How many inner simulations to compute conditional expectations with least-square Monte Carlo?

Aurélien Alfonsi, Bernard Lapeyre, Jérôme Lelong

How many inner simulations to compute conditional expectations with least-square Monte Carlo?

Aurélien Alfonsi* Bernard Lapeyre† Jérôme Lelong‡

May 11, 2023

Abstract

The problem of computing the conditional expectation $\mathbb{E}[f(Y)|X]$ with least-square Monte-Carlo is of general importance and has been widely studied. To solve this problem, it is usually assumed that one has as many samples of Y as of X. However, when samples are generated by computer simulation and the conditional law of Y given X can be simulated, it may be relevant to sample $K \in \mathbb{N}$ values of Y for each sample of X. The present work determines the optimal value of K for a given computational budget, as well as a way to estimate it. The main take away message is that the computational gain can be all the more important as the computational cost of sampling Y given X is small with respect to the computational cost of sampling X. Numerical illustrations on the optimal choice of K and on the computational gain are given on different examples including one inspired by risk management.

Keywords: Least square Monte-Carlo, Conditional expectation estimators, Variance reduction
AMS 2020: 65C05, 91G60
Acknowledgement: AA and BL acknowledge the support of the chaire "Risques financiers", Fondation du Risque.
Declarations:
• **Competing interests:** The authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.
• **Data availability:** Not applicable.

*CERMICS, Ecole des Ponts, Marne-la-Vallée, France. MathRisk, Inria, Paris, France.
email: aurelien.alfonsi@enpc.fr
†CERMICS, Ecole des Ponts, Marne-la-Vallée, France. MathRisk, Inria, Paris, France.
email: bernard.lapeyre@enpc.fr
‡Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.
email: jerome.lelong@univ-grenoble-alpes.fr

1 Introduction and Framework

We consider the classical problem of computing a conditional expectation using a least-square Monte Carlo approach. To be more precise, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $X : \Omega \to \mathbb{R}^d$ and $Y : \Omega \to \mathbb{R}^p$ be two random variables and $f : \mathbb{R}^p \to \mathbb{R}$ be a measurable function such that $\mathbb{E}[f(Y)^2] < \infty$. We are interested in computing $\mathbb{E}[f(Y)|X]$ by using a parametrized approximation. Thus, we introduce a family of measurable functions $(\varphi(\theta, \cdot))_{\theta \in \mathbb{R}^q}$ from $\mathbb{R}^d$ to $\mathbb{R}$ satisfying for all $\theta \in \mathbb{R}^q$, $\mathbb{E}[\varphi(\theta, X)^2] < \infty$. This family will be used to approximate the conditional expectation $\mathbb{E}[f(Y)|X]$. It is well known that $\mathbb{E}[f(Y)|X]$ solves the two following minimisation problems

$$\inf_{Z \in L^2(\Omega, \sigma(X))} \mathbb{E}[(Z - f(Y))^2], \quad \inf_{Z \in L^2(\Omega, \sigma(X))} \mathbb{E}[(Z - \mathbb{E}[f(Y)|X])^2],$$

where $L^2(\Omega, \sigma(X))$ denotes the set of square integrable random variables that are measurable with respect to the σ-algebra generated by X. Therefore, we are interested in the following minimization problems

$$\inf_{\theta \in \mathbb{R}^q} \mathbb{E}[(\varphi(\theta, X) - f(Y))^2], \quad \inf_{\theta \in \mathbb{R}^q} \mathbb{E}[(\varphi(\theta, X) - \mathbb{E}[f(Y)|X])^2]. \tag{1}$$

In practical cases, all these expectations are not explicit and it is often used Monte-Carlo estimators to approximate them. The classical problem of regression consists in minimizing $\frac{1}{N}\sum_{i=1}^N (\varphi(\theta, X_i) - f(Y_i))^2$ with respect to θ, where $(X_i, Y_i)_{i \geq 1}$ is a sequence of iid random variables with the same distribution as (X, Y). In this work, we consider the possibility of having for each X_i many samples of Y given X_i. This is the case when samples are generated by computer simulation and when the conditional law of Y given X can be simulated. More precisely, let $(X_i)_{i \geq 1}$ be a sequence of iid random variables following the distribution of X. For each $i \geq 1$, we introduce independent sequences $(Y_i^{(k)})_{k \geq 1}$ of iid random variables following the law $\mathcal{L}(Y|X = X_i)$ of Y conditionally on $X = X_i$. For $N, K \in \mathbb{N}^*$, we define the sequence of functions $v_N^K : \mathbb{R}^q \to \mathbb{R}$ by

$$v_N^K(\theta) = \frac{1}{N}\sum_{i=1}^N \left(\varphi(\theta, X_i) - \frac{1}{K}\sum_{k=1}^K f(Y_i^{(k)})\right)^2. \tag{2}$$

We are interested in finding θ_N^K minimizing v_N^K, so that $\varphi(\theta_N^K, X)$ will give an approximation of $\mathbb{E}[f(Y)|X]$. Formally, the two minimisation problems of Equation (1) correspond respectively to $N = \infty, K = 1$ and $N = K = \infty$. Note that the minimisation of (2) with $K = 1$ corresponds to the classical case, with as many samples of Y_i as of X_i. Up to our knowledge, most of the literature (if not all) considers the case of minimizing v_N^1 to approximate the conditional expectation, and we refer to Gyorfi et al. [GKKW02] for a nice presentation of the topic and references. This may be understood from the point of view of statistics: on empirical data, one usually have as many observations of X's and Y's. However, when X and Y are generated by computer simulation, it is relevant to consider the possibility of sampling $K \geq 2$ values of Y

for a given X. The natural question then is to understand how to choose N and K in order to achieve the best accuracy for a given computational time. This is the goal of this paper.

The problem of computing conditional expectations is an important problem that arises in many different fields of research, such as the approximation of backward stochastic differential equations [BT04, GLW05], the pricing of American options and more generally optimal stopping problems [LS15], and stochastic optimal control problems [BKS10] to mention a few. It has a particular relevance in risk management, see e.g. [BBR09, BDM15, KNK18], where financial institutions have to evaluate risk from a regulatory perspective. The valuation of future risks naturally involves conditional expectations. To be more precise, let us consider the case of insurance companies that have to calculate their Solvency Capital Requirement (SCR). This SCR can be calculated by computing expected losses under some stressed scenarios. This regulatory procedure to evaluate risk is called the "standard formula". If one aims at evaluating the SCR at a future date T with the same procedure, one has to compute conditional expected losses under the different stressed scenarios, given all the market information between the current date and T, see [ACIA21]. Therefore, one naturally has to deal with the numerical approximation of conditional expectations. Let us stress that it is usually natural in this context to be able to sample conditional laws: assets are usually modeled by a Markovian process that can be simulated, and we can then simulate as many paths as desired after T, from a given path up to time T.

Many works have developed numerical methods based on nested simulations and refinements to approximate an expectation that involves conditional expectations. Among them we can cite [GJ10] which optimize nested simulations to estimate a value at risk on a conditional expectation, [BRS12] which study nested simulations in the context of risk insurance modeling and [ACIA21] which use a multilevel approach on the same kind of insurance problem. But to the best of our knowledge, none of these works are interested in the nested approximation of conditional expectations using a parametric representation as done in this work.

The paper is structured as follows. First, Section 2 presents our main assumptions, under which we are able to show, by quite standard arguments, the convergence of θ_N^K as well as a Central Limit Theorem. Section 3 presents the main results of the paper. In particular, Theorem 3.2 gives a precise asymptotic of the suboptimality of θ_N^K (with respect to $\theta^\star$) as a function of K and N, and the optimal value of K for a given computational budget. It also gives estimators to approximate it. The computational gain is all the more important as the computational cost of sampling Y given X is small and the approximation family is close to the conditional expectation. Section 4 gives a focus on the particularly important case of linear regression. We are able to refine the result of Theorem 3.2 in this context. Besides, Proposition 4.3 shows for a particular choice of approximating functions that the optimal value of K can be arbitrarily large. Finally, Section 5 presents numerical results and shows the relevance of considering $K > 1$ on different examples. We also compare different estimators that approximate K, and it comes out that one estimator is more relevant for practical use.

2 Assumptions and Convergence results

In this section, we apply the general results on the convergence of the estimators of the optimal solutions presented by [RS93, Section 2.6]. We introduce the function

$$v^{\infty}(\theta) = \mathbb{E}\left[(\varphi(\theta, X) - \mathbb{E}[f(Y)|X])^2\right] \tag{3}$$

and make the following assumptions.

Assumptions Let $C \subset \mathbb{R}^q$ be a compact set with a non-empty interior $\mathring{C}$.

($\mathcal{H}$-1) Uniform integrability: $\mathbb{E}\left[\sup_{\theta \in C} |\varphi(\theta, X)|^2\right] < \infty$.

($\mathcal{H}$-2) The function $\theta \longmapsto \varphi(\theta, X)$ is a.s. continuous on C.

($\mathcal{H}$-3) The function v^{∞} admits on C a unique minimizer $\theta^{\star} \in \mathring{C}$.

($\mathcal{H}$-4) The function $\theta \longmapsto \varphi(\theta, X)$ is a.s. twice continuously differentiable on C and such that

$$\mathbb{E}\left[\sup_{\theta \in C} |\nabla \varphi(\theta, X)|^2\right] < \infty; \quad \mathbb{E}\left[\sup_{\theta \in C} \left|\nabla^2 \varphi(\theta, X)\right|^2\right] < \infty.$$

Here, and in the whole paper, the gradient ∇ is taken with respect to θ. Let us note that Hypotheses ($\mathcal{H}$-1), ($\mathcal{H}$-2) and ($\mathcal{H}$-4) are satisfied in the case of the linear regression, see Section 4 for further details.

To apply the results on the convergence of the estimators presented by [RS93, Section 2.6], we introduce the function

$$\Phi(\theta, Z) = \left(\varphi(\theta, X) - \frac{1}{K}\sum_{k=1}^{K} f(Y^{(k)})\right)^2,$$

with $Z = (X, Y^{(1)}, \ldots, Y^{(K)})$, where the sequence $(Y^{(k)})_{k \geq 1}$ is conditionally iid given X, and given $X = x$ follows the distribution $\mathcal{L}(Y|X = x)$[1]. We also define, for $K \in \mathbb{N}^*$, the function

$$v^K(\theta) = \mathbb{E}\left[\left(\varphi(\theta, X) - \frac{1}{K}\sum_{k=1}^{K} f(Y^{(k)})\right)^2\right], \tag{4}$$

and $v_N^K(\theta) = \frac{1}{N}\sum_{i=1}^{N} \Phi(\theta, Z_i)$, so that $v^K(\theta) = \mathbb{E}[v_N^K(\theta)]$. Since $|\Phi(\theta, Z)| \leq 2|\varphi(\theta, X)|^2 + \frac{2}{K}\sum_{k=1}^{K} f(Y^{(k)})^2$, we get the uniform integrability on C by using ($\mathcal{H}$-1). The continuity of Φ with respect to $\theta \in C$ is clear by ($\mathcal{H}$-2), and we get the following lemma from the uniform law of large numbers, see Lemma A.1.

[1] In practice, we typically have $Y = F(X, U)$ with U independent of X and F a measurable function, and the conditional independence means that $Y^{(k)} = F(X, U^{(k)})$ with $X, U^{(1)}, \ldots, U^{(K)}$ independent and $\mathcal{L}(U^{(i)}) = \mathcal{L}(U)$.

Lemma 2.1 *Under ($\mathcal{H}$-1) and ($\mathcal{H}$-2), for every fixed $K \in \mathbb{N}$, $\sup_{\theta\in C} |v_N^K(\theta) - v^K(\theta)| \to 0$ almost surely as $N \to \infty$.*

The next lemma makes explicit the link between $v^\infty(\theta)$ and $v^K(\theta)$ defined respectively by (3) and (4).

