\frac{\left(\log(s) - \left(\log(S_0) + (r - \sigma^2/2)t\right)\right)^2}{2\sigma^2 t}}}{s\sigma\sqrt{t}\sqrt{2\pi}} \le \overline{p}_t := \frac{e^{\left(\sigma^2 - r\right)t}}{S_0\sigma\sqrt{t}\sqrt{2\pi}} \in L^1([0, \tilde{T}]). \tag{3.22}$$

Proof. For $t \in (0, \tilde{T}]$, the derivative of p_t is

$$p_t'(s) = \frac{1}{\sigma\sqrt{t}\sqrt{2\pi}s^2} e^{-\frac{\left(\log(s/S_0) - (r - \sigma^2/2)t\right)^2}{2\sigma^2 t}} \left[-\frac{\log(s/S_0) - (r - \sigma^2/2)t}{\sigma^2 t} - 1\right], \ s \in (0, \infty).$$

It is easy to check that the density has a single maximum, located at $s_0(t) := S_0 e^{\left(r-3\sigma^2/2\right)t}$ where $p_t'(s_0(t)) = 0$. Therefore,

$$\overline{p}_t := p_t\left(s_0(t)\right) = \frac{e^{-\frac{\left(-\sigma^2 t\right)^2}{2\sigma^2 t}}}{S_0 e^{(r-3\sigma^2/2)t}\sigma\sqrt{t}\sqrt{2\pi}} = \frac{e^{\left(\sigma^2-r\right)t}}{S_0\sigma\sqrt{t}\sqrt{2\pi}} \in L^1([0,\tilde{T}]).$$

□

The rather lenghtly proofs of the following two Propositions are postponed to Appendix A.4 and A.5.

Proposition 3.7. *For every $t \in [0,\tilde{T}]$,*

1. *The function H_t defined in (3.11) has exactly two zeros denoted $\alpha(t), \beta(t)$ s.t. $t \to \alpha(t)$ and $t \to \beta(t)$are $\mathcal{C}^\infty\left([0,T)\right)$ and $\alpha(t) < 0 < \beta(t)$.*
2. *The function $\Gamma_t : s \in (0,\infty) \to \Delta_t'(s)$ has two zeros denoted $\tilde{\alpha}(t), \tilde{\beta}(t)$ given by*

$$\tilde{\alpha}(t) = Ke^{\sigma\alpha(t)\sqrt{T-t}-\left(r+\frac{\sigma^2}{2}\right)(T-t)}, \quad \tilde{\beta}(t) = Ke^{\sigma\beta(t)\sqrt{T-t}-\left(r+\frac{\sigma^2}{2}\right)(T-t)}. \tag{3.23}$$

   In particular, $t \to \tilde{\alpha}(t)$ and $t \to \tilde{\beta}(t) \in \mathcal{C}^\infty\left([0,T)\right)$.

3. *The delta of the butterfly option $s \in (0,\infty) \to \Delta_t(s)$ is increasing for $s \in (0,\tilde{\alpha}(t)) \cup \left(\tilde{\beta}(t),\infty\right)$, decreasing for $s \in \left(\tilde{\alpha}(t),\tilde{\beta}(t)\right)$ and has a unique zero denoted $\gamma(t)$. Furthermore, $t \to \gamma(t) \in \mathcal{C}^\infty\left([0,T)\right)$ is bounded by*

$$\forall t \in [0,T), \quad \tilde{\alpha}(t) < Ke^{-\left(r+\frac{\sigma^2}{2}\right)(T-t)} < \gamma(t) < \tilde{\beta}(t). \tag{3.24}$$

Remark 3.8 (Geometric butterfly option). The geometric butterfly option corresponds to the purchase of two calls with strike Ke^{-a} and Ke^{a} ($a > 0$) and the sale of two puts with strike K. In other words, K is the geometric mean of Ke^{-a} and Ke^{a} while for a plain butterfly option, it is the arithmetic mean. For the geometric butterfly option, we can obtain an explicit formula for the unique zero of the delta

$$\Delta_t^{\text{geom}}(T,K,a) = \mathcal{N}\left(d_1(t,T,S_t,K) + \tilde{A}(t)\right) + \mathcal{N}\left(d_1(t,T,S_t,K) - \tilde{B}(t)\right) - 2\mathcal{N}\left(d_1(t,T,S_t,K)\right),$$

$\tilde{A}(t) = \tilde{B}(t) = \frac{a}{\sigma\sqrt{T-t}}$, which is now given by $\gamma(t) = Ke^{-\left(r+\frac{\sigma^2}{2}\right)(T-t)}$. Unfortunately, for the more realistic case of the arithmetic butterfly option (3.18), we do not have such an explicit expression for the zero of the delta function; but the estimates provided in Proposition 3.7 will be sufficient for the remaining part of our analysis.

Proposition 3.9. *For every $t \in [0,\tilde{T}]$, denote $\mathcal{V}(t)$ the neighborhood of $\gamma(t)$ defined by $\mathcal{V}(t) = (\gamma(t)-\epsilon(t), \gamma(t)+\epsilon(t))$ where $\epsilon(t) := \left(\frac{\gamma(t)-\tilde{\alpha}(t)}{2}\right) \wedge \left(\frac{\tilde{\beta}(t)-\gamma(t)}{2}\right)$, and denote χ_t the density function of the random variable $\Delta_t(S_t)$. Then, χ_t is bounded on $\mathcal{W}(t) := \Delta_t\left(\mathcal{V}(t)\right)$, more precisely:*

$$\forall y \in \mathcal{W}(t), \quad \chi_t(y) \le C(t) < \infty,$$

where $C \in L^1([0,\tilde{T}])$.

Note that Proposition 3.9 provides an upper bound for the density of the random variable $\Delta_t(S_t)$ (related to the nested expectation problem via $\mathbb{E}[f(X,Y)|X] = e^{-rU}\sigma S_U \Delta_U(S_U)$, see Proposition 3.1 for notations) only on the bounded interval $\Delta_t(\mathcal{V}(t))$. Unfortunately, this won't be enough to justify Assumption 2.3: in the butterfly option case, additional contributions to the "small" probability $\mathbb{P}(\mathrm{dist}(E_f(X), D) \le z)$ will also come from the left and right tails of the function $s \mapsto \Delta_t(s)$, which tend to zero exponentially fast as $s \to 0$ and $s \to \infty$, see Figure 1. In the proof of Theorem 3.4, we precisely have to take care of this fact. As a result, the estimate we obtain on the probability $\mathbb{P}(\mathrm{dist}(E_f(X), D) \le z)$ is of order z^ν with $\nu < 1$, which is *worse* than what we would obtain if we knew that the random variable $\mathrm{dist}(E_f(X), D)$ had a bounded density around zero.

3.2.2 Proof of Theorem 3.4

$\triangleright$ Since $g(x) = |x|$ and $g'(x) = \mathrm{sgn}(x)$, g has only one singular point $d_1 = 0$ and Assumption 2.1 holds true with $\eta = 1$.

$\triangleright$ We now check Assumption 2.2. Recall that f is defined in (3.6), that $S_t = S_0 e^{\left(r-\frac{\sigma^2}{2}\right)t+\sigma\sqrt{t}\tilde{Z}}$ and $\tilde{S}_t = S_t e^{\left(r-\frac{\sigma^2}{2}\right)(T-t)+\sigma\sqrt{T-t}Y}$, where $\tilde{Z} \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$, $U \sim \mathcal{U}([0,\tilde{T}])$ are independent. First, notice that $\frac{e^{-rT}}{\sqrt{T-U}} \le \frac{e^{-rT}}{\sqrt{T-\tilde{T}}}$ and $\Phi(s) = (s-(K+a))_+ + (s-(K-a))_+ - 2(s-K)_+ \in [0,a]$. Therefore, we get

$$|f(X,Y)| \le 2\frac{e^{-rT}}{\sqrt{T-\tilde{T}}}a|Y|.$$

and Assumption 2.2 is obviously fulfilled for any positive p.

