



HAL
open science

Preliminary control variates to improve empirical regression methods

Tarik Benzineb, Emmanuel Gobet

► **To cite this version:**

Tarik Benzineb, Emmanuel Gobet. Preliminary control variates to improve empirical regression methods. Monte Carlo Methods and Applications, 2013, 19 (4), pp.331–354. hal-00802394

HAL Id: hal-00802394

<https://hal.science/hal-00802394>

Submitted on 19 Mar 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Preliminary control variates to improve empirical regression methods

Tarik Ben Zineb and Emmanuel Gobet

Abstract. We design a variance reduction method to reduce the estimation error in regression problems. It is based on an appropriate use of other known regression functions. Theoretical estimates are supporting this improvement and numerical experiments are illustrating the efficiency of the method.

Keywords. Empirical regression, variance reduction, distribution-free theory.

2010 Mathematics Subject Classification. 62G08, 65Cxx.

1 Introduction

Empirical regression methods to compute $\mathbb{E}(H|X = x)$ are widely used in Monte-Carlo algorithms to solve optimal stopping problems [4, 6, 12], Backward Stochastic Differential Equations [7, 8, 11] and various stochastic control problems [1, 2]: these algorithms are often referred to as Least Squares Monte-Carlo algorithms. However in some situations, the number of simulations is constrained to be relatively small (due to restriction on the computational time, see [3]). That case may cause significant inaccuracy in the least squares regression method, because the estimation error (also called statistical error) can be dominant, in particular if the conditional variance $\mathbb{V}\text{ar}(H|X)$ is large (see [9, Chapter 11] for details, or Theorem 2.1 below). The purpose of this work is to design a flexible method to significantly reduce it.

Depending on the use of empirical regression methods, the framework can be quite different. In the statistics field with applications to inference, data are rather given to the experimenter and in general, he can not rely on extra information about their distribution. In the probability field related to Monte-Carlo algorithms, the situation is different: the experimenter simulates data with possibly additional information. Here we consider this second framework by assuming the knowledge

This work has been done while the first author was preparing a CIFRE PhD Thesis at AXA and Ecole Polytechnique. The first author research is supported by AXA Fund Research and ANRT. The second author research is part of the Chair *Financial Risks* of the *Risk Foundation*, the Chair *Derivatives of the Future* and the Chair *Finance and Sustainable Development*.

of explicit regression functions that are going to be used to optimally reduce the estimation error of the empirical regression methods (see Theorems 2.2 and 2.3 below). These explicit regression functions are going to serve as control variates (called Preliminary Control Variates or PCV in short), which result into a two-stages algorithm presented in Section 2.3: firstly a L_2 -projection of the response H using the PCV, secondly a linear least squares regression applied to the residual. Our aim here is to prove numerically as well as mathematically how, in a context of few simulations, it reduces the estimation error and hence achieves the optimal bound (approximation error).

This paper is structured as follows. In the next section, we review an existing result of the error convergence of the standard linear least squares regression estimate; then we introduce the PCV method and state our main result about the global error estimates. The proofs are done in Section 3, where we recall the necessary tools of the Vapnik-Chervonenkis theory. Section 4 gathers numerical tests (in dimension 1 and 2 for X) showing the efficiency of the method. In some cases, the same accuracy is obtained using 50 times fewer simulations. We refer the reader to [3] for additional experiments.

2 Statement of the problem and main results

2.1 Setting

Our goal is to approximate

$$m(x) = \mathbb{E}[H|X = x] \quad \text{with} \quad H = h(Z),$$

where Z and X are two random variables taking values respectively in \mathbb{R}^{d_z} and \mathbb{R}^{d_x} ($1 \leq d_z, d_x < +\infty$) and $h : \mathbb{R}^{d_z} \rightarrow \mathbb{R}$ is a known function. To evaluate $m(\cdot)$, we make use of a data sample $D_N := (Z_i, X_i)_{1 \leq i \leq N}$ which consists of i.i.d simulations of (Z, X) . We set $H_i = h(Z_i)$.

2.2 Standard linear least squares regression method

Our presentation is inspired from [9].