Lemma 2.2 *We have for all $\theta \in C$,*

$$v^K(\theta) = v^\infty(\theta) + \frac{1}{K}\,\mathbb{E}[(f(Y) - \mathbb{E}[f(Y)|X])^2]. \tag{5}$$

Proof. We expand (4) and get

$$v^K(\theta) = \frac{1}{K}\,\mathbb{E}\left[(\varphi(\theta,X) - f(Y))^2\right] + \frac{1}{K^2}\sum_{k\neq k'}\mathbb{E}\left[\left(\varphi(\theta,X) - f(Y^{(k)})\right)\left(\varphi(\theta,X) - f(Y^{(k')})\right)\right].$$

On the one hand, using the conditional independence of the $Y^{(k)}$'s, we get for $k \neq k'$

$$\begin{aligned}
&\mathbb{E}\left[\left(\varphi(\theta,X) - f(Y^{(k)})\right)\left(\varphi(\theta,X) - f(Y^{(k')})\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\varphi(\theta,X) - f(Y^{(k)})|X\right]\mathbb{E}\left[\varphi(\theta,X) - f(Y^{(k')})|X\right]\right] \\
&= \mathbb{E}\left[(\varphi(\theta,X) - \mathbb{E}[f(Y)|X])^2\right].
\end{aligned}$$

On the other hand, as the conditional expectation is an orthogonal projection,

$$\begin{aligned}
\mathbb{E}\left[(\varphi(\theta,X) - f(Y))^2\right] &= \mathbb{E}\left[((\varphi(\theta,X) - \mathbb{E}[f(Y)|X]) + (\mathbb{E}[f(Y)|X] - f(Y)))^2\right] \\
&= \mathbb{E}\left[(\varphi(\theta,X) - \mathbb{E}[f(Y)|X])^2\right] + \mathbb{E}\left[(\mathbb{E}[f(Y)|X] - f(Y))^2\right].
\end{aligned}$$

This yields to the claim. ∎

Let θ_N^K (resp. $\theta^\star$) be a minimizer of v_N^K (resp. v^∞) on the compact set C, i.e.

$$v_N^K(\theta_N^K) = \inf_{\theta\in C} v_N^K(\theta) \quad \text{and} \quad v^\infty(\theta^\star) = \inf_{\theta\in C} v^\infty(\theta).$$

By Lemma 2.2, v^K and v^∞ differ only by a constant. So, $\theta^\star$ is also the unique minimizer of v^K for every K. Therefore, we have the following result from [RS93, Theorem A1, p. 67].

Proposition 2.3 *Under ($\mathcal{H}$-1), ($\mathcal{H}$-2), ($\mathcal{H}$-3), for every fixed K, $\theta_N^K \to \theta^\star$ a.s. when $N \to \infty$.*

Beside this almost sure convergence result, we also have a central limit theorem under additional assumptions.

Proposition 2.4 *Under the assumptions of Proposition 2.3, ($\mathcal{H}$-4) and if* $\mathbb{E}\left[\left(\varphi(\theta^\star,X)-\frac{1}{K}\sum_{k=1}^K f(Y^{(k)})\right)^2|\nabla\varphi(\theta^\star,X)|^2\right]<\infty$ *and the matrix* $H:=\nabla^2 v^\infty(\theta^\star)$ *is positive definite, we have*

$$\sqrt{N}(\theta_N^K-\theta^\star)\xrightarrow[N\to\infty]{\mathcal{L}}\mathcal{N}(0,4H^{-1}\Gamma^K H^{-1}) \tag{6}$$

with

$$\Gamma^K=A+B/K, \tag{7}$$

where $A,B\in\mathbb{R}^{q\times q}$ *are the following semi-definite positive matrices:*

$$A=\mathbb{E}\left[(\varphi(\theta^\star,X)-\mathbb{E}[f(Y)|X])^2\nabla\varphi(\theta^\star,X)\nabla\varphi(\theta^\star,X)^T\right], \tag{8}$$

$$B=\mathbb{E}\left[(f(Y)-\mathbb{E}[f(Y)|X])^2\nabla\varphi(\theta^\star,X)\nabla\varphi(\theta^\star,X)^T\right]. \tag{9}$$

Furthermore, we have

$$N(v^\infty(\theta_N^K)-v^\infty(\theta^\star))\xrightarrow[N\to\infty]{\mathcal{L}}2G^THG \text{ with } G\sim\mathcal{N}(0,H^{-1}\Gamma^K H^{-1}). \tag{10}$$

Proof. First, we check some properties on gradients. We have $\nabla\Phi(\theta,Z)=2(\varphi(\theta,X)-\frac{1}{K}\sum_{k=1}^K f(Y^{(k)}))\nabla\varphi(\theta,X)$ and $\nabla^2\Phi(\theta,Z)=2\nabla\varphi(\theta,X)\nabla\varphi(\theta,X)^T+2(\varphi(\theta,X)-\frac{1}{K}\sum_{k=1}^K f(Y^{(k)}))\nabla^2\varphi(\theta,X)$. From Cauchy-Schwarz inequality, ($\mathcal{H}$-4) and $\mathbb{E}[f(Y)^2]<\infty$, we get that $\sup_{\theta\in C}|\nabla\Phi(\theta,Z)|$ and $\sup_{\theta\in C}|\nabla^2\Phi(\theta,Z)|$ are integrable. Besides, the matrix

$$\Gamma^K=\mathbb{E}\left[\left(\varphi(\theta^\star,X)-\frac{1}{K}\sum_{k=1}^K f(Y^{(k)})\right)^2\nabla\varphi(\theta^\star,X)\nabla\varphi(\theta^\star,X)^T\right]. \tag{11}$$

is well defined since $\mathbb{E}\left[\left(\varphi(\theta^\star,X)-\frac{1}{K}\sum_{k=1}^K f(Y^{(k)})\right)^2|\nabla\varphi(\theta^\star,X)|^2\right]<\infty$, and we get (6) following the result from [RS93, Theorem A2, p. 74].

Let us check that $\Gamma^K=A+B/K$. We have

$$\begin{aligned}
\Gamma^K&=\frac{1}{K^2}\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{k=1}^K\varphi(\theta^\star,X)-f(Y^{(k)})\right)^2|X\right]\nabla\varphi(\theta^\star,X)\nabla\varphi(\theta^\star,X)^T\right]\\
&=\frac{1}{K}\mathbb{E}\left[\mathbb{E}\left[(\varphi(\theta^\star,X)-f(Y))^2|X\right]\nabla\varphi(\theta^\star,X)\nabla\varphi(\theta^\star,X)^T\right]\\
&\quad+\frac{K-1}{K}\mathbb{E}\left[(\varphi(\theta^\star,X)-\mathbb{E}[f(Y)|X])^2\nabla\varphi(\theta^\star,X)\nabla\varphi(\theta^\star,X)^T\right]\\
&=\mathbb{E}\left[(\varphi(\theta^\star,X)-\mathbb{E}[f(Y)|X])^2\nabla\varphi(\theta^\star,X)\nabla\varphi(\theta^\star,X)^T\right]\\
&\quad+\frac{1}{K}\mathbb{E}\left[(f(Y)-\mathbb{E}[f(Y)|X])^2\nabla\varphi(\theta^\star,X)\nabla\varphi(\theta^\star,X)^T\right]=A+\frac{B}{K}.
\end{aligned}$$

Last, we have $v^\infty(\theta)-v^\infty(\theta^\star)=\frac{1}{2}(\theta-\theta^\star)^TH(\theta-\theta^\star)+|\theta-\theta^\star|^2\varepsilon(|\theta-\theta^\star|)$ with $\varepsilon(h)\to0$ as $h\to0$. By Slutsky's theorem, Proposition 2.3 and (6), we get (10). ■

Remark 2.5 Note that the matrix A defined by (8) corresponds to the asymptotic variance of the optimal regression that we would obtain if we could directly sample $\mathbb{E}[f(Y)|X]$. The additional term B/K in the decomposition of Γ^K is the extra variance generated by the Monte Carlo approximation of $\mathbb{E}[f(Y)|X]$.

Unless in very specific cases where the function v^∞ is explicit, it is impossible in practice to numerically evaluate $N(v^\infty(\theta_N^K) - v^\infty(\theta^\star))$. The next proposition shows that $N(v_N^K(\theta^\star) - v_N^K(\theta_N^K))$ has the same asymptotics as (10). Roughly speaking, the suboptimality of $\theta^\star$ for v_N^K is of the same order as the suboptimality of θ_N^K for v^∞. This result will be used in the numerical section 5 to illustrate the convergence.

Proposition 2.6 *Under the same assumptions as in Proposition 2.4 and if C is convex, we have*

$$N(v_N^K(\theta^\star) - v_N^K(\theta_N^K)) \xrightarrow[N\to\infty]{\mathcal{L}} 2G^T H G, \text{ with } G \sim \mathcal{N}(0, H^{-1}\Gamma^K H^{-1}).$$

Proof. We have by Taylor's theorem

$$v_N^K(\theta^\star) - v_N^K(\theta_N^K) = (\theta_N^K - \theta^\star)^T \left(\int_0^1 (1-u) \nabla^2 v_N^K \left(\theta_N^K + u(\theta_N^K - \theta^\star) \right) du \right) (\theta_N^K - \theta^\star), \quad (12)$$

with

$$\nabla^2 v_N^K(\theta) = \frac{2}{N} \sum_{i=1}^N \nabla\varphi(\theta, X_i) \nabla\varphi(\theta, X_i)^T + \left(\varphi(\theta, X_i) - \frac{1}{K} \sum_{k=1}^K f(Y_i^{(k)}) \right) \nabla^2 \varphi(\theta, X_i).$$

By Lemma 2.2, we have

$$\nabla^2 v^K(\theta) = \nabla^2 v^\infty(\theta) = 2\mathbb{E}\left[\nabla\varphi(\theta, X) \nabla\varphi(\theta, X)^T + (\varphi(\theta, X) - \mathbb{E}[f(Y)|X]) \nabla^2 \varphi(\theta, X) \right].$$

By ($\mathcal{H}$-1), ($\mathcal{H}$-2) and ($\mathcal{H}$-4), we can apply [RS93, Lemma A1 p. 67] and get that $\sup_{\theta\in C} |\nabla^2 v_N^K(\theta) - \nabla^2 v^\infty(\theta)| \underset{N\to\infty}{\to} 0$, almost surely. Since $\theta^\star, \theta_N^K \in C$ and C is convex, we get $\int_0^1 (1-u) \nabla^2 v_N^K \left(\theta_N^K + u(\theta_N^K - \theta^\star) \right) du - \int_0^1 (1-u) \nabla^2 v^\infty \left(\theta_N^K + u(\theta_N^K - \theta^\star) \right) du \to 0$, almost surely. Since $\nabla^2 v^\infty$ is bounded on C, we get

$$\int_0^1 (1-u) \nabla^2 v_N^K \left(\theta_N^K + u(\theta_N^K - \theta^\star) \right) du \underset{N\to\infty}{\to} H = \nabla^2 v^K(\theta^\star), a.s.,$$

by using Proposition 2.3. This gives that $N(v_N^K(\theta^\star) - v_N^K(\theta_N^K))$ converges in law to $2G^T H G$ by using (12), Proposition 2.4 and Slutsky's theorem. ■

3 Main results

In this section, we present our main theorem that determines the optimal allocation between N and K to approximate the conditional expectation. Let us denote the computational time for

sampling X and $\mathcal{L}(Y|X)$ respectively by C_X and $C_{Y|X}$. With these notations, the cost for computing v_N^K is proportional to $NC_X + NKC_{Y|X}$. Without loss of generality, we will assume that $C_X = 1$. This means that the computational time for sampling X is one unit and that we express all the other computational times with respect to this unit.

We now discuss the computational cost of calculating θ_N^K by using a gradient descent type method. Let us observe from the definition of v_N^K in Equation (2) that $\left(\frac{1}{K}\sum_{k=1}^K f(Y_i^{(k)})\right)_{1\leq i\leq N}$ can be computed once and for all. Then, the gradient descent applied to v_N^K has exactly the same computational cost as the one applied to v_N^1. This is why we do not include the cost of calculating θ_N^K in our reasoning and only focus on the computational cost of v_N^K.

3.1 Optimal allocation between N and K

Definition 3.1 *For $x > 0$, we denote by $\nu(x) \in \mathbb{N}^*$ the unique natural number such that*

$$(\nu(x) - 1)\nu(x) < x \leq \nu(x)(\nu(x) + 1).$$

It is easy to check that $\forall x > 0, \lfloor\sqrt{x}\rfloor \leq \nu(x) \leq \lceil\sqrt{x}\rceil$. Now, we state our main result.