$\triangleright$ Let us now deal with Assumption 2.3. From (3.3)-(3.5)-(3.20), we have $E_f(X) = e^{-rU}S_U\sigma\Delta_U(S_U)$. Therefore, $\mathrm{dist}(E_f(X), D) = |E_f(X)| = e^{-rU}S_U\sigma\,|\Delta_U(S_U)|$. Since

$$\mathbb{P}\left(e^{-rU}\sigma S_U\,|\Delta_U(S_U)| < z\right) \le \mathbb{P}\left(\sigma S_U\,|\Delta_U(S_U)| < z e^{r\tilde{T}}/\sigma\right), \tag{3.25}$$

it is sufficient to study (writing z instead of $e^{r\tilde{T}}z/\sigma$) $\mathbb{P}\left(S_U\,|\Delta_U(S_U)| < z\right)$. We have

$$\begin{aligned}\mathbb{P}\left(S_U\,|\Delta_U(S_U)| < z\right) &\le \mathbb{P}\left(S_U < \sqrt{z}\right) + \mathbb{P}\left(|\Delta_U(S_U)| < \sqrt{z}\right)\\ &\le \frac{1}{\tilde{T}}\int_0^{\tilde{T}}\int_0^{\sqrt{z}} p_t(s)\,\mathrm{d}s\mathrm{d}t + \mathbb{P}\left(|\Delta_U(S_U)| < \sqrt{z}\right),\end{aligned}$$

and by Lemma 3.6

$$\frac{1}{\tilde{T}}\int_0^{\tilde{T}}\int_0^{\sqrt{z}} p_t(s)\,\mathrm{d}s\mathrm{d}t \le \frac{\sqrt{z}}{\tilde{T}}\int_0^{\tilde{T}}\frac{e^{\left(\sigma^2-r\right)t}}{S_0\sigma\sqrt{t}\sqrt{2\pi}}\mathrm{d}t \le \frac{2e^{|\sigma^2-r|\tilde{T}}}{\sqrt{\tilde{T}}S_0\sigma\sqrt{2\pi}}\sqrt{z}. \tag{3.26}$$

Furthermore, defining the sets $\mathcal{S}_1(t) := (0,\tilde{\alpha}(t))$, $\mathcal{S}_2(t) := \left(\tilde{\alpha}(t),\tilde{\beta}(t)\right)$, $\mathcal{S}_3(t) := \left(\tilde{\beta}(t),\infty\right)$ one has

$$\begin{aligned}\mathbb{P}\left(|\Delta_U(S_U)| < \sqrt{z}\right) &= \mathbb{P}\left(\Delta_U(S_U) < \sqrt{z}, S_U \in \mathcal{S}_1(U)\right) + \mathbb{P}\left(|\Delta_U(S_U)| < \sqrt{z}, S_U \in \mathcal{S}_2(U)\right)\\ &\quad + \mathbb{P}\left(-\Delta_U(S_U) < \sqrt{z}, S_U \in \mathcal{S}_3(U)\right)\\ &=: p_1 + p_2 + p_3.\end{aligned}$$

Estimate of the second term p_2. Taking $\sqrt{z} < \min\limits_{t\in[0,\tilde{T}]}\left\{\left|\Delta_t\left(\gamma(t) + \frac{\tilde{\beta}(t)-\gamma(t)}{2}\right)\right| \wedge \left|\Delta_t\left(\gamma(t) - \frac{\gamma(t)-\tilde{\alpha}(t)}{2}\right)\right|\right\} := \sqrt{z_2}$, using $\{|\Delta_t(S_t)| < \sqrt{z}, S_t \in \mathcal{S}_2(t)\} \subset \{S_t \in \mathcal{V}(t)\}$ and Proposition 3.9, we get

$$p_2 = \frac{1}{\tilde{T}}\int_0^{\tilde{T}}\mathbb{P}\left(|\Delta_t(S_t)| < \sqrt{z}, S_t \in \mathcal{S}_2(t)\right)\mathrm{d}t$$

$$\leq \frac{1}{\tilde{T}} \int_0^{\tilde{T}} \int_{-\sqrt{z}}^{\sqrt{z}} \chi_t(y) \mathrm{d}y \mathrm{d}t \leq \left(\frac{2}{\tilde{T}} \int_0^{\tilde{T}} C(t) dt \right) \sqrt{z}. \tag{3.27}$$

Estimate of the first term p_1. Recall (3.10) and (3.21), so that $\Delta_t(S_t) = \delta_t\left(d_1(t, T, S_t, K)\right)$. Using (3.13), we have :

$$\begin{aligned}
\frac{\mathcal{N}\left(y - B(t)\right)}{\mathcal{N}\left(y + A(t)\right)} &\underset{y \to -\infty}{\sim} \frac{\exp\left(-\frac{(y-B(t))^2}{2}\right)}{\exp\left(-\frac{(y+A(t))^2}{2}\right)} \frac{y + A(t)}{y - B(t)} \\
&\underset{y \to -\infty}{\sim} \exp\left(\frac{1}{2}\left(A(t) + B(t)\right)\left(2y - B(t) + A(t)\right)\right) \xrightarrow[y \to -\infty]{} 0, \\
\frac{\mathcal{N}(y)}{\mathcal{N}\left(y + A(t)\right)} &\underset{y \to -\infty}{\sim} \exp\left(\frac{A(t)}{2}\left(2y + A(t)\right)\right) \xrightarrow[y \to -\infty]{} 0.
\end{aligned}$$

Consequently, $\delta_t(y) \underset{y \to -\infty}{\sim} \mathcal{N}\left(y + A(t)\right)$ and using that $\forall t \in [0, \tilde{T}]$, $\mathcal{N}(y + A(t)) > \mathcal{N}(y)$, we deduce the existence of $\theta_1 < 0$ s.t. :

$$t \in [0, \tilde{T}], \quad y < \theta_1 \implies \delta_t(y) > \frac{1}{2}\mathcal{N}(y).$$

W.l.o.g. we can assume that $\theta_1 < \inf_{t \in [0, \tilde{T}]} \tilde{\alpha}(t)$, possibly taking a smaller θ_1. Set $\sqrt{z_1} := \min_{t \in [0, \tilde{T}]} \delta_t\left(\theta_1\right)$. Combining the estimate above yields (for $\sqrt{z} \leq \sqrt{z_1}$)

$$\left\{\Delta_U(S_U) = \delta_U(d_1(U, T, S_U, K)) < \sqrt{z}, S_U \in \mathcal{S}_1(U)\right\} \subset \left\{\frac{1}{2}\mathcal{N}(d_1(U, T, S_U, K)) < \sqrt{z}\right\},$$

i.e. $p_1 \leq \mathbb{P}\left(\mathcal{N}\left(d_1(U, T, S_U, K)\right) < 2\sqrt{z}\right)$. Furthermore, noticing that $d_1(U, T, S_U, K) | U \sim \mathcal{N}\left(\mu_U, \sigma_U^2\right)$ where :

$$\underline{\mu} := \frac{\log\left(\frac{S_0}{K}\right) + rT + \sigma^2\left(T/2 - \tilde{T}\right)}{\sigma\sqrt{T}} < \mu_U := \frac{\log\left(\frac{S_0}{K}\right) + \left(r + \sigma^2/2\right)T - \sigma^2 U}{\sigma\sqrt{T - U}} < \bar{\mu} := \frac{\log\left(\frac{S_0}{K}\right) + \left(r + \sigma^2/2\right)T}{\sigma\sqrt{T - \tilde{T}}},$$

$$0 < \sigma_U := \sqrt{\frac{U}{T - U}} < \sigma_{\tilde{T}} := \sqrt{\frac{\tilde{T}}{T - \tilde{T}}} < \infty,$$

we obtain:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{N}\left(d_1(U, T, S_U, K)\right) < 2\sqrt{z}\right) &= \mathbb{P}\left(\mathcal{N}\left(\mu_U + \sigma_U N\right) < 2\sqrt{z}, N > 0\right) + \mathbb{P}\left(\mathcal{N}\left(\mu_U + \sigma_U N\right) < 2\sqrt{z}, N < 0\right) \\
&\leq \mathbb{P}\left(\mathcal{N}\left(\underline{\mu} - \sigma_{\tilde{T}} N\right) < 2\sqrt{z}, N > 0\right) + \mathbb{P}\left(\mathcal{N}\left(\underline{\mu} + \sigma_{\tilde{T}} N\right) < 2\sqrt{z}, N < 0\right) \\
&\leq 2\mathbb{P}\left(\mathcal{N}\left(\underline{\mu} + \sigma_{\tilde{T}} N\right) < 2\sqrt{z}\right),
\end{aligned}$$

with $N \sim \mathcal{N}(0, 1) \perp U$. Hence, from Lemma 3.5 we obtain that for every $A > 0$, there exists $z_{0,1} > 0$ s.t. :

$$\forall 0 < z < z_{0,1} \wedge z_1, \quad p_1 \leq 2^{1 + \frac{1}{(1+A)\sigma_{\tilde{T}}^2}} z^{\frac{1}{2(1+A)\sigma_{\tilde{T}}^2}} = 2^{1 + \frac{T - \tilde{T}}{\tilde{T}(1+A)}} z^{\frac{T - \tilde{T}}{2\tilde{T}(1+A)}}. \tag{3.28}$$

Estimate of the third term p_3. Using the identity $\mathcal{N}\left(y\right) + \mathcal{N}\left(-y\right) = 1$, the function $-\delta_t$ can be written as

$$-\delta_t(y) = \mathcal{N}\left(-y - A(t)\right) + \mathcal{N}\left(-y + B(t)\right) - 2\mathcal{N}\left(-y\right).$$