- Let μ_N be the empirical measure associated to the sample D_N :

$$\mu_N(dz, dx) := \frac{1}{N} \sum_{i=1}^N \delta_{(Z_i, X_i)}(dz, dx).$$

- The L_2 -norm of a function g , measured with respect to μ_N , is denoted by $\|g\|_{L_2(\mu_N)}$:

$$\|g\|_{L_2(\mu_N)}^2 := \frac{1}{N} \sum_{i=1}^N g^2(Z_i, X_i).$$

- Similarly, by denoting μ the law of (Z, X) , we define $\|g\|_{L_2(\mu)}$ by

$$\|g\|_{L_2(\mu)}^2 := \int g^2(z, x) \mu(dz, dx).$$

- The unknown regression function $m(\cdot) : \mathbb{R}^{d_x} \mapsto \mathbb{R}$ is approximated within the linear vector space

$$\mathcal{F}_N = \text{Span}(\Phi_k : 1 \leq k \leq K_{\mathcal{F}}),$$

with $K_{\mathcal{F}} \in \mathbb{N}^*$ and where $\Phi_k : \mathbb{R}^{d_x} \mapsto \mathbb{R}$ may depend on (X_1, \dots, X_N) (see for instance the examples of data-driven basis functions in [4]).

The standard linear least squares regression method approximates¹ m by $m_N = \sum_{k=1}^{K_{\mathcal{F}}} \tilde{\gamma}_k \Phi_k$ where

$$(\tilde{\gamma}_k)_{1 \leq k \leq K_{\mathcal{F}}} = \underset{(\gamma_k)_k}{\text{arginf}} \frac{1}{N} \sum_{i=1}^N \left| H_i - \sum_{k=1}^{K_{\mathcal{F}}} \gamma_k \Phi_k(X_i) \right|^2.$$

The following result (see [9, Theorem 11.1]) provides a standard control on the convergence rate of m_N . For more recent results, see [10].

Theorem 2.1. Assume $\Sigma^2 = \sup_{x \in \mathbb{R}^{d_x}} \text{Var}(H|X = x) < +\infty$. Then

$$\mathbb{E} \left[\|m_N - m\|_{L_2(\mu_N)}^2 \right] \leq \Sigma^2 \frac{K_{\mathcal{F}}}{N} + \mathbb{E} \left[\inf_{f \in \mathcal{F}_N} \|f - m\|_{L_2(\mu_N)}^2 \right].$$

The above global error reads as the usual bias-variance decomposition, consisting of a sum of two terms.

- The first term is the *estimation error*: it is due to the finite number of simulations.
- The second term is the *approximation error*: it measures how well the regression function m can be approximated by functions of \mathcal{F}_N .

¹ The minimizing function m_N is unique in $L_2(\mu_N)$ but there may be multiple minimizing coefficients: in that case, we choose for $\tilde{\gamma}$ the SVD-optimal one, corresponding to that with minimal norm. This choice does not affect the empirical regression function m_N , see [8] for details.

- The larger $K_{\mathcal{F}}$, the smaller the approximation error but the larger the estimation error: hence, $K_{\mathcal{F}}$ and N have to be tuned optimally to achieve optimal convergence rates [13, 15].

Numerical experiments in [3] show that the upper bound in Theorem 2.1 is quite tight. But, in a context of few simulations or when the variance $\sup_{x \in \mathbb{R}^{d_x}} \text{Var}(H|X = x)$ is large, the estimation error is dominant, making the global error be far from achieving the approximation error. Hence in such a context, designing a new regression method which accelerates the convergence of the global error to the approximation error becomes an essential concern: this is the subject of the next subsection where we present the PCV method and the related results.

2.3 PCV least squares regression method

Heuristics

Suppose that the available additional information is the knowledge of K_{pcv} regression functions $x \mapsto \mathbb{E}[P_k(Z)|X = x]$ for some known functions $P_k : \mathbb{R}^{d_z} \mapsto \mathbb{R}$ ($1 \leq K_{\text{pcv}} < +\infty$). There is no loss of generality² to assume that they are conditionally centered:

$$\forall 1 \leq k \leq K_{\text{pcv}} : \mathbb{E}[P_k(Z)|X] = 0.$$

Based on this extra information and on the same sample data D_N , we wish to reduce the estimation error term $\Sigma^2 \frac{K_{\mathcal{F}}}{N}$ in the previous theorem, that is to improve the factor $\sup_{x \in \mathbb{R}^{d_x}} \text{Var}(H|X = x)$.