Theorem 3.2 *Under the assumptions of Proposition 2.4 and if the sequence $N(v^\infty(\theta_N^K) - v^\infty(\theta^\star))_{N\geq 1}$ is uniformly integrable, we have*

$$\mathbb{E}[v^\infty(\theta_N^K)] = v^\infty(\theta^\star) + \frac{\mathrm{tr}(\Gamma^K H^{-1})}{N} + o(1/N),$$

as $N \to \infty$. If $A \neq 0$, the asymptotic optimal choice minimizing $\mathbb{E}[v^\infty(\theta_N^K)]$ for a computational budget $c \to \infty$ is to take

$$N^\star = \left\lfloor \frac{c}{1 + K^\star C_{Y|X}} \right\rfloor,\ K^\star = \nu\left(\frac{\mathrm{tr}(BH^{-1})}{C_{Y|X}\,\mathrm{tr}(AH^{-1})}\right).$$

Note that if $A = 0$ and $\mathbb{P}(\nabla\varphi(\theta^\star, X) \neq 0) > 0$, then $\varphi(\theta^\star, X) = \mathbb{E}[f(Y)|X]$. The condition $A \neq 0$ ensures that $\mathrm{tr}(AH^{-1}) > 0$.

Proof. We have by Proposition 2.4, $N(v^\infty(\theta_N^K) - v^\infty(\theta^\star)) \xrightarrow[N\to\infty]{\mathcal{L}} 2G^T HG$ with $G \sim \mathcal{N}(0, H^{-1}\Gamma^K H^{-1})$. From this convergence in distribution and the uniform integrability assumption, we get $\mathbb{E}[N(v^\infty(\theta_N^K) - v^\infty(\theta^\star))] \to 2\mathbb{E}[G^T HG]$. Let $C = \sqrt{H^{-1}\Gamma^K H^{-1}}$. Then, G has the same law as $C\tilde{G}$ with $\tilde{G} \sim \mathcal{N}(0, I_q)$ and thus $\mathbb{E}[G^T HG] = \mathbb{E}[\tilde{G}^T CHC\tilde{G}] = \mathrm{tr}(CHC) = \mathrm{tr}(HC^2) = \mathrm{tr}(\Gamma^K H^{-1})$, which gives the first claim.

As $N \to \infty$, the minimization of $\mathbb{E}[v^\infty(\theta_N^K)]$ with respect to K amounts to minimizing $\mathrm{tr}(\Gamma^K H^{-1})$ with respect to K. For a large enough budget c, the problem becomes

$$\inf_{\substack{N,K\in\mathbb{N} \\ s.t.\ N+NKC_{Y|X}=c}} \frac{\mathrm{tr}(AH^{-1})}{N} + \frac{\mathrm{tr}(BH^{-1})}{KN}.$$

Then, we apply Lemma A.2 to get the claim. ∎

Remark 3.3 Theorem 3.2 gives the asymptotic optimal allocation to minimize $\mathbb{E}[v^\infty(\theta_N^K)]$. Unfortunately, it involves the matrix H which is in general unknown and may be difficult to estimate. When θ is a one dimensional parameter, A, B and H are scalar values and thus $K^\star = \nu\left(\frac{B}{C_{Y|X}A}\right)$. Otherwise, since H is a definite positive matrix, we have $\underline{\lambda}_H I_q \leq H \leq \overline{\lambda}_H I_q$ and thus

$$\overline{\lambda}_H^{-1} \operatorname{tr}(\Gamma^K) \leq \operatorname{tr}(\Gamma^K H^{-1}) \leq \underline{\lambda}_H^{-1} \operatorname{tr}(\Gamma^K).$$

Therefore, it is reasonable (though not optimal) to minimize $\operatorname{tr}(\Gamma^K)$ under the same computational budget constraint, which then leads to

$$N' = \left\lfloor \frac{c}{1 + K' C_{Y|X}} \right\rfloor, \; K' = \nu\left(\frac{\operatorname{tr}(B)}{C_{Y|X} \operatorname{tr}(A)}\right).$$

The next corollary gives a bound on the computational gain that can be obtained by the optimization of K given by Theorem 3.2.

Corollary 3.4 (Comparison of the estimators $\theta_{N^\star}^{K^\star}$ and θ_N^1 for a fixed computational budget) *Let c be the computational budget. Under the assumptions of Theorem 3.2 and with $N = \lfloor c/(1 + C_{Y|X}) \rfloor$, we have*

$$\mathbb{E}[v^\infty(\theta_{N^\star}^{K^\star})] - v^\infty(\theta^\star) \sim_{c\to\infty} r^\star (\mathbb{E}[v^\infty(\theta_N^1)] - v^\infty(\theta^\star)),$$

with

$$r^\star = \frac{(1 + \nu(\xi) C_{Y|X}) \left(1 + \frac{\xi}{\nu(\xi)} C_{Y|X}\right)}{(1 + C_{Y|X})(1 + \xi C_{Y|X})}, \; \xi = \frac{\operatorname{tr}(BH^{-1})}{C_{Y|X} \operatorname{tr}(AH^{-1})}. \tag{13}$$

This multiplicative gain $r^\star \in (0, 1]$ satisfies $r^\star \geq \frac{C_{Y|X}}{1+C_{Y|X}}$ and $\lim_{\xi\to\infty} r^\star = \frac{C_{Y|X}}{1+C_{Y|X}}$.

Note that if $C_{Y|X} = 1$, i.e. the computation time of sampling $\mathcal{L}(Y|X)$ is the same as the one of sampling X, we cannot reduce the computation time more than by a factor $1/2$. Besides, the smaller is $C_{Y|X}$, the more we may hope a significant reduction of computational time, and this really occurs if ξ is large, so that $r^\star \approx \frac{C_{Y|X}}{1+C_{Y|X}}$.

Proof. Since we are comparing θ_N^1 and $\theta_{N^\star}^{K^\star}$ for the same computation budget, we have $c \sim_{c\to\infty} N(1 + C_{Y|X}) \sim_{c\to\infty} N^\star(1 + K^\star C_{Y|X})$. By Theorem 3.2, the multiplicative gain in precision is

$$r^\star = \frac{\operatorname{tr}(\Gamma^{K^\star} H^{-1})/N^\star}{\operatorname{tr}(\Gamma^1 H^{-1})/N} \underset{c\to\infty}{\to} \frac{\operatorname{tr}(\Gamma^{K^\star} H^{-1})}{\operatorname{tr}(\Gamma^1 H^{-1})} \times \frac{1 + K^\star C_{Y|X}}{1 + C_{Y|X}}.$$

Since $\Gamma^K = A + B/K$ and $K^\star = \nu(\xi)$, we get (13) after simple calculations. We have $r^\star \leq 1$ since $\nu(\xi) + \frac{\xi}{\nu(\xi)} \leq 1 + \xi$, $r^\star \geq \frac{C_{Y|X}}{1+C_{Y|X}}$ since $\nu(\xi) \geq 1$ and $\lim_{\xi\to\infty} r^\star = \frac{C_{Y|X}}{1+C_{Y|X}}$ since $\nu(\xi) \sim_{\xi\to\infty} \sqrt{\xi}$. ∎

Remark 3.5 By the same reasoning as in the proof of Corollary 3.4, we can define

$$r^K = \frac{\operatorname{tr}(\Gamma^K H^{-1})}{\operatorname{tr}(\Gamma^1 H^{-1})} \times \frac{1 + K C_{Y|X}}{1 + C_{Y|X}} \tag{14}$$

as the multiplicative gain resulting from using θ_N^K instead of θ_N^1. Note that this is indeed a gain if $r^K < 1$ and that we have $r^\star = r^{K^\star}$.

3.2 Estimation of the matrices A and B

In practice, to calculate the value of $K^\star$ given by Theorem 3.2, we need to estimate the matrices A and B. Let $(X_i, Y_i^{(1)}, \dots, Y_i^{(K)})_i$ be iid samples such that for all i, $X_i \sim X$ and $Y_i^{(k)} \sim \mathcal{L}(Y|X = X_i)$ for $k = 1, \dots, K$ being sampled independently given X_i. From these samples, we can compute θ_N^K and define

$$\hat{\Gamma}_N^K = \frac{1}{N}\sum_{i=1}^{N}\left(\varphi(\theta_N^K, X_i) - \frac{1}{K}\sum_{k=1}^{K} f(Y_i^{(k)})\right)^2 \nabla\varphi(\theta_N^K, X_i)\nabla\varphi(\theta_N^K, X_i)^T. \tag{15}$$

Proposition 3.6 *Assume ($\mathcal{H}$-1), ($\mathcal{H}$-2), ($\mathcal{H}$-3), ($\mathcal{H}$-4) and*

$$\mathbb{E}\left[\sup_{\theta\in C}|\varphi(\theta, X)\nabla\varphi(\theta, X)|^2\right] < \infty, \quad \mathbb{E}\left[f(Y)^2 \sup_{\theta\in C}|\nabla\varphi(\theta, X)|^2\right] < \infty. \tag{16}$$

Then, we have $\hat{\Gamma}_N^K \underset{N\to\infty}{\to} \Gamma^K$ almost surely.

Proof. We define the function $g_K : \mathbb{R}^q \times \mathbb{R}^d \times (\mathbb{R}^p)^K \to \mathbb{R}^{q\times q}$ by

$$g(\theta, x, (y^{(k)})_{1\le k\le K}) = \left(\varphi(\theta, x) - \frac{1}{K}\sum_{k=1}^{K} f(y^{(k)})\right)^2 \nabla\varphi(\theta, x)\nabla\varphi(\theta, x)^T.$$

Assumptions ($\mathcal{H}$-1), ($\mathcal{H}$-2), ($\mathcal{H}$-4) ensure that the function $\theta \mapsto g_K(\theta, X, (Y^{(k)})_{1\le k\le K})$ is a.s. continuous, while Assumption (16) gives the integrability of $\sup_{\theta\in C}|g_K(\theta, X, (Y^{(k)})_{1\le k\le K})|$. From Lemma A.1, we get that

$$\sup_{\theta\in C}\left|\frac{1}{N}\sum_{i=1}^{N} g_K(\theta, X_i, (Y_i^{(k)})_{1\le k\le K}) - \mathbb{E}[g_K(\theta, X, (Y^{(k)})_{1\le k\le K})]\right| \to 0, \ a.s.$$

From Proposition 2.3, $\theta_N^K \to \theta^\star$ a.s. Hence, we deduce that $\hat{\Gamma}_N^K \to \Gamma^K$ a.s. ■

Estimators for A and B From Proposition 3.6 and Equation (7), we deduce estimators of A and B. For $K_1, K_2 \in \mathbb{N}^*$ such that $K_1 < K_2$, we have by Proposition 3.6 when N tends to $+\infty$

$$\frac{K_2\hat{\Gamma}_N^{K_2} - K_1\hat{\Gamma}_N^{K_1}}{K_2 - K_1} \to A \ a.s., \ \frac{K_1K_2\left(\hat{\Gamma}_N^{K_1} - \hat{\Gamma}_N^{K_2}\right)}{K_2 - K_1} \to B \ a.s.$$

We will mainly use $K_1 = \bar{K}$ and $K_2 = 2\bar{K}$ for a given $\bar{K} \in \mathbb{N}^*$, which leads to simpler formulas

$$\hat{A}_{\bar{K}} := 2\hat{\Gamma}_N^{2\bar{K}} - \hat{\Gamma}_N^{\bar{K}}, \hat{B}_{\bar{K}} = 2\bar{K}\left(\hat{\Gamma}_N^{\bar{K}} - \hat{\Gamma}_N^{2\bar{K}}\right).$$