Note that we now have

$$\frac{\mathcal{N}\left(-y - A(t)\right)}{\mathcal{N}\left(-y + B(t)\right)} \underset{y \to +\infty}{\sim} \exp\left(\frac{1}{2}\left(A(t) + B(t)\right)\left(-2y - A(t) + B(t)\right)\right) \xrightarrow[y \to +\infty]{} 0,$$

$$\frac{\mathcal{N}(-y)}{\mathcal{N}(-y+B(t))} \underset{y\to+\infty}{\sim} \exp\left(\frac{1}{2}B(t)(-2y+B(t))\right) \xrightarrow[y\to+\infty]{} 0.$$

Hence, there exists $\theta_3 > 0$ such that $\forall y > \theta_3$, $-\delta_t(y) > \frac{1}{2}\mathcal{N}(-y+B(t)) > \frac{1}{2}\mathcal{N}(-y) > 0$. We can assume w.l.o.g. that $\theta_3 > \sup_{t\in[0,\tilde{T}]}\tilde{\beta}(t)$ and set $\sqrt{z_3} := \min_{t\in[0,\tilde{T}]} -\delta_t(\theta_3)$. Then, with similar arguments to the analysis of the term p_1 above, we get (for $z < z_3$)

$$\begin{aligned} p_3 &= \mathbb{P}\left(-\delta_U\left(d_1(U,T,S_U,K)\right) < \sqrt{z}, S_U \in \mathcal{S}_3(U)\right) \\ &\le \mathbb{P}\left(\mathcal{N}\left(-d_1(U,T,S_U,K)\right) < 2\sqrt{z}\right) \\ &= \mathbb{P}\left(\mathcal{N}\left(-\mu_U - \sigma_U N\right) < 2\sqrt{z}\right) \le 2\mathbb{P}\left(\mathcal{N}\left(-\bar{\mu} + \sigma_{\tilde{T}} N\right) < 2\sqrt{z}\right). \end{aligned}$$

Therefore, owing to Lemma 3.5, for any $A > 0$ there exists $z_{0,3} > 0$ such that

$$\forall 0 < z < z_{0,3} \wedge z_3, \quad p_3 \le 2^{1+\frac{T-\tilde{T}}{\tilde{T}(1+A)}} z^{\frac{T-\tilde{T}}{2\tilde{T}(1+A)}}. \tag{3.29}$$

Finally, putting estimates (3.26), (3.27), (3.28) and (3.29) together, we obtain an upper bound for (3.25): defining $\bar{z} := (z_{0,1} \wedge z_1 \wedge z_2 \wedge z_3 \wedge z_{0,3})\, e^{r\tilde{T}}/\sigma$, we conclude that for every $0 < z < \bar{z} \wedge 1 := z_0$,

$$\begin{aligned} \mathbb{P}\left(|E_f(X)| < z\right) &\le \left(\frac{2}{\tilde{T}}\int_0^{\tilde{T}} C(t)\mathrm{d}t + \frac{2e^{|\sigma^2-r|\tilde{T}}}{\sqrt{\tilde{T}}S_0\sigma\sqrt{2\pi}}\right)\left(\frac{z}{\sigma e^{-r\tilde{T}}}\right)^{\frac{1}{2}} + 2^{2+\frac{T-\tilde{T}}{\tilde{T}(1+A)}}\left(\frac{z}{\sigma e^{-r\tilde{T}}}\right)^{\frac{T-\tilde{T}}{2\tilde{T}(1+A)}} \\ &\le K_\nu z^\nu \end{aligned}$$

for ν as in the statement of the Theorem. Assumption 2.3 is therefore proved, and the proof of Theorem 3.4 follows. □

3.3 Numerical experiments

In this Section, we test and compare numerically the antithetic ML estimator (2.3), the non-antithetic ML estimator (2.8), the NMC estimator (2.2) and the upper and lower bounds (2.19) for the estimation of the initial margin correction (3.5) in different situations: a single call/ put/ butterfly option, and several portfolios containing different combinations of butterfly options. In order to evaluate the **MSE** of the different estimators, we will make use of the available closed formulas I^{call} and I^{put} (3.17) in the single call and put cases. Unfortunately, we do not possess analogous formulae for the expectation $I^{butterfly}$ associated to the butterfly option, or to butterfly option portfolios. In this case, since we still have an explicit expression for the conditional expectation $E_f(X)$ (based on (3.20)), we can construct a benchmark value by applying an intensive standard Monte-Carlo method to evaluate $\mathbb{E}\left[g(E_f(X))\right] = \mathbb{E}\left[|E_f(X)|\right]$.

3.3.1 The nested estimators

We estimate the mean-squared errors (denoted respectively $\mathbf{MSE}^{\mathrm{ML}}$, $\mathbf{MSE}^{\mathrm{ML2}}$ and $\mathbf{MSE}^{\mathrm{NMC}}$) with $\frac{1}{n_{\mathbf{MSE}}}\sum_{j=1}^{n_{\mathbf{MSE}}}\left(\hat{I}_j - I\right)^2$, where $(\hat{I}_j)_{1\le j\le n_{\mathbf{MSE}}}$ are independent copies of the respective estimator and $n_{\mathbf{MSE}} = 200$. We will denote the related computational complexities $\mathbf{C}^{\mathrm{ML}}$, $\mathbf{C}^{\mathrm{ML2}}$ and $\mathbf{C}^{\mathrm{NMC}}$.

In order to achieve $\mathbf{MSE} = \mathcal{O}(\epsilon^2)$ for $\epsilon > 0$, Theorem 2.6, Propositions 2.9 and 2.8 provide the following optimal choices of parameters for each estimator.

Construction of the antithetic ML estimator. The optimal ML estimator obtained in Theorem 2.6 depends on some problem-specific parameters ν, η, p. In our application to IM computation presented in Section 3, we can take $\eta = 1$, p arbitrary large and $\nu = \frac{1}{2}\left(1 \wedge \frac{T-\tilde{T}}{\tilde{T}(1+A)}\right)$ for any $A > 0$.

If $\tilde{T} < T/2$, then $\frac{T-\tilde{T}}{\tilde{T}} > 1$, and choosing $A = \frac{1}{2}\left(\frac{T-\tilde{T}}{\tilde{T}} - 1\right)$ we get $\frac{T-\tilde{T}}{\tilde{T}(1+A)} > 1$. As a consequence, $\nu = \frac{1}{2}$. Since $1 + \frac{(p-2)\nu}{4(p+\nu)} \wedge \frac{\eta}{2} \to \frac{9}{8}$ as $p \uparrow +\infty$, following Theorem 2.6 we choose the outer Monte-Carlo samples M_l as

$$M_l = M_0 2^{-\frac{9}{8}l}.$$

If $\tilde{T} \geq T/2$, then $\frac{T-\tilde{T}}{\tilde{T}(1+A)} < 1$ for any $A > 0$, yielding $\nu = \frac{T-\tilde{T}}{2\tilde{T}(1+A)}$. Since $1 + \frac{(p-2)\nu}{4(p+\nu)} \wedge \frac{\eta}{2} \to 1 + \frac{T-\tilde{T}}{8\tilde{T}(1+A)}$ as $p \uparrow +\infty$, again following Theorem 2.6 we can choose

$$M_l = M_0 2^{-\left(1+\frac{T-\tilde{T}}{8\tilde{T}(1+A)}\right)l}, \qquad A > 0.$$

For such choices of M_l, the overall computational complexity is $\mathbf{C}^{\mathrm{ML}} = \sum_{l=0}^{L} n_l M_l = \frac{1-2^{-(L+1)\nu/4}}{1-2^{-\nu/4}} M_0 n_0$ where $\nu = \frac{1}{2}$ if $\tilde{T} < \frac{T}{2}$ and $\nu = \frac{T-\tilde{T}}{2\tilde{T}(1+A)}$ if $\tilde{T} \geq \frac{T}{2}$. In our experiments, we will consider options with maturity $T = 1$ and set $\tilde{T} = \frac{252-5}{252}T$ (1 week before the maturity, 252 trading days in a year) and $A = 0.05$. As a consequence, we have $\tilde{T} \geq \frac{T}{2}$ and $\nu = \frac{T-\tilde{T}}{2\tilde{T}(1+A)}$.

Construction of the two other estimators. The NMC estimator (2.2) only depends on the two sample sizes N and M; its complexity is $\mathbf{C}^{\mathrm{NMC}} = MN$. The non-antithetic ML estimator ML2 is fixed by Propositions 2.9: recall that we make the choice $\tilde{M}_l = \tilde{M}_0 2^{-l}$, which implies that the product $\tilde{n}_l \tilde{M}_l = \tilde{n}_0 \tilde{M}_0$ is independent of l. The resulting computational complexity is $\mathbf{C}^{\mathrm{ML2}} = \tilde{n}_0 \tilde{M}_0(\tilde{L}+1)$.