We expect such an improvement by replacing H_i by $H_i - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k^* P_k(Z_i)$ where $(\alpha_k^*)_{1 \leq k \leq K_{\text{pcv}}}$ are the optimal coefficients

$$(\alpha_k^*)_{1 \leq k \leq K_{\text{pcv}}} = \underset{\alpha \in \mathcal{A}}{\text{arginf}} \mathbb{E} \left[\left| H - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k P_k(Z) \right|^2 \right] \quad (1)$$

(the set \mathcal{A} of admissible parameters is defined later on). Indeed, as $\mathbb{E}[P_k(Z)|X] = 0$ for any k , observe that α^* is the minimizer

$$\underset{\alpha \in \mathcal{A}}{\text{arginf}} \text{Var} \left(H - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k P_k(Z) \right) = \underset{\alpha \in \mathcal{A}}{\text{arginf}} \mathbb{E} \left[\text{Var} \left(H - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k P_k(Z) | X \right) \right], \quad (2)$$

² Indeed, we can rewrite the regression problem using extended variables $(\widehat{Z}, \widehat{X}, \widehat{H}, \widehat{m})$ where $\widehat{Z} := (X, Z)$, $\widehat{X} := X$, $\widehat{H} := h(Z)$, $\widehat{m}(\widehat{X}) := \mathbb{E}(\widehat{H}|\widehat{X}) = m(X)$, $\widehat{P}_k(\widehat{Z}) := P_k(Z) - \mathbb{E}[P_k(Z)|X]$ that satisfies $\mathbb{E}(\widehat{P}_k(\widehat{Z})|\widehat{X}) = 0$. See also further examples.

which might be close to minimizing $\sup_{x \in \mathbb{R}^{d_x}} \mathbb{V}\text{ar}(H - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k P_k(Z) | X = x)$ over α . Actually, to achieve a fully implementable scheme, the minimization w.r.t. the $L_2(\mu)$ -norm in (1) is going to be replaced by a minimization over the sample data D_N .

Examples of known regression functions

Our examples are inspired by applications related to a given stochastic process $(Y_s)_{0 \leq s \leq T}$, that is for computing $\mathbb{E}[h(Y_T) | Y_t]$.

We first consider *control variates with local support*.

- a) Consider a standard Brownian motion W ; if h were well approximated by piecewise constant functions, one might take advantage of the control variates

$$P_k(Z) := P_k(W_t, W_T) := p_{x_k, \Delta}^0(W_T) - \mathbb{E}[p_{x_k, \Delta}^0(W_T) | W_t],$$

with

$$p_{x_k, \Delta}^0 := \mathbf{1}_{]x_k - \Delta, x_k + \Delta]}$$

and $x_k := -2\sqrt{T} + (k - \frac{1}{2})\frac{4\sqrt{T}}{K_{\text{pcv}}}$, $\Delta := \frac{2\sqrt{T}}{K_{\text{pcv}}}$, so that the full support of $(p_{x_k, \Delta}^0)_k$ is $] -2\sqrt{T}, 2\sqrt{T}]$. Besides, one has $\mathbb{E}[p_{x_k, \Delta}^0(W_T) | W_t = x] = \mathcal{N}\left(\frac{x_k + \Delta - x}{\sqrt{T-t}}\right) - \mathcal{N}\left(\frac{x_k - \Delta - x}{\sqrt{T-t}}\right)$ and $X = W_t$, $Z = (W_t, W_T)$.

- b) Now suppose that h could be well approximated by hat functions. Define

$$p_{x_k, \Delta}^1(x) := (1 - |\frac{x - x_k}{\Delta}|)_+$$

where $x_k := -2\sqrt{T} + k\Delta$, $\Delta := \frac{4\sqrt{T}}{K_{\text{pcv}}+1}$, and set

$$P_k(Z) := P_k(W_t, W_T) := p_{x_k, \Delta}^1(W_T) - \mathbb{E}[p_{x_k, \Delta}^1(W_T) | W_t],$$

where

$$\begin{aligned} \mathbb{E}[p_{x_k, \Delta}^1(W_T) | W_t = x] &= \frac{x - x_{k-1}}{\Delta} \left[\mathcal{N}\left(\frac{x_k - x}{\sqrt{T-t}}\right) - \mathcal{N}\left(\frac{x_{k-1} - x}{\sqrt{T-t}}\right) \right] \\ &+ \frac{x_{k+1} - x}{\Delta} \left[\mathcal{N}\left(\frac{x_{k+1} - x}{\sqrt{T-t}}\right) - \mathcal{N}\left(\frac{x_k - x}{\sqrt{T-t}}\right) \right] \\ &+ \frac{\sqrt{T-t}}{\Delta\sqrt{2\pi}} \left[e^{\frac{(x_{k-1}-x)^2}{2(T-t)}} - 2e^{\frac{(x_k-x)^2}{2(T-t)}} + e^{\frac{(x_{k+1}-x)^2}{2(T-t)}} \right]. \end{aligned}$$