Besides, we rather work with the following antithetic estimators:

$$\hat{A}^{anti}_{\bar{K}} = \frac{1}{N}\sum_{i=1}^{N}\left[2\left(\varphi(\theta_N^{2\bar{K}},X_i) - \frac{1}{2\bar{K}}\sum_{k=1}^{2\bar{K}} f(Y_i^{(k)})\right)^2 - \frac{1}{2}\left(\varphi(\theta_N^{2\bar{K}},X_i) - \frac{1}{\bar{K}}\sum_{k=1}^{\bar{K}} f(Y_i^{(k)})\right)^2 \right.$$
$$\left. - \frac{1}{2}\left(\varphi(\theta_N^{2\bar{K}},X_i) - \frac{1}{\bar{K}}\sum_{k=\bar{K}+1}^{2\bar{K}} f(Y_i^{(k)})\right)^2\right]\nabla\varphi(\theta_N^{2\bar{K}},X_i)\nabla\varphi(\theta_N^{2\bar{K}},X_i)^T$$

$$\hat{B}^{anti}_{\bar{K}} = 2\bar{K}\left(\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\left(\varphi(\theta_N^{2\bar{K}},X_i) - \frac{1}{\bar{K}}\sum_{k=1}^{\bar{K}} f(Y_i^{(k)})\right)^2 + \frac{1}{2}\left(\varphi(\theta_N^{2\bar{K}},X_i) - \frac{1}{\bar{K}}\sum_{k=\bar{K}+1}^{2\bar{K}} f(Y_i^{(k)})\right)^2 \right.\right.$$
$$\left.\left. - \left(\varphi(\theta_N^{2\bar{K}},X_i) - \frac{1}{2\bar{K}}\sum_{k=1}^{2\bar{K}} f(Y_i^{(k)})\right)^2\right]\nabla\varphi(\theta_N^{2\bar{K}},X_i)\nabla\varphi(\theta_N^{2\bar{K}},X_i)^T\right).$$

Note that the same value of $\theta_N^{2\bar{K}}$ is used. Similarly, we have the almost sure convergence of these estimators respectively to A and B as $N \to \infty$. Thanks to the convexity of the square function, $\hat{B}^{anti}_{\bar{K}}$ is a semi-definite positive matrix. Unfortunately, the matrix $\hat{A}^{anti}_{\bar{K}}$ may not be semi-definite positive. More generally, the matrix A is in general difficult to estimate. From its definition (8), we see that the better is the approximation family $\varphi(\theta, X)$, the smaller is the matrix A for the natural order (Löwner order). Thus, when the conditional expectation is well approximated, the matrix A is small and may be smaller than the noise in $O(N^{-1/2})$, so that the estimated matrix $\hat{A}^{anti}_{\bar{K}}$ may have negative eigenvalues. Thus, in practice, we use

$$\hat{K}^A_H = \nu\left(\frac{\operatorname{tr}(\hat{B}^{anti}_{\bar{K}}\hat{H}^{-1})}{C_{Y|X}\operatorname{tr}((\hat{A}^{anti}_{\bar{K}}\hat{H}^{-1})_+)}\right) \tag{17}$$

to approximate $K^\star$. An alternative is to approximate A by $\Gamma_N^{2\bar{K}}$ for a (fixed) large value of $\bar{K}$: it is a nonnegative estimator of $\Gamma = A + B/\bar{K} \geq A$, and therefore

$$\hat{K}^\Gamma_H = \nu\left(\frac{\operatorname{tr}(\hat{B}^{anti}_{\bar{K}}\hat{H}^{-1})}{C_{Y|X}\operatorname{tr}(\Gamma_N^{2\bar{K}}\hat{H}^{-1})}\right) \tag{18}$$

underestimates $K^\star$. These estimators are discussed and illustrated in the numerical section 5.

4 The linear regression framework

In this section, we rephrase some results of Section 3 in the framework of linear regression as they actually take simpler forms. In particular, we show that the uniform integrability assumption of Theorem 3.2 is always satisfied.

4.1 Main results for the linear regression framework

We consider in this section a function $u : \mathbb{R}^d \to \mathbb{R}^q$ such that $\mathbb{E}[|u(X)|^2] < \infty$ and

$$\varphi(\theta, X) = \theta \cdot u(X),\ \theta \in \mathbb{R}^q.$$

In this case, we have $\nabla\varphi(\theta, X) = u(X)$, $\nabla^2\varphi(\theta, X) = 0$, $\nabla v^\infty(\theta) = 2\mathbb{E}[(\theta \cdot u(X) - \mathbb{E}[f(Y)|X])u(X)]$ and $\nabla^2 v^\infty(\theta) = 2\mathbb{E}[u(X)u^T(X)]$ does not depend on θ. Therefore, Assumptions ($\mathcal{H}$-1), ($\mathcal{H}$-2) and ($\mathcal{H}$-4) are clearly satisfied for any compact C, while ($\mathcal{H}$-3) holds if, and only if

$$H = 2\mathbb{E}[u(X)u^T(X)] \text{ is positive definite and } \theta^\star = 2H^{-1}\mathbb{E}[f(Y)u(X)] \in \mathring{C}. \qquad (\mathcal{H}\text{-3-lin})$$

We also get a simpler expression for θ_N^K and $\hat{\Gamma}_N^K$:

$$\theta_N^K = \left(\frac{1}{N}\sum_{i=1}^N u(X_i)u(X_i)^T\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N\left(\frac{1}{K}\sum_{k=1}^K f(Y_i^{(k)})\right)u(X_i)\right),$$
$$\hat{\Gamma}_N^K = \frac{1}{N}\sum_{i=1}^N\left(\theta_N^K \cdot u(X_i) - \frac{1}{K}\sum_{k=1}^K f(Y_i^{(k)})\right)^2 u(X_i)u(X_i)^T.$$

Here, we assume that H is positive definite and N is large enough so that $\frac{1}{N}\sum_{i=1}^N u(X_i)u(X_i)^T$ is positive definite by the law of large numbers. However, it may be convenient to slightly modify the estimator as in the next proposition. This new estimator satisfies in particular the uniform integrability assumption of Theorem 3.2, as shown in the proof of Proposition 4.2.

Definition 4.1 *For a positive semi-definite matrix $S \in \mathbb{R}^{q\times q}$ and $\epsilon \in \mathbb{R}_+^*$, $S \vee (\epsilon I_q)$ is the positive definite matrix such that $(S \vee (\epsilon I_q))e_l = \max(\lambda_l, \epsilon)e_l$, where $(e_l)_{1\le l\le q}$ is an orthonormal basis of eigenvectors with respective eigenvalues $(\lambda_l)_{1\le l\le q}$.*

Proposition 4.2 *We assume ($\mathcal{H}$-3-lin), $\mathbb{E}[|u(X)|^{4+\eta}] < \infty$ and $\mathbb{E}[f(Y)^{2+\eta}|u(X)|^{2+\eta}] < \infty$ for some $\eta > 0$. Let $\epsilon > 0$ be such that $H - 2\epsilon I_q$ is positive definite and define*

$$\theta_N^{K,\epsilon} = 2\left(\left(\frac{2}{N}\sum_{i=1}^N u(X_i)u(X_i)^T\right) \vee (\epsilon I_q)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N\left(\frac{1}{K}\sum_{k=1}^K f(Y_i^{(k)})\right)u(X_i)\right).$$

Then, we have $\theta_N^{K,\epsilon} \to \theta^\star$ a.s., $\sqrt{N}(\theta_N^{K,\epsilon} - \theta^\star) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0, 4H^{-1}\Gamma^K H^{-1})$ and

$$\mathbb{E}[v^\infty(\theta_N^K)] \underset{N\to\infty}{=} v^\infty(\theta^\star) + \frac{\mathrm{tr}(\Gamma^K H^{-1})}{N} + o(1/N).$$

The conclusions of Theorem 3.2 hold.

Proof. By the law of large numbers, $\frac{2}{N}\sum_{i=1}^N u(X_i)u(X_i)^T \to H$, almost surely. Since $H - 2\epsilon I_q$ is positive definite, there exists, almost surely, $\bar{N}$ such that for $N \geq \bar{N}$, $\frac{2}{N}\sum_{i=1}^N u(X_i)u(X_i)^T - \epsilon I_q$ is positive definite and thus $\theta_N^{K,\epsilon} = \theta_N^K$. This gives $\theta_N^{K,\epsilon} \to \theta^\star$ a.s. by Proposition 2.3 and $\sqrt{N}(\theta_N^{K,\epsilon} - \theta^\star) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0, 4H^{-1}\Gamma^K H^{-1})$ by Proposition 2.4.

Now, we check the uniform integrability of the sequence $N|\theta_N^{K,\epsilon} - \theta^\star|^2$. We have

$$\theta_N^{K,\epsilon} - \theta^\star = 2\left[\left(\left(\frac{2}{N}\sum_{i=1}^N u(X_i)u(X_i)^T\right) \vee (\epsilon I_q)\right)^{-1} - H^{-1}\right]\mathbb{E}[f(Y)u(X)]$$
$$+ 2\left(\left(\frac{2}{N}\sum_{i=1}^N u(X_i)u(X_i)^T\right) \vee (\epsilon I_q)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N\left(\frac{1}{K}\sum_{k=1}^K f(Y_i^{(k)})\right)u(X_i) - \mathbb{E}[f(Y)u(X)]\right)$$

Note that for two symmetric matrices M_1, M_2 such that $M_1 - \epsilon I_q$ and $M_2 - \epsilon I_q$ are definite positive, we have $|M_1^{-1} - M_2^{-1}| = |M_1^{-1}(M_2 - M_1)M_2^{-1}| \leq \frac{1}{\epsilon^2}|M_2 - M_1|$. Thus, we obtain

$$\begin{aligned}|\theta_N^{K,\epsilon} - \theta^\star| \leq & \frac{2}{\epsilon^2}\left|\left(\frac{2}{N}\sum_{i=1}^N u(X_i)u(X_i)^T\right) \vee (\epsilon I_q) - H\right| |\mathbb{E}[f(Y)u(X)]| \\ & + \frac{2}{\epsilon}\left|\frac{1}{N}\sum_{i=1}^N\left(\frac{1}{K}\sum_{k=1}^K f(Y_i^{(k)})\right)u(X_i) - \mathbb{E}[f(Y)u(X)]\right|.\end{aligned}$$

Then, Lemma A.3 with the assumptions on the moments gives the uniform integrability of the sequence $(N|\theta_N^{K,\epsilon} - \theta^\star|^2)_{N\geq 1}$. Last, observe that $0 \leq v^\infty(\theta) - v^\infty(\theta^\star) = (\theta - \theta^\star)^T H(\theta - \theta^\star) \leq |H||\theta - \theta^\star|^2$. Therefore, the sequence $N(v^\infty(\theta_N^{K,\epsilon}) - v^\infty(\theta^\star))$ is uniformly integrable, and we get $\mathbb{E}[v^\infty(\theta_N^K)] \underset{N\to\infty}{=} v^\infty(\theta^\star) + \frac{\mathrm{tr}(\Gamma^K H^{-1})}{N} + o(1/N)$ as in the proof of Theorem 3.2. ■

4.2 Piecewise constant approximation framework

We now specify our results in the linear case when X takes its values in $[0,1]^d$, with $q = M^d$ and the basis

$$u_n(x) = \prod_{j=1}^d \mathbf{1}_{I_{a_j}}(x_j), \tag{19}$$

for $n - 1 = a_1 + a_2 M + \cdots + a_d M^{d-1}$ with $a_1, \ldots, a_d \in \{0, \ldots, M-1\}$, and $I_a = [\frac{a}{M}, \frac{a+1}{M})$ for $a = 0, \ldots, M-2$ and $I_{M-1} = [\frac{M-1}{M}, 1]$.

The next proposition shows that with this choice of basis, the optimal number of inner simulations is (under suitable assumptions) at least of order M. This illustrates that the sharper is the family $\varphi(\theta, x)$ to approximate the conditional expectation $\mathbb{E}[f(Y)|X]$, the larger is the optimal number of inner simulations.

Proposition 4.3 *Let us assume that $\mathbb{E}[f(Y)|X] = \psi(X)$ with $\psi : [0,1]^d \to \mathbb{R}$ being a Lipschitz function with Lipschitz constant L. Let us assume that $\mathbb{E}[f(Y)^2|X] - (\mathbb{E}[f(Y)|X])^2 = \sigma^2(X)$*

for a function $\sigma : [0,1]^d \to [\underline{\sigma}, +\infty)$ for some $0 < \underline{\sigma} < \infty$. We consider $\varphi(\theta, X) = \theta \cdot u(X)$ with $\theta \in \mathbb{R}^q$, $q = M^d$ with the basis defined by (19)*.*

Then, $\mathrm{tr}(AH^{-1}) \leq \frac{1}{2}L^2 M^{d-2}$ and $\frac{1}{2}\underline{\sigma}^2 M^d \leq \mathrm{tr}(BH^{-1})$. In particular, $\frac{\mathrm{tr}(BH^{-1})}{\mathrm{tr}(AH^{-1})} \geq \underline{\sigma}^2 M^2$ and thus $K^\star \geq \gamma M$ for some $\gamma > 0$, with $K^\star$ given by Theorem 3.2.