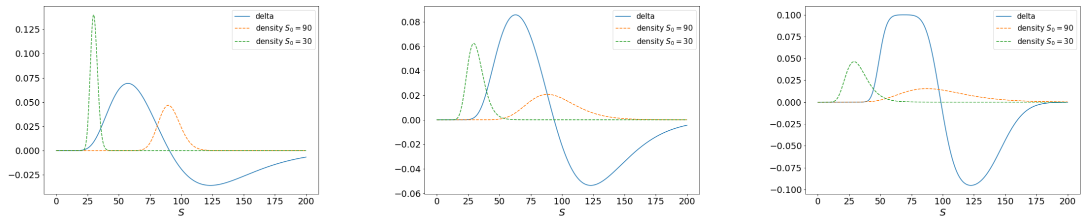
We fix the model parameters $r = 0.1$, $\sigma = 0.3$ in (3.1) and consider four different Portfolios $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$, described below. To estimate a reference value for I, we use an unbiased Monte-Carlo estimation of $\mathbb{E}\left[g(E_f(X))\right]$ (recall that $E_f(X)$ is explicit in the examples below) with $n_{\mathrm{MC}} = 5 \times 10^7$ samples, and provide the associated 95%-confidence interval. Denote $\Phi(s, K, a) = (s-(K-a))_+ + (s+(K+a))_+ - 2(s-K)_+$ the butterfly option payoff with strikes $K-a$, K, and $K+a$.

Portfolio $\mathcal{A}$. We consider one butterfly option with payoff $\Phi(s, 100, 50)$, and set $S_0 = 90$. Here $I_{\mathcal{A}} = \mathbb{E}\left[g(E_f^{butterfly}(X))\right]$ is worth 10.720 ± 0.002.

Portfolio $\mathcal{B}$. We choose Portfolio $\mathcal{A}$, but now with $S_0 = 30$, which implies $S_0 \ll K = 100$ and therefore the resulting samples will be far from the singularity of g with high probability (see Figure 2). Here $I_{\mathcal{B}} = 0.998 \pm 5 * 10^{-4}$.

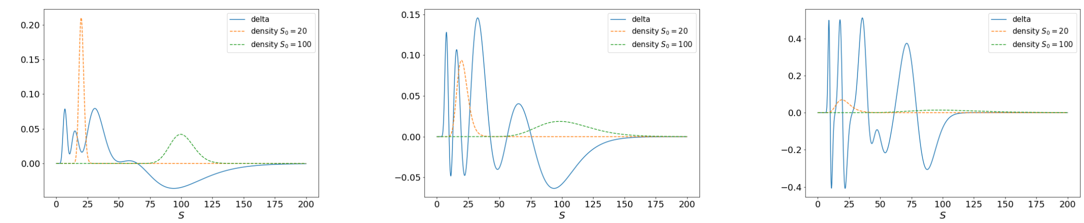
Portfolio $\mathcal{C}$. We choose a more diversified portfolio, made of a linear combination of five different butterfly options. The final payoff is of the form
$2\Phi(s, 10, 1) + 2\Phi(s, 20, 2) + 2\Phi(s, 40, 4) + \Phi(s, 50, 5) + 1.5\Phi(s, 80, 8)$. In Figure 3 one can observe that the Portfolio delta, as a function of s, now has several zeros (playing the role of singularities for $g(z) = |z|$). We set $S_0 = 20$ (close to the singularities) and get $I_{\mathcal{C}} = 0.507 \pm 0.0002$ as a reference value.

Portfolio $\mathcal{D}$. We choose Portfolio $\mathcal{C}$, but now setting $S_0 = 100$ (further away from the singularities of the option's delta, see Figure 3 again). Here $I_{\mathcal{D}} = 1.263 \pm 0.0004$.



(a) $t = 0.1T$ (b) $t = 0.5T$ (c) $t = 0.9T$

Figure 2: Functions $s \to 0.1\Delta_t(s)$ and $s \to p_t(s)$ for the Portfolios $\mathcal{A}, \mathcal{B}$ and different values of t.



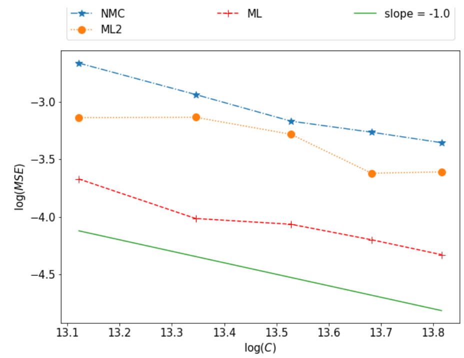
(a) $t = 0.1T$ (b) $t = 0.5T$ (c) $t = 0.9T$

Figure 3: Functions $s \to \Delta_t(s)$ and $s \to p_t(s)$ for the Portfolios $\mathcal{C}, \mathcal{D}$ and different values of t.

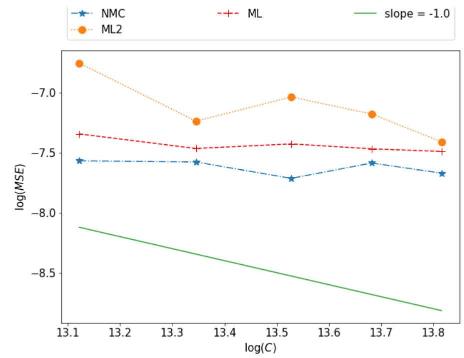
In Figures 2 and 3 we plot the Portfolio delta (as a function of s) for different times t, together with the probability density function of S_t, denoted p_t. Note that in Figure 2, we have rescaled the delta so to fit all the functions on the same graph. Observing the support of p_t, we see that in Portfolios $\mathcal{A}$ and $\mathcal{C}$ there is a high probability that $\Delta_t(S_t)$ will change sign, as opposed to Portfolios $\mathcal{B}$ and $\mathcal{D}$.

In Figure 4, we plot the **MSE**'s of the different estimators in terms of their respective computational costs, on a log-log plot. We have to fix their (free) parameters: (M, N) for the NMC estimator, (M_0, n_0, L) for the ML estimator and $\left(\tilde{M}_0, \tilde{n}_0, \tilde{L}\right)$ for the ML2 estimator. Recall from the analysis above that $\mathbf{C}^{\text{ML}} = \frac{1-2^{-(L+1)\nu/4}}{1-2^{-\nu/4}} M_0 n_0$, $\mathbf{C}^{\text{ML2}} = \tilde{n}_0 \tilde{M}_0(\tilde{L}+1)$ and $\mathbf{C}^{\text{NMC}} = MN$. For a fixed value of the cost $\mathbf{C}$ (here ranging from 5×10^5 to 1×10^6), we look for values of $M, M_0, \tilde{M}_0$ on a grid of evenly spaced values between 10^3 and 10^4 and for values of $L, \tilde{L}$ in $\{2, 3, 4, 5\}$ that minimize the estimated **MSE**. We fix the corresponding parameters $N, n_0, \tilde{n}_0$ using the identities

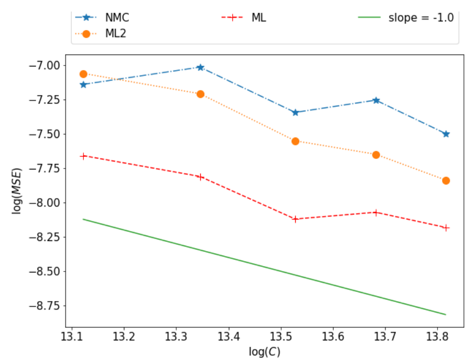
$$N = \frac{\mathbf{C}}{M}, \; n_0 = \frac{(1-2^{-\nu/4})\mathbf{C}}{(1-2^{-(L+1)\nu/4})M_0}, \; \tilde{n}_0 = \frac{\mathbf{C}}{(\tilde{L}+1)\tilde{M}_0}.$$



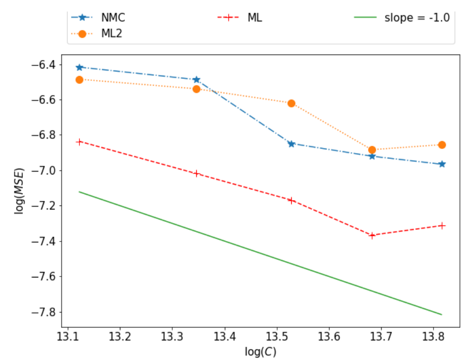
(a) Portfolio $\mathcal{A}$



(b) Portfolio $\mathcal{B}$



(c) Portfolio $\mathcal{C}$



(d) Portfolio $\mathcal{D}$

Figure 4: $\log(\mathbf{MSE})$ against log of the computational cost for the Portfolios $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$. ML stands for the antithetic Multi-level estimator (2.3), ML2 for the non-antithetic ML estimator (2.8), and NMC for the plain nested estimator (2.2).