Proof. We have $u_n(x) = \mathbf{1}_{C_n}(x)$, with $C_n = I_{a_1} \times \cdots \times I_{a_d}$. Since $C_n \cap C_{n'} = \emptyset$ for $n \neq n'$, we have that $u(x)u(x)^T$ is a diagonal matrix. Then, the matrices H, A and B are diagonal and we get:

$$H_{nn} = 2\mathbb{P}(X \in C_n),\ A_{nn} = \mathbb{E}[(\theta^\star \cdot u(X) - \mathbb{E}[f(Y)|X])^2 \mathbf{1}_{X \in C_n}],$$
$$B_{nn} = \mathbb{E}[(f(Y) - \mathbb{E}[f(Y)|X])^2 \mathbf{1}_{X \in C_n}] = \mathbb{E}[\mathbb{E}[f(Y)^2|X] - (\mathbb{E}[f(Y)|X])^2 \mathbf{1}_{X \in C_n}].$$

Therefore, we get $\mathrm{tr}(AH^{-1}) = \frac{1}{2}\sum_{n=1}^q \mathbb{E}[(\theta^\star \cdot u(X) - \mathbb{E}[f(Y)|X])^2 | X \in C_n]$ and $\mathrm{tr}(BH^{-1}) = \frac{1}{2}\sum_{n=1}^q \mathbb{E}[\mathbb{E}[f(Y)^2|X] - (\mathbb{E}[f(Y)|X])^2 | X \in C_n] = \frac{1}{2}\sum_{n=1}^q \mathbb{E}[\sigma(X)^2 | X \in C_n] \geq \frac{\underline{\sigma}^2 q}{2}$. Besides, we observe that $\theta_n^\star = \mathbb{E}[\psi(X)|X \in C_n]$ since $\theta^\star$ minimizes $v^\infty(\theta) = \sum_{n=1}^q \mathbb{E}[\mathbf{1}_{C_n}(X)(\theta_n - \psi(X))^2]$, and therefore $|\theta_n^\star - \psi(x)| \leq L/M$ for $x \in C_n$ by the triangular inequality. This gives $\mathrm{tr}(AH^{-1}) \leq \frac{1}{2}(L/M)^2 q$, and then the claim. ■

Remark 4.4 (Asymptotic optimal tuning of M, K and N) We work under the assumptions of Proposition 4.3 and assume in addition that $\sigma(x) \leq \overline{\sigma} < \infty$. We are interested in minimizing

$$\mathcal{E} := \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \left(\theta_N^K \cdot u(X_i) - \frac{1}{K}\sum_{k=1}^K f(Y_i^{(k)})\right)^2\right],$$

which is the averaged quadratic error on the sample. Following the lines of [Gob16, Theorem 8.2.4], we get $\mathcal{E} = \mathcal{E}_1 + \mathcal{E}_2$, with

$$\mathcal{E}_1 = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \left(\theta_N^K \cdot u(X_i) - \psi(X_i)\right)^2\right] \text{ and } \mathcal{E}_2 = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \left(\psi(X_i) - \frac{1}{K}\sum_{k=1}^K f(Y_i^{(k)})\right)^2\right],$$

representing respectively the approximation error and the statistical error. We show in the same manner that $\mathcal{E}_1 \leq \frac{L^2}{M^2}$ and $\mathcal{E}_2 \leq \frac{\overline{\sigma}^2 M^d}{KN}$. To achieve a precision of order $\varepsilon > 0$ (and a quadratic error of order ε^2), we then take $\frac{L^2}{M^2} = \varepsilon^2$ and $\frac{\overline{\sigma}^2 M^d}{KN} = \varepsilon^2$. This leads to $M = c\varepsilon^{-1}$ and $KN = c'\varepsilon^{-(2+d)}$ for some constants $c, c' > 0$, for an overall computational cost of $O(\varepsilon^{-(3+d)})$. Taking the optimal choice $K^\star$ of Theorem 3.2 does not change the order of convergence but improves its multiplicative rate.

5 Numerical experiments

The present numerical section is organized as follows. We first present the different estimators to approximate $K^\star$ and describe the Monte-Carlo algorithm that is used for all the examples. Then,

we begin with a toy example that illustrates that $K^\star$ may be arbitrarily large. In this case, as the computational time $C_{Y|X}$ is equal to C_X, the theoretical multiplicative gain $r^\star \approx 1/2$, and we almost reach this bound in practice. The second example shows a case where the computational time $C_{Y|X}$ is much smaller than C_X, and therefore where the multiplicative gain may be larger. The last example is more practically oriented and deals with risk management concerns.

In this section, we compare the performances of the following estimations of $K^\star$ based on Theorem 3.2,

$$K_H^\star = \nu\left(\frac{\operatorname{tr}(BH^{-1})}{C_{Y|X}\operatorname{tr}(AH^{-1})}\right)$$

with the one suggested by Remark 3.3,

$$K_{\not H}^\star = \nu\left(\frac{\operatorname{tr}(B)}{C_{Y|X}\operatorname{tr}(A)}\right),$$

which is much simpler to estimate. Following (17) and (18), we approximate $K_H^\star$ or $K_{\not H}^\star$ by the four estimators

$$\begin{aligned}
\hat K_H^A &= \nu\left(\frac{\operatorname{tr}(\hat B_{\bar K}^{anti}\hat H^{-1})}{C_{Y|X}\operatorname{tr}((\hat A_{\bar K}^{anti}\hat H^{-1})_+)}\right); \quad & \hat K_{\not H}^A &= \nu\left(\frac{\operatorname{tr}(\hat B_{\bar K}^{anti})}{C_{Y|X}\operatorname{tr}((\hat A_{\bar K}^{anti})_+)}\right)\\
\hat K_H^\Gamma &= \nu\left(\frac{\operatorname{tr}(\hat B_{\bar K}^{anti}\hat H^{-1})}{C_{Y|X}\operatorname{tr}(\Gamma_N^{2\bar K}\hat H^{-1})}\right); & \hat K_{\not H}^\Gamma &= \nu\left(\frac{\operatorname{tr}(\hat B_{\bar K}^{anti})}{C_{Y|X}\operatorname{tr}(\Gamma_N^{2\bar K})}\right),
\end{aligned} \tag{20}$$

where $\hat H = \nabla^2 v_N^K(\theta_N^K)$. Note that in the linear regression framework, we simply have $\hat H = \frac{1}{N}\sum_{i=1}^N u(X_i)u(X_i)^T$. When $q = 1$, matrices are scalar, and we take $\hat K_H^A = \hat K_{\not H}^A = \nu\left(\frac{\hat B_{\bar K}^{anti}}{C_{Y|X}|\hat A_{\bar K}^{anti}|}\right)$ and $\hat K_H^\Gamma = \hat K_{\not H}^\Gamma = \nu\left(\frac{\hat B_{\bar K}^{anti}}{C_{Y|X}\Gamma_N^{2\bar K}}\right)$. Since $\Gamma^{2\bar K} \geq A$ is a semi-definite positive matrix, $\hat K_H^\Gamma$ and $\hat K_{\not H}^\Gamma$ will slightly underestimate $K_H^\star$ and $K_{\not H}^\star$. However, as we will see they have a much smaller variance and give a nearly optimal computational gain.

For each example, we run our algorithm $20,000$ times to approximate $\mathbb{E}[v_N^K(\theta^\star)] - \mathbb{E}[v_N^K(\theta_N^K)]$, which is (under uniform integrability assumption) an estimator of $\mathbb{E}[v^\infty(\theta_N^K)] - \mathbb{E}[v^\infty(\theta^\star)]$ by Proposition 2.6. Namely, we calculate for $J = 20,000$ the estimator $\frac{1}{J}\sum_{j=1}^J v_{N,j}^K(\theta^\star) - v_{N,j}^K(\theta_{N,j}^K)$, where $(v_{N,j}^K)_{1\leq j\leq J}$ are iid samples of (2) and, for each j, $\theta_{N,j}^K$ is the minimum of $v_{N,j}^K$. This minimum is computed explicitly for linear regression and can be approximated by a gradient descent otherwise. The value of $\theta^\star$ is approximated by minimizing $v_N^1(\theta)$ for $N = 100,000$. In comparison, the values of (N, K) to sample $v_{N,j}^K$ and then $\theta_{N,j}^K$ are such that $N\frac{1+KC_{Y|X}}{1+C_{Y|X}} \approx 5000$. This means that the simulation computational cost is fixed at 5000 times the cost of simulating X, across the different values of K.

Using the $20,000$ runs, we compute as many samples of the estimators $\hat K_H^A$, $\hat K_{\not H}^A$, $\hat K_H^\Gamma$ and $\hat K_{\not H}^\Gamma$ and plot their empirical distributions on the window $0, \ldots, 110$. To do so, we use $N = 50000$ samples in the formulas (20) and indicate the value of $\bar K$ in the captions of each related figure. Separately, we also calculate on $20,000$ runs the multiplicative computational gain r^K defined

by Remark 3.5 by using the estimator

$$\hat{r}^K = \frac{\frac{1}{J}\sum_{j=1}^J v^K_{N'(N,K),j}(\theta^\star) - v^K_{N'(N,K),j}(\theta^K_{N'(N,K),j})}{\frac{1}{J}\sum_{j=1}^J v^1_{N,j}(\theta^\star) - v^1_{N,j}(\theta^1_{N,j})} \tag{21}$$

with $N = 5000$ and $N'(N,K) = \left\lfloor N\frac{1+C_{Y|X}}{1+KC_{Y|X}} \right\rfloor$. In fact, assuming the uniform integrability of the family $N'(N,K)(v^K_{N'(N,K)}(\theta^\star) - v^K_{N'(N,K)}(\theta^K_{N'(N,K)}))$, we get by Proposition 2.6 that $\mathbb{E}[v^K_{N'(N,K)}(\theta^\star)] - \mathbb{E}[v^K_{N'(N,K)}(\theta^K_{N'(N,K)})] \sim_{N\to\infty} \frac{\mathrm{tr}(\Gamma^K H^{-1})}{N'(N,K)}$ exactly as in the proof of Theorem 3.2. By (14), this gives

$$\frac{\mathbb{E}[v^K_{N'(N,K)}(\theta^\star)] - \mathbb{E}[v^K_{N'(N,K)}(\theta^K_{N'(N,K)})]}{\mathbb{E}[v^1_N(\theta^\star)] - \mathbb{E}[v^1_N(\theta^1_N)]} \to_{N\to\infty} \frac{1+KC_{Y|X}}{1+C_{Y|X}} \frac{\mathrm{tr}(\Gamma^K H^{-1})}{\mathrm{tr}(\Gamma^1 H^{-1})} = r^K.$$

5.1 Toy example in a Gaussian framework

Consider a one dimensional toy example in a Gaussian framework. Let (X,Y) be a Gaussian vector such that X and Y are two standard normal random variables with covariance $\rho \in [-1,1]$. Let f be the square function, $f : x \in \mathbb{R} \longmapsto x^2$. We consider a constant approximation meaning that the function φ is defined by $\varphi(\theta, x) = \theta$, for $\theta \in \mathbb{R}$ and $x \in \mathbb{R}$. Easy computations lead to explicit formulas

$$\mathbb{E}[f(Y)|X] = \rho^2 X^2 + (1-\rho^2)$$
$$\theta^\star = 1; \quad A = \Gamma_\infty = 2\rho^4; \quad B = 2(1-\rho^4).$$

In this case, the value of $K^\star$ is given by

$$K^\star = \nu\left(\frac{1-\rho^4}{\rho^4}\right).$$

This very simple example shows that the optimal number of inner samples $K^\star$ can vary from 1 to arbitrary large values. As the parameter θ is one dimensional, the Hessian matrix H is scalar valued and therefore $K^\star = K^\star_{\not H}$. Thus, the four estimators reduce to two.