We observe that the antithetic ML estimator gives the best results in terms of size of the **MSE** for a fixed cost $\mathbf{C}$ for the Portfolios $\mathcal{A}, \mathcal{C}$ and $\mathcal{D}$, for which we have singularities in the delta. We retrieve a slope close to -1 for the ML estimator, whereas the other estimators show a slightly smaller slope (in absolute value), which is in line with the theoretical results. On the other side, recall that Portfolio $\mathcal{B}$ was constructed in such a way that the probability of hitting the zeros of the delta function is very small, so that the function g essentialy does not introduce any bias. In this case, the three estimators provide similar results (and the antithetic estimator ML actually does not display the best results anymore): the use of a multilevel estimator in this setting does not seem to provide concrete advantage with respect to a plain nested MC procedure. In all the other situations, the antithetic MLMC estimator has the best performance.

3.3.2 Non-nested lower and upper estimators

With reference to the notation of Section 2.3 and Proposition 3.1, we have to evaluate the regression coefficients (2.16) and (2.17) with respect to $X = (U, Z) \in \mathbb{R}^2$. To do so, we choose as a projection basis the tensor product of (first-kind) Chebychev polynomials T_i and Hermite polynomials H_i, precisely: we set $L_{X,k_X} := L_{U,k_U} \otimes L_{Z,k_Z}$ for $k_U \geq 0$ and $k_Z \geq 0$, where

$$k_X := k_U \times k_Z, \quad L_{U,k_U} := \operatorname{span}\left(T_i(U/\tilde{T}) : 0 \leq i \leq k_U\right), \quad L_{Z,k_Z} := \operatorname{span}\left(H_i(Z) : 0 \leq i \leq k_Z\right),$$

where the polynomials T_i and H_i are defined by

$$T_0(x) = 1, \quad T_1(x) = x, \quad \forall i \geq 1, T_{i+1}(x) = 2xT_i(x) - T_{i-1}(x),$$
$$H_0(x) = 1, \quad H_1(x) = x, \quad \forall i \geq 1, H_{i+1}(x) = xH_i(x) - iH_{i-1}(x).$$

Choosing such smooth basis functions is motivated by the intuition that, in the present setting, $\mathbb{E}_f(X)$ is a reasonably smooth function of X. For every $0 \leq j_u \leq k_U$ and $0 \leq j_z \leq k_Z$, we solve the least-squares problems introduced in Section 2.3, the first one being

$$(l^*_{j_u j_z}) = \underset{\substack{l_{j_u j_z} \in \mathbb{R}, \\ 0 \leq j_u \leq k_U, 0 \leq j_z \leq k_Z}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left(f(\tilde{X}_i, \tilde{Y}_i) - \sum_{0 \leq j_u \leq k_U, 0 \leq j_z \leq k_Z} l_{j_u j_z} T_{j_u}(\tilde{U}_i/\tilde{T}) H_{j_z}(\tilde{Z}_i) \right)^2.$$

For the construction of the random variable ε^*_d, we use the same kind of basis for X, that is $L_{X,d_X} = L_{U,d_U} \otimes L_{Z,d_Z}$ (possibly with different values of $d_U \geq 0, d_Z \geq 0$), and choose as an approximation space for Y the space $L_{Y,d_Y} = \operatorname{span}(H_i(Y) : 0 \leq i \leq d_Y)$, $d_Y \geq 0$. For every $0 \leq a_u \leq d_U$, $0 \leq a_z \leq d_Z$ and $0 \leq b_y \leq d_Y$, we have to solve the least-squares problem

$$(u^*_{a_u a_z b_y}) = \underset{\substack{u_{a_u a_z b_y} \in \mathbb{R}, 0 \leq a_u \leq d_U, \\ 0 \leq a_z \leq d_Z, 0 < b_y \leq d_Y}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left(f(\tilde{X}_i, \tilde{Y}_i) - \sum_{\substack{0 \leq a_u \leq d_U, \\ 0 \leq a_z \leq d_Z, \\ 0 < b_y \leq d_Y}} u_{a_u a_z b_y} T_{a_u}(\tilde{U}_i/\tilde{T}) H_{a_z}\left(\tilde{Z}_i\right) H_{b_y}\left(\tilde{Y}_i\right) \right)^2,$$

where we have used the zero-mean property $H_{b_y}(\tilde{Y}) - \mathbb{E}\left[H_{b_y}(\tilde{Y})\right] = \mathbf{1}_{b_y \neq 0} H_{b_y}(\tilde{Y})$ following from the orthogonality of the Hermite polynomials with respect to the Gaussian measure [AS64, 22.2.15, p.775], which means that we can restrict the sum to $0 < b_y \leq d_Y$.

To construct the upper and lower biased estimators, we generate N independent copies $\left(\tilde{X}_i, \tilde{Y}_i\right)_{1 \leq i \leq N}$ of (X, Y) in order to solve both least-squares problems above, and define the regression functions φ^*_k and ε^*_d from (2.18). Once these functions are fixed, we generate another sample of $\tilde{N}$ independent copies of (X, Y) and estimate the left and right expectations in (2.19) with a standard Monte Carlo procedure.

Impact and choice thereof of the polynomial degrees. In Table 1, we list the resulting values of the lower/upper bounds together with their 95% confidence interval, for different choices of the polynomial degrees (k_U, k_Z) and (d_U, d_Z, d_Y). We have run the algorithm on call and put options with parameters $S_0 = K = 100$, $T = 1$, $r = 0.1$, $\sigma = 0.3$. In Table 2, we provide the corresponding lower/upper bounds obtained for Portfolios $\mathcal{A}$ and $\mathcal{C}$.

(k_U, k_Z)	Call	Put	(d_U, d_Z, d_Y)	Call	Put
$(1, 1)$	20.52 ± 0.033	9.415 ± 0.014	$(1, 1, 1)$	22.963 ± 0.028	11.066 ± 0.012
$(2, 2)$	20.551 ± 0.033	9.421 ± 0.014	$(2, 2, 2)$	22.461 ± 0.012	10.969 ± 0.006
$(3, 3)$	20.551 ± 0.033	9.396 ± 0.014	$(3, 3, 3)$	21.104 ± 0.009	10.177 ± 0.005
$(4, 4)$	20.554 ± 0.033	9.424 ± 0.015	$(4, 4, 4)$	20.909 ± 0.01	9.924 ± 0.005
$(5, 5)$	20.535 ± 0.073	9.409 ± 0.032	$(5, 5, 5)$	20.883 ± 0.01	9.859 ± 0.006
$(6, 6)$	20.548 ± 0.033	9.429 ± 0.015	$(6, 6, 6)$	20.874 ± 0.02	9.797 ± 0.009

Table 1: Lower/upper bounds for the values of $I^{call} = 20.567$ and $I^{put} = 9.433$. The sample sizes are $N = 10^5, \tilde{N} = 5 \times 10^6$.

(k_U, k_Z)	Portfolio $\mathcal{A}$	Portfolio $\mathcal{C}$	(d_U, d_Z, d_Y)	Portfolio $\mathcal{A}$	Portfolio $\mathcal{C}$
$(6,6)$	10.686 ± 0.036	0.467 ± 0.004	$(6,6,6)$	11.695 ± 0.013	1.424 ± 0.003
$(7,7)$	10.689 ± 0.036	0.495 ± 0.004	$(7,7,7)$	11.342 ± 0.038	1.415 ± 0.005
$(8,8)$	10.702 ± 0.036	0.498 ± 0.004	$(8,8,8)$	11.207 ± 0.017	1.383 ± 0.004
$(9,9)$	10.707 ± 0.036	0.503 ± 0.004	$(9,9,9)$	11.142 ± 0.04	1.375 ± 0.007

Table 2: Lower/upper bounds for the values of $I_{\mathcal{A}} = 10.720 \pm 0.002$ and $I_{\mathcal{C}} = 0.507 \pm 0.0002$. Sample sizes are $N = 10^6$, $\tilde{N} = 2 \times 10^6$.