For our numerical experiments on this toy example, we fix $\rho = 0.1$, in which case the theoretical value of $K^\star$ is 100. Figure 1 clearly shows that the gain r_K is almost constant for any $K \geq 20$. Even though from a theoretical point of view $K^\star = 100$, any values of K larger than 20 are equally good in practice. Note that from Corollary 3.4, $r^\star \geq \frac{1}{2}$; this lower bound is almost attained by r_K for $K \geq 20$. Figure 2 shows a comparison of the distributions of the two estimators $\hat{K}^A_H$, and $\hat{K}^\Gamma_H$. The estimator $\hat{K}^A_H$ has a very large standard deviation (equal to 79) and may take values as large as 4275, whereas the estimator $\hat{K}^\Gamma_H$ is much more concentrated and only takes two values 8 and 9. These are typical behaviours of these estimators: as discussed at the end of Section 3.2, the estimated matrix $\hat{A}^{anti}_{\bar{K}}$ may have negative eigenvalues coming from a too large variance in the Monte Carlo computation and leading to non reliable estimations of A. On the contrary, $\hat{K}^\Gamma_H$ uses $\Gamma^{2\bar{K}}_N$ as an approximation of A from above leading to a conservative

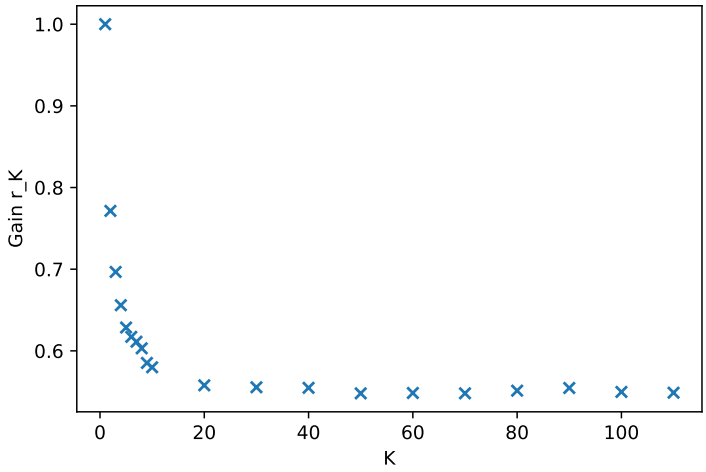
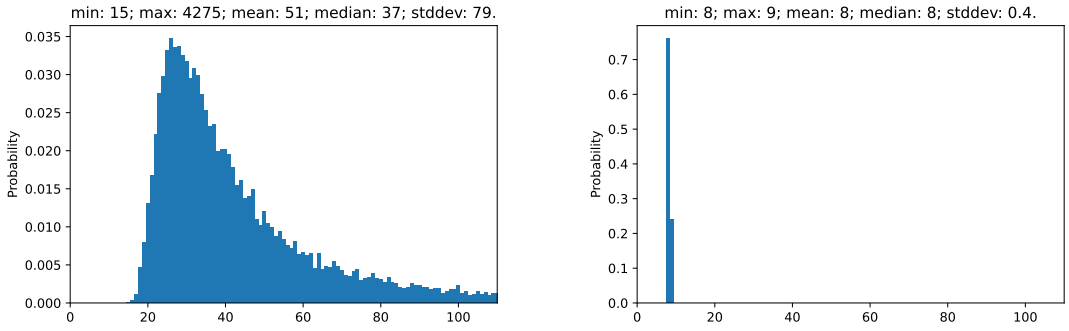


Figure 1: Computational multiplicative gain as a function of K estimated with (21) for the Gaussian toy example ($\rho = 0.1$) with regression on the constant function.



(a) Distribution of $\hat{K}_H^A$ with $\bar{K} = 4$

(b) Distribution of $\hat{K}_H^\Gamma$ with $\bar{K} = 32$

Figure 2: Comparison of the 2 estimators $\hat{K}_H^A$ and $\hat{K}_H^\Gamma$ for the Gaussian toy example.

estimation of $K^\star$ from below with far less variance. The gains r_K reported in Figure 1 for $K = 8$ or $K = 9$ are very close to the best possible gains. In the example, we can conclude that $\hat{K}_H^\Gamma$ with $\bar{K} = 32$ is much better than $\hat{K}_H^A$, since it has small fluctuations and gives a nearly optimal computation gain.

5.2 A SDE conditioned on an intermediate date

We consider the following SDE

$$dX_t = \cos(X_t)dW_t; \quad X_0 = 0$$

where W is a real valued Brownian motion. We aim at estimating $\mathbb{E}[X_{t_2}^2|X_{t_1}]$ with $t_2 = 10$ and $t_1 = 9$. This amounts to take $Y = X_{t_2}$, $f(x) = x^2$ and $X = X_{t_1}$ in (1). The SDE is discretized using the Euler scheme with 200 time-steps, hence inner simulations are cheaper than outer simulations; their relative cost is $C_{Y|X} = \frac{1}{9}$. We consider two different settings for the family of functions $(\varphi(\theta;\cdot))_\theta$: a polynomial with degree 3 (see Figures 3 and 5) and a piecewise constant approximation (see Figures 4 and 6). In both settings, the parameter θ is multi-dimensional, so the Hessian matrix is a true matrix and the estimators with and without H are actually different.

We build the piecewise constant approximation on $\mathbb{R}$ in the following way. First, we center the samples of X around their mean and rescale them by their standard deviation, then we apply the function $x \mapsto \frac{1}{\sqrt{\pi}} \int_{-\infty}^{x} \mathrm{e}^{-t^2}\, dt$ to map $\mathbb{R}$ into $(0, 1)$. Finally, we split the interval $[0, 1]$ into M regular sub-intervals.

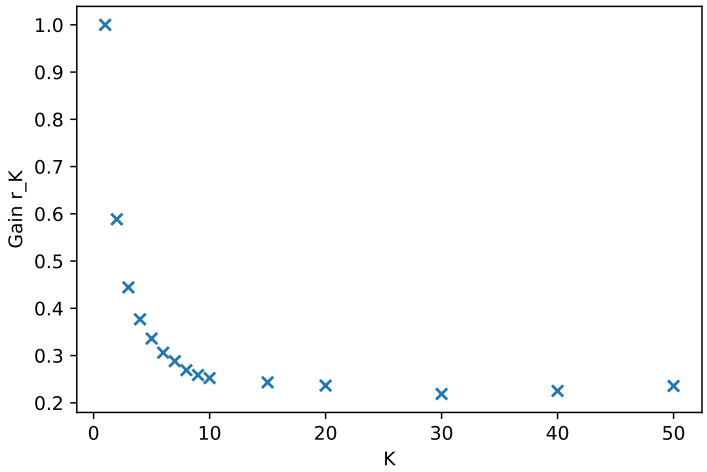


Figure 3: Computational multiplicative gain as a function of K estimated with (21) for the SDE example with a polynomial regression of order 3 and $t_1 = 9$.

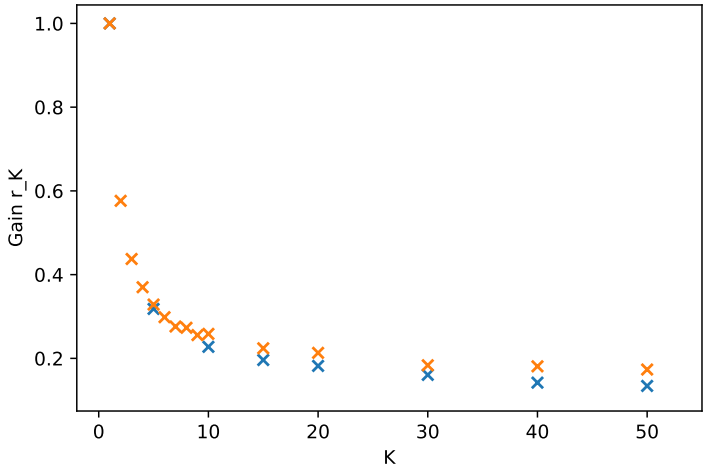
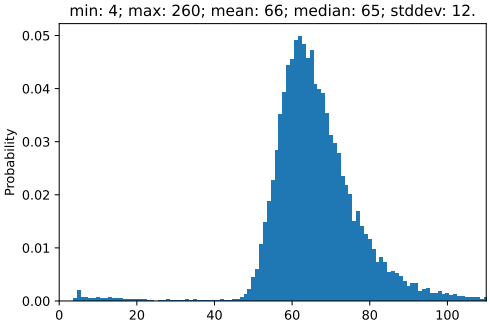


Figure 4: Computational multiplicative gain as a function of K estimated with (21) for the SDE example with a local regression with $t_1 = 9$ and $M = 50$ (orange crosses) or $M = 100$ (blue crosses).

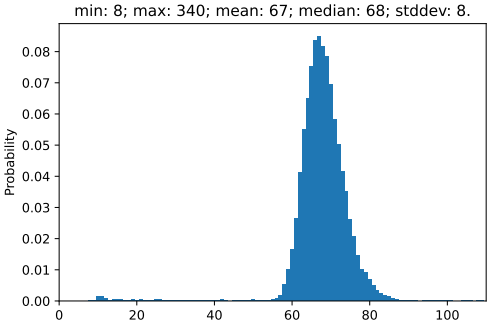
Figures 3 and 4 exhibit very similar gain profiles for r^K. The polynomial regression of order 3 works quite well on this example, which is related to the choice of $f(x) = x^2$. Figure 4 compares the multiplicative gain for two different local regressions: one with $M = 50$ intervals and the other one with $M = 100$. As expected, increasing the number of cells improves the approximation and thus the gain, according to Corollary 3.4. The multiplicative gain obtained is around 0.2 for $M = 50$ and 0.15 for $M = 100$, to be compared with the best gain $\frac{1}{10}$ given by Corollary 3.4.

Figure 5 shows a comparison of the four estimators defined in (20) in the polynomial regression setting. The estimators $\hat{K}^A_H$ (resp. $\hat{K}^\Gamma_H$) and $\hat{K}^A_{\not H}$ (resp. $\hat{K}^\Gamma_{\not H}$) have very similar distributions. Simplifying H in the ratio $\frac{\mathrm{tr}(BH^{-1})}{C_{Y|X}\,\mathrm{tr}(AH^{-1})}$ even tends to slightly reduce the variance of the estimator without significantly changing its mean. The estimators $\hat{K}^A_H$ and $\hat{K}^A_{\not H}$ based on the use of $\hat{A}^{anti}_K$ have larger variances and may return very extreme values (between 4 and 340). On the contrary, the estimators $\hat{K}^\Gamma_H$ and $\hat{K}^\Gamma_{\not H}$ have very small standard deviation and show a much more concentrated probability function than the estimators based on $\hat{A}^{anti}_K$. The use of Γ_{2K} as an approximation of A tends to produce smaller approximations of $K^\star$: their empirical means are shifted by approximately -20. However, the gain profiles of Figure 3 are almost flat for $K \geq 20$, hence this shift does not change the best gain attained by our method. As a conclusion, we recommend to use $\hat{K}^K_{\not H}$ to approximate $K^\star$.

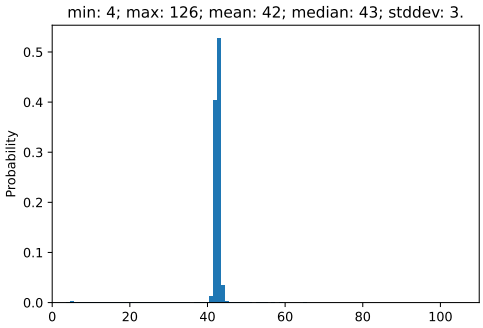
In Figure 6, we observe very similar behaviours for the piecewise constant approximation setting as the ones we described above for the polynomial regression framework. However, we note that the estimation of the matrix H is more difficult than in the previous polynomial framework, especially for the intervals with few data. This explains heuristically why the estimators



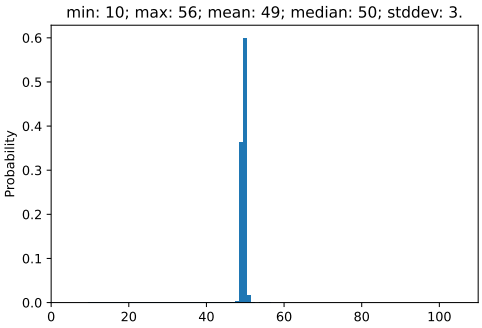
(a) Distribution of $\hat{K}_H^A$ with $\bar{K} = 4$



(b) Distribution of $\hat{K}_{\not{H}}^A$ with $\bar{K} = 4$



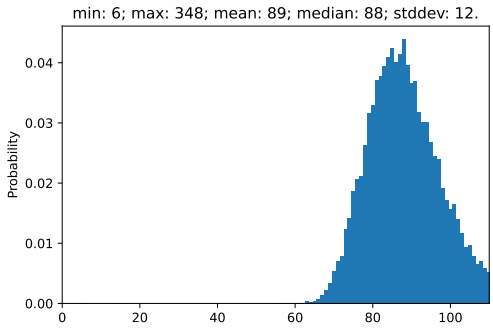
(c) Distribution of $\hat{K}_H^\Gamma$ with $\bar{K} = 32$



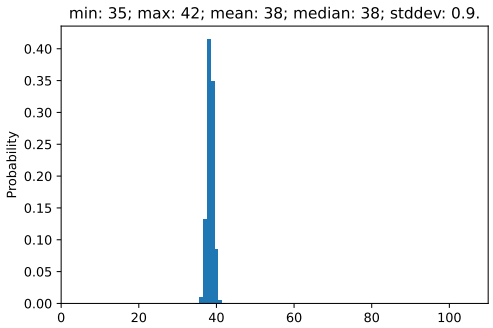
(d) Distribution of $\hat{K}_{\not{H}}^\Gamma$ with $\bar{K} = 32$

Figure 5: Comparison of the 4 estimators $\hat{K}_H^A$, $\hat{K}_{\not{H}}^A$, $\hat{K}_H^\Gamma$ and $\hat{K}_{\not{H}}^\Gamma$ for the SDE example with a polynomial regression with degree 3 and $t_1 = 9$.