Notice that an automatic choice of the degrees (k_U, k_Z) and (d_U, d_Z, d_Y) could be done through a cross-validation procedure so that the out-of-sample regression error is minimized, see [GKKW02, Chapter 8] and [FHT08, Chapter 8] for instance. A minimal out-of-sample regression error (a.k.a. generalization error) should intuitively give the best bounds. Here, since we evaluate both bounds at once, we can also make a choice based on the difference between their values: we choose polynomial degrees (k_U, k_Z) and (d_U, d_Z, d_Y) such that the distance between the two estimated bounds is minimized. In Table 1, this would correspond to the highest degree $d_U = d_Z = d_Y = 6$ for the upper bound; the lower bound does not seem to be very sensitive to the value of k_U, k_Z and is well estimated already with $k_U = k_Z = 1$. Using the reference explicit values of I^{call} and I^{put}, we can see that the results are overall very accurate; when $k_U = k_Z = d_U = d_Z = d_Y = 6$, the relative error between the two bounds is $\frac{20.874-20.548}{20.567} = 1.6\%$ for the call option and $\frac{9.797-9.429}{9.433} = 3.9\%$ for the put option. In Table 2, we observe a similar behaviour for Portfolio $\mathcal{A}$; here the relative error for the higher tested degree is $\frac{11.142-10.707}{10.720} = 4.0\%$. Note that we had to do use higher polynomial degrees and larger sample size N than in Table 1 in order to obtain comparable relative errors, which we can attribute to the more singular behavior of this portfolio. For Portfolio $\mathcal{C}$ (containing even more singularities), we see that the two bounds are still quite far from each other; notably, while the lower bounds provide accurate results, the upper bounds remain relatively far from the true value. Obtaining tighter bounds in this case would require higher values for N, $\tilde{N}$ and the polynomial degrees.

4 Conclusion

For the evaluation of $\mathbb{E}\left[g(\mathbb{E}\left[f(X,Y)|X\right])\right]$ with a piece-wise C^1 function g, we have designed an antithetic Multi-Level Monte-Carlo estimator and lower/upper biased non-nested Monte-Carlo estimators. Theoretical results support the convergence of all these estimators (Theorem 2.6 and Theorem 2.11). Globally, our numerical experiments show better performance of the antithetic MLMC estimator compared to the non-antithetic one and to a plain nested Monte-Carlo estimator (as expected). The lower/upper bounds can be implemented using empirical regression methods, and this leads to a scheme for "sandwiching" the quantity of interest, provided that the conditional expectation $x \mapsto \mathbb{E}\left[f(X,Y)|X=x\right]$ can be efficiently approximated over a space of basis functions.

A Technical proofs

A.1 Proof of Proposition 2.8 on the complexity of the NMC estimator

First notice that by independence, $\mathbb{E}\left[\hat{I}_{M,N}\right] = \mathbb{E}\left[g\left(\frac{1}{N}\sum_{j=1}^{N} f\left(X, Y_j\right)\right)\right]$ and by Proposition 2.4 there exists a positive constant κ s.t.

$$\forall N \geq 1, \qquad \left|\mathbb{E}\left[\hat{I}_{M,N} - g\left(E_f(X)\right)\right]\right| \leq \frac{\kappa}{N^{\frac{1}{2}\left(1+\frac{(p-1)\nu}{p+\nu} \wedge \eta\right)}}.$$

By Lemma 2.14, applying (2.20), we have that the identity

$$g\left(\hat{E}_{f,N}(X)\right) = g\left(E_f(X)\right) + \delta E_N(X) \int_0^1 \left(g'\left(E_f(X) + \lambda \delta E_N(X)\right)\right) \mathrm{d}\lambda$$

holds on the set $E_f(X) \notin \{d_1, \ldots, d_\theta\}$ with probability one by Assumption 2.3. Whence, we have :

$$\mathbb{E}\left[g\left(\hat{E}_{f,N}(X)\right)^2\right] \le 2\left(\mathbb{E}\left[g\left(E_f(X)\right)^2\right] + ||g'||_\infty^2 \mathbb{E}\left[\left(\delta E_N(X)\right)^2\right]\right) \le 2\left(\mathbb{E}\left[g\left(E_f(X)\right)^2\right] + ||g'||_\infty^2 \frac{C_p^2}{N}\right),$$

using Lemma 2.13. By independence, we get the following estimate for the variance :

$$\mathbf{Var}\left(\hat{I}_{M,N}\right) = \frac{1}{M}\mathbf{Var}\left(g\left(\hat{E}_{f,N}(X)\right)\right) \le \frac{1}{M}\mathbb{E}\left[g\left(\hat{E}_{f,N}(X)\right)^2\right] \le \frac{2}{M}\left(\mathbb{E}\left[g\left(E_f(X)\right)^2\right] + ||g'||_\infty^2 \frac{C_p^2}{N}\right).$$

Finally, by the bias-variance decomposition of $\mathbf{MSE}^{\mathrm{NMC}}$, we get :

$$\mathbf{MSE}^{\mathrm{NMC}} \le C_{NMC}\left(\frac{1}{N^{\left(1+\frac{(p-1)\nu}{p+\nu}\wedge\eta\right)}} + \frac{1}{MN} + \frac{1}{M}\right),$$

where $C_{NMC} = \kappa^2 \vee \left(2\mathbb{E}\left[g\left(E_f(X)\right)^2\right]\right) \vee \left(2||g'||_\infty^2 C_p^2\right)$. The statement on the complexity of the NMC estimator finally follows from the fact that the computational cost is of order $\mathcal{O}(MN)$. □

A.2 Proof of Proposition 2.9 on the complexity of the Multi level estimator ML2

Since g is Lipschitz continuous, proceeding as in the proof of Proposition 2.8, there exists a constant $C_p > 0$ such that

$$\mathbb{E}\left[\left(g(\hat{Z}) - g(\hat{Z}_1)\right)^2\right] \le ||g'||_\infty^2 \mathbb{E}\left[\left(\frac{\delta\hat{Z}_2 - \delta\hat{Z}_1}{2}\right)^2\right] \le ||g'||_\infty^2 \frac{C_p^2}{n_l}.$$

Now noticing that the variance at level l for the estimator (2.8) is given by

$$V_l = \mathbf{Var}\left(g\left(\frac{1}{n_l}\sum_{j=1}^{n_l} f\left(X, Y_j\right)\right) - g\left(\frac{1}{n_{l-1}}\sum_{j=1}^{n_{l-1}} f\left(X, Y_j\right)\right)\right) = \mathbf{Var}\left(g(\hat{Z}) - g(\hat{Z}_1)\right),$$

we conclude as in the Proof of Theorem 2.6, now using the estimate $V_l \le C n_l^{-1}$ for a positive constant C and the case $\beta = \gamma$ in [Gil15, Theorem 1]. □

A.3 Proof of Lemma 3.5

Since $\mathcal{N}(.)$ is increasing on $\mathbb{R}$, we have $\mathbb{P}\left(\mathcal{N}\left(\mu + \sigma Z\right) < z\right) = \mathcal{N}\left(\frac{\mathcal{N}^{-1}(z)-\mu}{\sigma}\right)$ for every $z > 0$. For given $A > 0$, define $\sigma' := (1 + A/2)^{1/2}\sigma$ and note that, owing to (3.13)),

$$\begin{aligned}
\frac{\mathcal{N}\left(\frac{\mathcal{N}^{-1}(z)-\mu}{\sigma}\right)}{\mathcal{N}\left(\frac{\mathcal{N}^{-1}(z)}{\sigma'}\right)} &\underset{z\to 0}{\sim} \frac{\sigma}{\sigma'} \frac{\exp\left(-\frac{1}{2}\left(\frac{\mathcal{N}^{-1}(z)-\mu}{\sigma}\right)^2\right)}{\exp\left(-\frac{1}{2}\left(\frac{\mathcal{N}^{-1}(z)}{\sigma'}\right)^2\right)} \frac{1}{1 - \frac{\mu}{\mathcal{N}^{-1}(z)}} \\
&\underset{z\to 0}{\sim} \frac{\sigma}{\sigma'} \exp\left(-\frac{1}{2}\left(\frac{\sigma' - \sigma}{\sigma\sigma'}\mathcal{N}^{-1}(z) - \frac{\mu}{\sigma}\right)\left(\frac{\sigma' + \sigma}{\sigma\sigma'}\mathcal{N}^{-1}(z) - \frac{\mu}{\sigma}\right)\right).
\end{aligned}$$

Then, using $\sigma' > \sigma > 0$ and $\mathcal{N}^{-1}(z) \xrightarrow[z\to 0^+]{} -\infty$, we obtain $\frac{\mathcal{N}\left(\frac{\mathcal{N}^{-1}(z)-\mu}{\sigma}\right)}{\mathcal{N}\left(\frac{\mathcal{N}^{-1}(z)}{\sigma'}\right)} \xrightarrow[z\to 0^+]{} 0$. Hence, there exists $\lambda_0(A) > 0$ s.t. for every $0 < z < \lambda_0(A)$:

$$\mathcal{N}\left(\frac{\mathcal{N}^{-1}(z)-\mu}{\sigma}\right) < \mathcal{N}\left(\frac{\mathcal{N}^{-1}(z)}{\sigma'}\right).$$