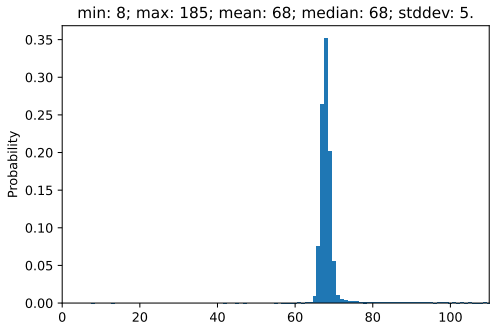
without H are less noisy.



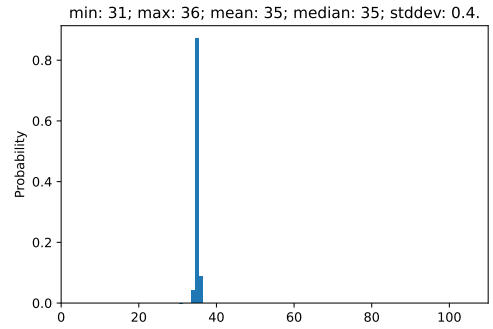
(a) Distribution of $\hat{K}_H^A$ with $\bar{K} = 4$ and $t_1 = 9$



(b) Distribution of $\hat{K}_{\not{H}}^A$ with $\bar{K} = 4$



(c) Distribution of $\hat{K}_H^\Gamma$ with $\bar{K} = 32$



(d) Distribution of $\hat{K}_{\not{H}}^\Gamma$ with $\bar{K} = 32$

Figure 6: Comparison of the 4 estimators $\hat{K}_H^A$, $\hat{K}_{\not{H}}^A$, $\hat{K}_H^\Gamma$ and $\hat{K}_{\not{H}}^\Gamma$ for the SDE example with a local regression with $M = 50$ and $t_1 = 9$.

5.3 An introductory example to risk management in insurance

In the introduction of the present paper, we have indicated the relevance of computing conditional expectations for risk management. Here, we take back this example from [ACIA21] that mimics the methodology of the standard formula to calculate the Solvency Capital Requirement, in the sense that it applies a shock to the underlying asset (we refer to [ACIA21] for further details). We now describe this example and consider an asset whose price follows the Black-Scholes model:

$$S_t = S_0 \exp\left(\sigma W_t - \frac{\sigma^2}{2} t\right), \ t \geq 0,$$

where $S_0, \sigma > 0$ and W is a standard Brownian motion. In practice, insurance companies are interested in computing the losses of their portfolio when a shock occurs in the economy. Here,

for simplicity, we will consider a butterfly option as a crude approximation of a true insurance portfolio. Thus, we are interested in a butterfly option with strikes $0 < \mathbf{K}_1 < \mathbf{K}_2$ that pays

$$\psi(S_T) = (S_T - \mathbf{K}_1)^+ + (S_T - \mathbf{K}_2)^+ - 2\left(S_T - \frac{\mathbf{K}_1 + \mathbf{K}_2}{2}\right)^+$$

at time $T > 0$. The price of such an option at time $t \in [0, T]$ is given by $\mathbb{E}[\psi(S_T)|S_t]$. Solvency II in its standard model assumes that there is a shock on the asset at time $t \in (0, T)$ that multiplies its value by $1 + s$, $s \in (-1, +\infty)$. Then, in the Black-Scholes model, we have to compute the following quantity

$$\mathcal{L} = \mathbb{E}[\max(\mathbb{E}[\psi(S_T) - \psi((1+s)S_T)|S_t], 0)], \tag{22}$$

which can be seen as the expected loss generated by the shock. In this particular example, $\mathbb{E}[\psi(S_T) - \psi((1+s)S_T)|S_t]$ has an explicit form by using the Black-Scholes formula, that we can use as a benchmark to compute the mean square error of our estimator of (22). Note that since $x \mapsto \max(x, 0)$ is 1-Lipschitz, we have

$$\left|\mathbb{E}[\max(\varphi(\theta, S_t), 0)] - \mathbb{E}[\max(\mathbb{E}[\psi(S_T) - \psi((1+s)S_T)|S_t], 0)]\right| \\ \leq \sqrt{\mathbb{E}\left[(\mathbb{E}[\psi(S_T) - \psi((1+s)S_T)|S_t] - \varphi(\theta, S_t))^2\right]}.$$

The estimator θ_N^K minimizes empirically the right hand side, which gives at the same time an upper bound on the approximation error of the expected loss.

Here, we have used our approach to compute $\mathbb{E}[\psi(S_T) - \psi((1+s)S_T)|S_t]$. Thus, we have $X = S_t$, $Y = X \exp\left(\sigma(W_T - W_t) - \frac{\sigma^2}{2}(T-t)\right)$ and $C_X = C_{Y|X}$ (the simulation of X and of Y given X both require to sample one normal random variable)[2]. We have taken $s = 0.2$, $t = 1$ and $T = 2$ and consider the local regression with $M = 50$, using the same transformation as the one used for the SDE example presented in Subsection 5.2.

Figure 7 plots the multiplicative computational gain as a function of K, while Figure 8 shows the empirical distribution of the different estimators (20). We see from Figure 8 that most of the computational gain is realized for $K \geq 5$. Similarly to the previous example, Figure 8 shows that the estimator $\hat{K}^{\Gamma}_{\mathcal{H}}$ is a good one to choose K: it has few fluctuations and avoid the issue of estimating $\hat{H}$.

We now focus on the numerical approximation of (22). Figure 9 illustrates the mean square error on the estimated expected loss as a function of (N, K) for a given computational budget, as explained in the introduction of Section 5. More precisely, from the sample $(\theta^K_{N'(N,K),j}, 1 \leq j \leq J)$, we compute:

$$\frac{1}{J}\sum_{j=1}^{J}\left(\mathbb{E}[\max(\varphi(\theta^K_{N'(N,K),j}, S_t), 0)|\theta^K_{N'(N,K),j}] - \mathcal{L}\right)^2,$$

[2]Note that $C_X = C_{Y|X}$ is particular to the Black-Scholes model for which exact simulation is possible. For a more general diffusion, one typically uses a discretization scheme to approximate it, like in Subsection 5.2. Then, we rather get $C_{Y|X} \approx \frac{T-t}{t} C_X$ and the computational gain may be important when $t \to T$.

and plot the different values. Here, we compute

$$\mathbb{E}[\max(\varphi(\theta^K_{N'(N,K),j}, S_t), 0)|\theta^K_{N'(N,K),j}] = \int_{\mathbb{R}} \max(\varphi(\theta^K_{N'(N,K),j}, S_0 e^{\sigma\sqrt{t}x - \sigma^2 t/2}), 0) \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$$

and $\mathcal{L}$ by numerical integration, using the Black-Scholes formula for $\mathcal{L}$. We find $\mathcal{L} \approx 3.077$. We first note from Figure 9 that in this example, as in all the other ones in Section 5, the choice $K = 1$ that is commonly used is suboptimal. Numerically, the optimal choice of K seems to be $K^\star = 8$ or $K^\star = 9$, which is in line with the estimators $\hat{K}^\Gamma_H$ and $\hat{K}^\Gamma_{\not H}$. However, any choice of K between 5 and 20 leads to an MSE that is close to the optimal one, which confirms that a precise estimation of $K^\star$ is not needed to take the benefit of the proposed method.

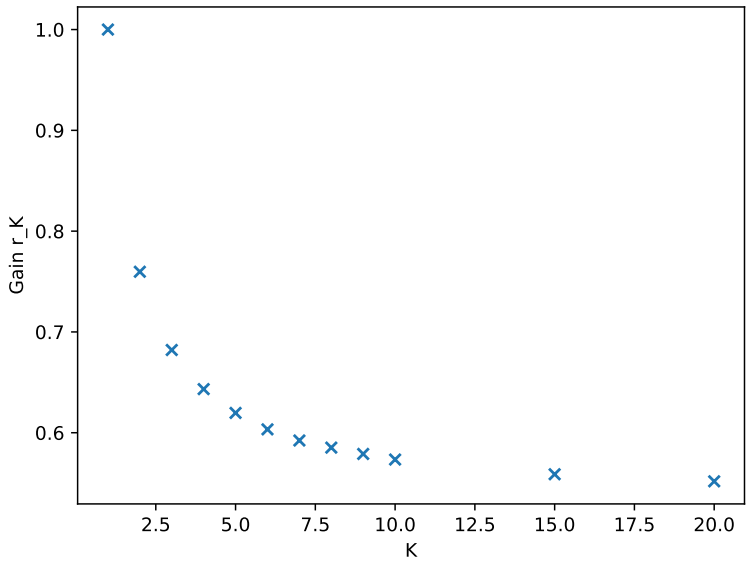
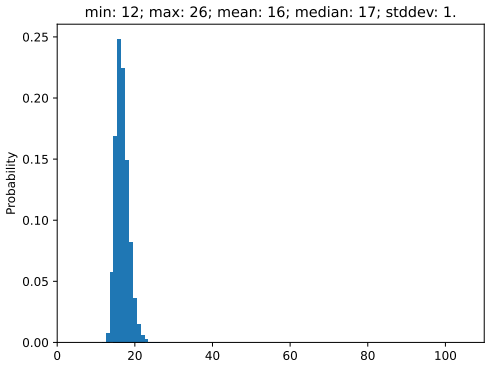
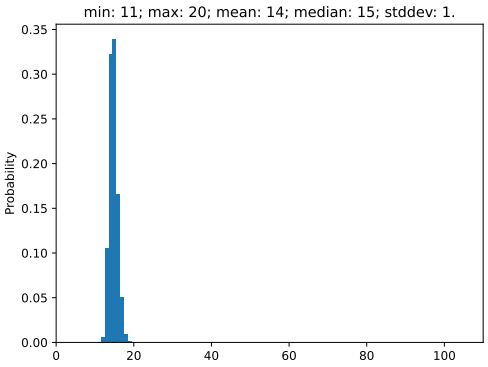


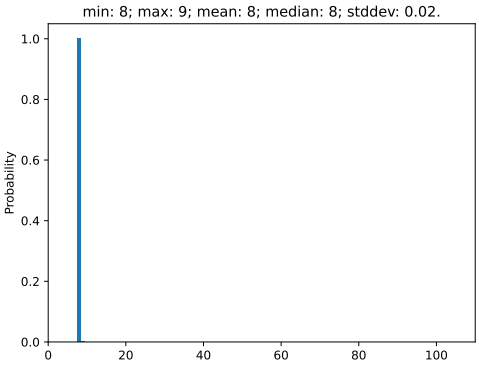
Figure 7: Computational multiplicative gain as a function of K estimated with (21) for the butterfly example with a local regression with $M = 50$.