Now, using the fact for every $z \in (0, \frac{1}{2}]$, $\mathcal{N}^{-1}(z) \le 0$, we have that for every $z \in (0, 2^{-(1+A)\sigma^2}]$,

$$\mathcal{N}\left(\frac{\mathcal{N}^{-1}(z)}{\sigma'}\right) \le z^{\frac{1}{(1+A)\sigma^2}} \iff \frac{\mathcal{N}^{-1}(z)}{\sigma'} \le \mathcal{N}^{-1}\left(z^{\frac{1}{(1+A)\sigma^2}}\right) \le 0 \iff \left(\frac{\mathcal{N}^{-1}(z)}{\sigma'}\right)^2 \ge \left(\mathcal{N}^{-1}\left(z^{\frac{1}{(1+A)\sigma^2}}\right)\right)^2.$$

Now using the asymptotic expansion for the inverse c.d.f. of the standard normal distribution (see [Dom03]),

$$\left(\mathcal{N}^{-1}(z)\right)^2 \underset{z\to 0^+}{\sim} \log\left(\frac{1}{2\pi z^2}\right) - \log\left(\log\left(\frac{1}{2\pi z^2}\right)\right) \underset{z\to 0^+}{\sim} -2\log(z),$$

the following limit holds:

$$\frac{\left(\frac{\mathcal{N}^{-1}(z)}{\sigma'}\right)^2}{\left(\mathcal{N}^{-1}\left(z^{\frac{1}{(1+A)\sigma^2}}\right)\right)^2} \xrightarrow[z\to 0^+]{} \frac{1}{\frac{1}{(1+A)\sigma^2}(\sigma')^2} = \frac{1+A}{1+A/2} > 1.$$

We have just shown that there exists $\lambda_1(A) > 0$ s.t. for every $0 < z < \lambda_1(A)$, one has $\left(\frac{\mathcal{N}^{-1}(z)}{\sigma'}\right)^2 \ge \left(\mathcal{N}^{-1}\left(z^{\frac{1}{(1+A)\sigma^2}}\right)\right)^2$. Finally, choosing $z_0(A) := \lambda_0(A) \wedge \lambda_1(A) \wedge 2^{-\frac{1}{(1+A)\sigma^2}}$, we obtained the announced claim. $\square$

A.4 Proof of Proposition 3.7

$\triangleright$ Let us justify point 1. For notational simplicity, we omit the dependence with respect to t of the functions A and B. Taking the derivative of H_t yields

$$H_t'(x) = -Ae^{-\frac{A^2}{2}}e^{-Ax} + Be^{-\frac{B^2}{2}}e^{Bx}.$$

Now

$$H_t'(x) = 0 \iff x = x_0(t) := \frac{\log\left(\frac{A}{B}\right)}{A+B} - \frac{1}{2}(A-B), \tag{A.1}$$

$H_t'(x) > 0$ for $x > x_0(t)$ and $H_t'(x) < 0$ for $x < x_0(t)$. Recall that $A > B > 0$. Using $H_t(x) \xrightarrow[x\to\pm\infty]{} +\infty$, H_t is decreasing on $(-\infty, x_0(t))$ and increasing on $(x_0(t), +\infty)$. Furthermore, let us show that

$$H_t(x_0(t)) = e^{-A\frac{\log\left(\frac{A}{B}\right)}{A+B} - \frac{AB}{2}} + e^{B\frac{\log\left(\frac{A}{B}\right)}{A+B} - \frac{AB}{2}} - 2 < 0.$$

Let us define $y := \frac{A}{B} > 1$ so that $H_t(x_0(t)) = e^{-\frac{AB}{2}}\left(e^{-\frac{y\log(y)}{1+y}} + e^{\frac{\log(y)}{1+y}} - 2\right) := e^{-\frac{AB}{2}}(h(y) - 2)$. Since $e^{\frac{-AB}{2}} < 1$, it is now sufficient to prove that

$$\forall y > 1, \quad h(y) < 2.$$

Set $g(y) := \frac{y\log(y)}{1+y}$ so that $h(y) = \frac{1+y}{e^{\frac{y\log(y)}{1+y}}} = \frac{1+y}{e^{g(y)}}$. A direct computation yields $g'(y) = \frac{\log(y)}{(1+y)^2} + \frac{1}{(1+y)}$ and $h'(y) = e^{-g(y)}\left(1-(1+y)\,g'(y)\right) = \frac{-\log(y)e^{-g(y)}}{(1+y)} < 0$ for $y > 1$. Hence h is continuous and decreasing on $(1,\infty)$, thus strictly upper bounded by $h(1) = 2$, which is the announced claim.

Now using the intermediate value theorem, there exist only two roots $\alpha(t), \beta(t) \in \mathbb{R}$ of H_t. Summing up, we have :

$$\alpha(t) < x_0(t) < \beta(t), \qquad H_t(\alpha(t)) = H_t(\beta(t)) = 0.$$

Using $H_t(0) = e^{-\frac{A^2}{2}} + e^{-\frac{B^2}{2}} - 2 < 0$, we finally get:

$$\alpha(t) < 0 < \beta(t). \tag{A.2}$$

Let us now prove that $\alpha \in \mathcal{C}^\infty([0,T))$ (similar arguments can be used for β). Note that H which is $\mathcal{C}^\infty$ can be defined for negative t as well, say on $(-\eta, T)$ with $\eta > 0$ and we keep writing H for this extended defintion. The roots α and β are also well defined on $(-\eta, T)$. Now let $t_1 \in (-\eta, T)$ and using $\forall t \in (-\eta, T)\,, \quad x_0(t) \neq \alpha(t)$, we have:

$$H(t_1, \alpha(t_1)) = 0, \quad \frac{\partial H}{\partial x}(t_1, \alpha(t_1)) = -A(t_1)\,e^{-\frac{A(t_1)^2}{2}}e^{-A(t_1)\alpha(t_1)} + B(t_1)\,e^{-\frac{B(t_1)2}{2}}e^{B(t_1)\alpha(t_1)} \neq 0.$$

Using the implicit function theorem (see Thereom 10.2.2 in [Die90]), there exist neighborhoods $\mathcal{U}_{t_1}$ and $\mathcal{U}_{\alpha(t_1)}$ of t_1 and $\alpha(t_1)$, a function $\chi : \mathcal{U}_{t_1} \to \mathcal{U}_{\alpha(t_1)}$ of class $\mathcal{C}^\infty$ s.t.

$$\forall t \in \mathcal{U}_{t_1},\ H(t, \chi(t)) = 0.$$

Hence $\alpha(t_1) = \chi(t_1)$ and α is $\mathcal{C}^\infty$ at t_1 (for every $t_1 \in (-\eta, T)$), so that $\alpha \in \mathcal{C}^\infty((-\eta,T))$ and consequently $\alpha \in \mathcal{C}^\infty([0,T))$.

▷ We now prove point 2. Let us denote $d_1(t,s)$ instead of $d_1(t,T,s,K)$. Notice that $\forall t \in [0,T),\ s \in (0,\infty) \to d_1(t,s) = \frac{\log\left(\frac{s}{K}\right)+\left(r+\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}$ is increasing and:

$$\begin{aligned} x = d_1(t,s) &\Longleftrightarrow s = Ke^{\sigma\sqrt{T-t}x-\left(r+\frac{\sigma^2}{2}\right)(T-t)}, \\ \Gamma_t(S_t) := \Delta_t'(S_t) &\overset{(3.21)}{=} \delta_t'\left(d_1(t,T,S_t,K)\right)\frac{1}{S_t\sigma\sqrt{T-t}}. \end{aligned} \tag{A.3}$$

In view of the relation (3.12) (i.e. $\sqrt{2\pi}\delta_t'(x) = e^{-\frac{y^2}{2}}H_t(x)$), the zeros $\alpha(t), \beta(t)$ of $H_t(.)$ transfer to those of $\delta_t'(d_1(t,.))$ (and thus $\Gamma_t(.)$) via (A.3). This leads to the announced expression (3.23) of two zeros $\tilde{\alpha}(t), \tilde{\beta}(t)$ for $\Gamma_t(.)$. The time-smoothness of $\alpha(t), \beta(t)$ is clearly transferred to that of $t \to \tilde{\alpha}(t), \tilde{\beta}(t)$.

▷ Finally, let us prove point 3. The function $x \to \delta_t(x)$ is increasing on $(-\infty, \alpha(t)) \cup (\beta(t), \infty)$ and decreasing on $(\alpha(t), \beta(t))$ since H_t is positive on $(-\infty, \alpha(t)) \cup (\beta(t), \infty)$, negative on $(\alpha(t), \beta(t))$ and has two unique zeros $\alpha(t), \beta(t)$. Using that $s \to d_1(t,s)$ is an increasing function and by (3.21), we conclude that $s \to \Delta_t(s)$ is increasing on $(0, \tilde{\alpha}(t)) \cup \left(\tilde{\beta}(t), \infty\right)$ and decreasing on $\left(\tilde{\alpha}(t), \tilde{\beta}(t)\right)$. Noticing that, $\forall t \in [0,T)$, $\Delta_t(s) \xrightarrow[s\to 0^+]{} 0$ and $\Delta_t(s) \xrightarrow[s\to+\infty]{} 0$, by the intermediate value theorem, there exists a unique function $t \in [0,T) \to \gamma(t)$ s.t.:

$$\forall t \in [0,T), \quad \tilde{\alpha}(t) < \gamma(t) < \tilde{\beta}(t), \quad \Delta_t(\gamma(t)) = 0.$$

Also, using the monotonicity of $\mathcal{N}$, the definitions (3.9) and $\delta_t(0) = \Delta_t\left(Ke^{-\left(r+\frac{\sigma^2}{2}\right)(T-t)}\right) = \mathcal{N}(A(t)) - \mathcal{N}(B(t)) > 0$, we obtain:

$$\forall t \in [0,T),\ \tilde{\alpha}(t) < Ke^{-\left(r+\frac{\sigma^2}{2}\right)(T-t)} < \gamma(t).$$

It remains to justify that $t \to \gamma(t) \in \mathcal{C}^\infty([0,T))$: this can be done using the implicit function theorem, exactly as we have done for point 1. $\square$

A.5 Proof of Proposition 3.9

By Proposition 3.7, $s \in \mathcal{V}(t) \to \Delta_t(s)$ is decreasing and takes its values in $\mathcal{W}(t)$, hence we can define Δ_t^{-1} (the inverse function of $\Delta_t(.)$) and Δ_t is a C^1- diffeomorphism from $\mathcal{V}(t)$ to $\mathcal{W}(t)$. Keeping on with the notation p_t used in Lemma 3.6, the density χ_t restricted to $\mathcal{W}(t)$ is defined as:

$$\forall t \in [0,T),\ \forall y \in \mathcal{W}(t), \quad \chi_t(y) = \left.\frac{p_t(s)}{|\Delta_t'(s)|}\right|_{s=\Delta_t^{-1}(y)},$$

where for the sake of clarity we write $d_1(t,s)$ instead of $d_1(t,T,s,K)$ and (see (A.3) and (3.12))

$$|\Delta_t'(s)| = \frac{e^{-\frac{d_1(t,s)^2}{2}}}{s\sigma\sqrt{T-t}\sqrt{2\pi}} |H_t(d_1(t,s))|.$$

Proposition 3.7 gives us all information about the sign and the variation of $H_t(.)$. Combining the previous remark and the fact that $s \to d_1(t,s)$ is increasing on $(0,\infty)$, we easily deduce that for $t \in [0,\tilde{T}]$ the function $s \to |H_t(d_1(t,s))|$ is increasing on $(\tilde{\alpha}(t), \tilde{x_0}(t))$ and decreasing on $\left(\tilde{x_0}(t), \tilde{\beta}(t)\right)$ where $\tilde{x}_0(t) := Ke^{\sigma\sqrt{T-t}x_0(t)-\left(r+\frac{\sigma^2}{2}\right)(T-t)}$. Consequently,

$$\forall s \in \mathcal{V}(t),\ |H_t(d_1(t,s))| \geq |H_t(d_1(t,\gamma(t)+\epsilon(t)))| \wedge |H_t(d_1(t,\gamma(t)-\epsilon(t)))| := C_1(t).$$

Observe that on $[0,\tilde{T}]$, $C_1(.)$ is continuous and positive because of the choice of $\epsilon(t)$.

Furthermore, using that $s \in (0,\infty) \to e^{-\frac{d_1(t,s)^2}{2}}$ is increasing on $\left(0, Ke^{-(r+\frac{\sigma^2}{2})(T-t)}\right)$, decreasing on $\left(Ke^{-(r+\frac{\sigma^2}{2})(T-t)}, \infty\right)$ and the inequality $1/(s\sqrt{T-t}) \geq 1/(\sqrt{T}(\gamma(t)+\epsilon(t)))$ on $\mathcal{V}(t)$, we have:

$$\forall s \in \mathcal{V}(t), \quad \frac{e^{-\frac{d_1(t,s)^2}{2}}}{s\sigma\sqrt{T-t}\sqrt{2\pi}} \geq \frac{e^{-\frac{d_1(t,\gamma(t)+\epsilon(t))^2}{2}} \wedge e^{-\frac{d_1(t,\gamma(t)-\epsilon(t))^2}{2}}}{\sqrt{T}(\gamma(t)+\epsilon(t))\sigma\sqrt{2\pi}} := C_2(t).$$

It is clear that $C_2(.)$ is also continuous and positive on $[0,\tilde{T}]$. Owing to Lemma 3.6, we finally obtain the following upper bound for χ_t :

$$\forall y \in \mathcal{W}(t),\ \chi_t(y) \leq \frac{e^{\left(\sigma^2-r\right)t}}{S_0\sigma\sqrt{2\pi}\sqrt{t}} \frac{1}{C_1(t)C_2(t)} := C(t).$$

Summing up, we have shown that the function $C(.)$ is bounded by $1/\sqrt{t}$ up to a constant. Therefore, $C(.)$ is in $L^1([0,\tilde{T}])$, and this concludes the proof. □

References

[ADG+19] A. Agarwal, S. De Marco, E. Gobet, J. López-Salas, F. Noubiagain, and A. Zhou. Numerical approximations of McKean anticipative backward stochastic differential equations arising in Initial Margin requirements. *ESAIM: Proceedings And Surveys*, 65:1–26, Feb. 2019. 1, 3

[AS64] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964. 3.3.2

[Avi09] R. Avikainen. On irregular functionals of SDEs and the Euler scheme. *Finance and Stochastics*, 13(3):381–401, Sep 2009. 2.6, 2.6, 2.6

[BDM15] M. Broadie, Y. P. Du, and C. C. Moallemi. Risk estimation via regression. *Operations Research*, 63(5):1077–1097, 2015. 1

[BG96] M. Broadie and P. Glasserman. Estimating security price derivatives using simulation. *Management science*, 42(2):269–285, 1996. 3

[BHR15] K. Bujok, B. M. Hambly, and C. Reisinger. Multilevel simulation of functionals of Bernoulli random variables with application to basket credit derivatives. *Methodology and Computing in Applied Probability*, 17(3):579–604, Sep 2015. 1

[Die90] J. Dieudonné. *Eléments d'analyse*. Jacques Gabay, Paris, 1990. Tome 1, (chapitres I à XI), Fondements de l'Analyse moderne. A.4

[Dom03] D. E. Dominici. The inverse of the cumulative standard normal probability function. *Integral Transforms and Special Functions*, 14(4):281–292, 2003. A.3

[FHT08] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning: Data Mining, Inference, and Prediction*. Springer series in statistics New York, second edition, 2008. 3.3.2

[GG19] M.B. Giles and T. Goda. Decision-making under uncertainty: using MLMC for efficient estimation of EVPPI. *Statistics and Computing*, 29(4):739–751, 2019. 1

[GH19] M.B. Giles and A.-L. Haji-Ali. Multilevel nested simulation for efficient risk estimation. *SIAM/ASA Journal on Uncertainty Quantification*, 7(2):497–525, 2019. 1

[Gil08] M.B. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008. (document), 1

[Gil15] M.B. Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015. 1, 1, 2.1, 2.1, 2.6, A.2

[GJ10] M. B. Gordy and S. Juneja. Nested simulation in portfolio risk measurement. *Management Science*, 56(10):1833–1848, 2010. 1

[GKKW02] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics, 2002. 2.3, 3.3.2

[GL18] I. Guo and G. Loeper. Pricing bounds for volatility derivatives via duality and least squares Monte Carlo. *Journal of Optimization Theory and Applications*, 179(2):598–617, 2018. 2.3

[GLP18] D. Giorgi, V. Lemaire, and G. Pagès. Weak error for nested multilevel Monte Carlo. *preprint hal-01817386*, June 2018. 1, 2

[GM12] E. Gobet and A. Makhlouf. The tracking error rate of the Delta-Gamma hedging strategy. *Mathematical Finance*, 22(2):277–309, 2012. 3.1

[Gor41] R. D. Gordon. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941. 3

[HH80] P. Hall and C. C. Heyde. *Martingale Limit Theory and its Application*. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press, 1980. 2.4

[HK04] M.B. Haugh and L. Kogan. Pricing American options: A duality approach. *Operations Research*, 52(2):258–270, 2004. 2.3

[HL17] P. Henry-Labordere. Deep primal-dual algorithm for BSDEs: application of Machine Learning to CVA and IM. *Available at SSRN https://ssrn.com/abstract=3071506*, 2017. 1