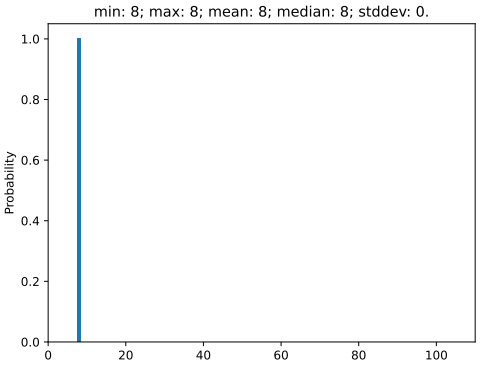
(a) Distribution of $\hat{K}_H^A$ with $\bar{K} = 4$



(b) Distribution of $\hat{K}_{\not{H}}^A$ with $\bar{K} = 4$



(c) Distribution of $\hat{K}_H^\Gamma$ with $\bar{K} = 32$



(d) Distribution of $\hat{K}_{\not{H}}^\Gamma$ with $\bar{K} = 32$

Figure 8: Comparison of the 4 estimators $\hat{K}_H^A$, $\hat{K}_{\not{H}}^A$, $\hat{K}_H^\Gamma$ and $\hat{K}_{\not{H}}^\Gamma$ for the butterfly example with a local regression with $M = 50$.

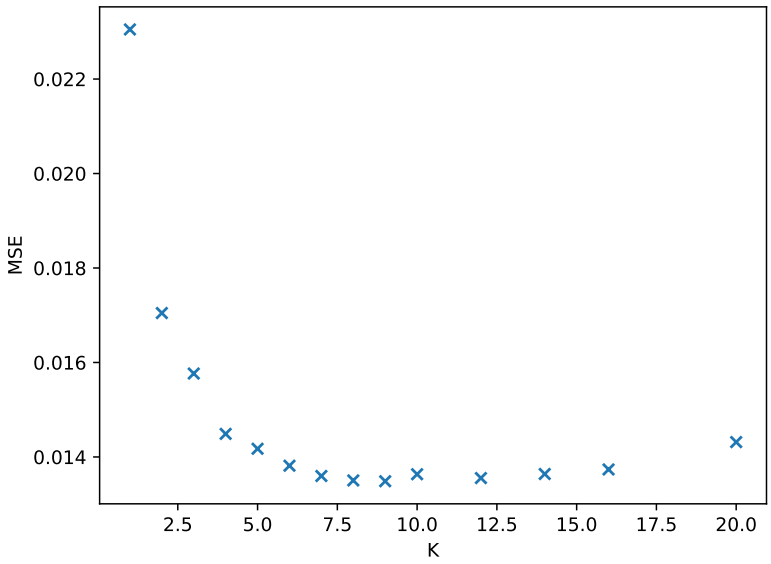


Figure 9: Computation of the mean square error as a function of K with a local regression with $M = 50$.

6 Conclusion

In this work, we have investigated how to balance the computational effort between inner and outer simulations when computing conditional expectations with least-square Monte Carlo. The computational gain can be significant when the computational cost $C_{Y|X}$ is small with respect to C_X, and when the family $(\varphi(\theta, X))_\theta$ well approximates the conditional expectation $\mathbb{E}[f(Y)|X]$.

We have proposed several estimators to approximate the optimal number of inner simulations in practice. Numerical simulations have shown that the estimators $\hat{K}^\Gamma_H$ and $\hat{K}^\Gamma_{\not H}$ have much smaller standard deviations. Although they provide smaller estimations of the optimal number of inner samples $K^\star$, they almost attain the best gain and should be used in practice in favour of those relying on $\hat{A}^{anti}$. When it comes to choosing between $\hat{K}^\Gamma_H$ and $\hat{K}^\Gamma_{\not H}$, one should keep in mind that H is a Hessian matrix, whose computation may be extremely costly and noisy. The effect of removing H in $\hat{K}^\Gamma$ is to reduce the noise, and in all our experiments, this estimator almost reaches the optimal gain. Then, as the best trade-off between accuracy and ease of computation, we suggest to use $K^\Gamma_{\not H}$.

A Technical results

We start by recalling the uniform law of large number, that can be found e.g. in [RS93, Lemma A1, p. 67].

Lemma A.1 *Let $X : \Omega \to \mathbb{R}^d$ be a random variable, $\psi : \mathbb{R}^q \times \mathbb{R}^d \to \mathbb{R}$ a measurable function and $C \subset \mathbb{R}^q$ a compact set. Let $(X_i)_{i\geq 1}$ be an iid sequence with the same distribution as X. If $C \ni \theta \mapsto \psi(\theta, X)$ is a.s. continuous and $\mathbb{E}[\sup_{\theta\in C} |\psi(\theta, X)|] < \infty$, then*

$$\sup_{\theta\in C} \left| \frac{1}{N}\sum_{i=1}^{N} \psi(\theta, X_i) - \mathbb{E}[\psi(\theta, X)] \right| \underset{N\to\infty}{\to} 0, \ a.s.$$

The next lemma solves the optimisation problem that arises to find the best estimator for a given computational budget.

Lemma A.2 *Let $a, b, \bar{c} > 0$. As $c \to +\infty$, we have*

$$\inf_{x,y\in\mathbb{N}^*: x+xy\bar{c}\leq c} \frac{a}{x} + \frac{b}{xy} \sim_{c\to\infty} \frac{1}{c}\left[a + b\bar{c} + \frac{b}{a\bar{c}}\left(\nu\left(\frac{b}{a\bar{c}}\right) + \nu\left(\frac{b}{a\bar{c}}\right)^{-1}\right)\right],$$

and the following solution is asymptotically optimal

$$x^\star(c) = \left\lfloor \frac{c}{1 + y^\star\bar{c}} \right\rfloor, \ y^\star = \nu\left(\frac{b}{a\bar{c}}\right), \tag{23}$$

in the sense that it satisfies $x^\star(c) + x^\star(c)y^\star\bar{c} \leq c$ and

$$\inf_{x,y\in\mathbb{N}^*: x+xy\bar{c}\leq c} \frac{a}{x} + \frac{b}{xy} \sim_{c\to\infty} \frac{a}{x^\star(c)} + \frac{b}{x^\star(c)y^\star}.$$

Proof. We consider the semi-discrete minimization problem $\inf_{x>0,y\in\mathbb{N}^*: x+xy\bar{c}\leq c} \frac{a}{x} + \frac{b}{xy}$. For each $y \in \mathbb{N}^*$, the optimal choice is to take $x = \frac{c}{1+y\bar{c}}$, and the infimum is given by

$$g(y) := \frac{a + b\bar{c}}{c} + \frac{b}{cy} + \frac{a\bar{c}}{c}y.$$

Let $y_0^\star = \sqrt{\frac{b}{a\bar{c}}}$. We check easily that g is decreasing on $(0, y_0^\star)$ and increasing on $(y_0^\star, +\infty)$. Therefore the minimum on $\mathbb{N}^*$ is reached by 1 if $y_0^\star \leq 1$, and by p or $p+1$ if $y_0^\star \in [p, p+1]$ for some $p \in \mathbb{N}^*$. We compare these two candidates and rewrite $g(y) = \frac{a}{c} + \frac{b\bar{c}}{c} + \frac{a\bar{c}}{c}\left(y + \frac{(y_0^\star)^2}{y}\right)$. Since

$$p + \frac{z^2}{p} \leq p + 1 + \frac{z^2}{p+1} \iff z^2 \leq p(p+1),$$

we get that p is optimal if $(y_0^\star)^2 \leq p(p+1)$ and $p+1$ is optimal if $(y_0^\star)^2 \geq p(p+1)$ (both are optimal for $(y_0^\star)^2 = p(p+1)$). Therefore, the infimum is reached by $y^\star = \nu\left(\frac{b}{a\bar{c}}\right)$ (see Definition 3.1), and we have

$$\inf_{x>0, y\in\mathbb{N}^*: x+xy\bar{c}=c} \frac{a}{x} + \frac{b}{xy} \sim_{c\to\infty} \frac{1}{c}\left[a + b\bar{c} + \frac{b}{a\bar{c}}\left(\nu\left(\frac{b}{a\bar{c}}\right) + \nu\left(\frac{b}{a\bar{c}}\right)^{-1}\right)\right].$$

Now, we simply notice that $\inf_{x>0,y\in\mathbb{N}^*:x+xy\bar{c}=c} \frac{a}{x} + \frac{b}{xy} \leq \inf_{x,y\in\mathbb{N}^*:x+xy\bar{c}=c} \frac{a}{x} + \frac{b}{xy}$ and that $\frac{a}{x^\star(c)} + \frac{b}{x^\star(c)y^\star} \sim_{c\to\infty} \frac{1}{c}\left[a + b\bar{c} + \frac{b}{a\bar{c}}\left(\nu\left(\frac{b}{a\bar{c}}\right) + \nu\left(\frac{b}{a\bar{c}}\right)^{-1}\right)\right]$. ■

The next lemma gives a sufficient condition to get some uniform integrability in the central limit theorem.

Lemma A.3 *Let $(Z_i)_{i\geq 1}$ be an iid sequence of random variables in $\mathbb{R}^d$ such that $\mathbb{E}[|Z_1|^{2+\eta}] < \infty$ for some $\eta > 0$. Let $\bar{Z}_N = \frac{1}{N}\sum_{i=1}^N Z_i$. Then, the sequence $(N|\bar{Z}_N - \mathbb{E}[Z_1]|^2)_{N\geq 1}$ is uniformly integrable.*

Proof. This is a direct application of [GHJvW23, Proposition 2.4] that gives

$$\mathbb{E}[(N|\bar{Z}_N - \mathbb{E}[Z_1]|^2)^{1+\eta/2}] = \mathbb{E}[(\sqrt{N}|\bar{Z}_N - \mathbb{E}[Z_1]|)^{2+\eta}] \leq C_{2+\eta}\mathbb{E}[|Z_1 - \mathbb{E}[Z_1]|^{2+\eta}],$$

for some constant $C_{2+\eta} < \infty$. ■

References

[ACIA21] Aurélien Alfonsi, Adel Cherchali, and Jose Arturo Infante Acevedo. Multilevel Monte-Carlo for computing the SCR with the standard formula and other stress tests. *Insurance Math. Econom.*, 100:234–260, 2021.

[BBR09] Daniel Bauer, Daniela Bergmann, and Andreas Reuss. Solvency II and nested simulations – a least-squares Monte Carlo approach. *Preprint Universität Ulm*, 2009.

[BDM15] Mark Broadie, Yiping Du, and Ciamac C. Moallemi. Risk estimation via regression. *Oper. Res.*, 63(5):1077–1097, 2015.

[BKS10] Denis Belomestny, Anastasia Kolodko, and John Schoenmakers. Regression methods for stochastic control problems and their convergence analysis. *SIAM J. Control Optim.*, 48(5):3562–3588, 2009/10.

[BRS12] Daniel Bauer, Andreas Reuss, and Daniela Singer. On the calculation of the solvency capital requirement based on nested simulations. *ASTIN Bulletin*, 42(2):453–499, 2012.

[BT04] Bruno Bouchard and Nizar Touzi. Discrete-time approximation and Monte-Carlo simulation of backward stochastic differential equations. *Stochastic Process. Appl.*, 111(2):175–206, 2004.

[GHJvW23] Philipp Grohs, Fabian Hornung, Arnulf Jentzen, and Philippe von Wurstemberger. A Proof that Artificial Neural Networks Overcome the Curse of Dimensionality in the Numerical Approximation of Black-Scholes Partial Differential Equations. *Mem. Amer. Math. Soc.*, 284(1410):1–106, 2023.

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

[GKKW02] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[GLW05] Emmanuel Gobet, Jean-Philippe Lemor, and Xavier Warin. A regression-based Monte Carlo method to solve backward stochastic differential equations. *Ann. Appl. Probab.*, 15(3):2172–2202, 2005.

[Gob16] Emmanuel Gobet. *Monte-Carlo methods and stochastic processes*. CRC Press, Boca Raton, FL, 2016. From linear to non-linear.

[KNK18] Anne-Sophie Krah, Zoran Nikolić, and Ralf Korn. A least-squares Monte Carlo framework in proxy modeling of life insurance companies. *Risks*, 6(2), 2018.

[LS15] Francis A. Longstaff and Eduardo S. Schwartz. Valuing American Options by Simulation: A Simple Least-Squares Approach. *The Review of Financial Studies*, 14(1):113–147, 06 2015.

[RS93] Reuven Y. Rubinstein and Alexander Shapiro. *Discrete event systems*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester, 1993.