- c) The extension of the previous example to a d -dimensional Brownian Motion $W = (W^1, \dots, W^d)$ is directly obtained by tensorization, leading to $K_{\text{pcv}} = (K_{\text{pcv}}^{(d=1)})^d$ control variates:

$$P_k(Z) := P_{k_1, \dots, k_d}(W_t, W_T) := \prod_{i=1}^d p_{x_{k_i}, \Delta}(W_T^i) - \prod_{i=1}^d \mathbb{E} \left[p_{x_{k_i}, \Delta}(W_T^i) | W_t^i \right],$$

where $1 \leq k_i \leq K_{\text{pcv}}^{(d=1)}$, $1 \leq i \leq d$.

We now consider *control variates with full support*.

- d) Associated to a standard Brownian motion $(W_s)_{0 \leq s \leq T}$, put $X = W_t, Z = (W_t, W_T)$ and let $(H_k)_k$ be the Hermite polynomials given by $H_k(x) := e^{x^2/2} \frac{d^k}{dx^k} (e^{-x^2/2})$. Because of the martingale property of $\left(t^{k/2} H_k \left(\frac{W_t}{\sqrt{t}} \right) \right)_{t \geq 0}$ (see [14, Chapter 4]), one can take as control variates

$$P_k(Z) := T^{k/2} H_k \left(\frac{W_T}{\sqrt{T}} \right) - t^{k/2} H_k \left(\frac{W_t}{\sqrt{t}} \right).$$

- e) Similarly, for a Poisson process $(N_s)_{0 \leq s \leq T}$, let $(C_k)_k$ be the Charlier polynomials defined by $\sum_{i=0}^{\infty} C_i(y, a) \frac{w^i}{i!} = e^w (1 - \frac{w}{a})^y$. Since $(t^k C_k(N_t, t))_{0 \leq t \leq T}$ is a martingale,

$$P_k(Z) := T^k C_k(N_T, T) - t^k C_k(N_t, t)$$

defines a control variate. Other examples of orthogonal polynomials in relation with stochastic processes are provided in [14].

- f) Recently in [5], Cuchiero *et al.* have proved that polynomials of stochastic processes are preserved by conditional expectations, for a large class of models including affine processes, processes with quadratic diffusion coefficients, Lévy-driven SDEs with affine vector fields... all these examples naturally lead to PCVs of polynomial type.

- g) More generally, let $(Y_s)_{0 \leq s \leq T}$ be a \mathbb{R}^d -valued diffusion process

$$dY_t = b(t, Y_t)dt + \sigma(t, Y_t)dW_t$$

(assuming globally Lipschitz coefficients), and let $(p_k)_{1 \leq k \leq K_{\text{pcv}}}$ be a family of smooth functions. Put $X = Y_t, Z = (Y_t, U_{t,T}, Y_{U_{t,T}}, Y_T)$ where $U_{t,T}$ is a random variable uniformly distributed on $[t, T]$, independent of $(Y_s)_{0 \leq s \leq T}$. Denoting by \mathcal{L} the infinitesimal generator and using Itô's formula, it is easy to check that

$$P_k(Z) := p_k(T, Y_T) - (T - t)(\partial_t + \mathcal{L})p_k(U_{t,T}, Y_{U_{t,T}}) - p_k(t, Y_t)$$

is such $\mathbb{E}[P_k(Z)|Y_t] = 0$, thus it is a control variate.

Notations and algorithm description

Let \mathcal{A} be a non-empty closed convex subset of $\mathbb{R}^{K_{\text{pcv}}}$ and for $\alpha \in \mathcal{A}$, put

$$H^\alpha = H - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k P_k(Z).$$

The PCV algorithm is as follows.

STEP 1. *Variance Reduction.* We approximate $(\alpha_k^*)_{1 \leq k \leq K_{\text{pcv}}}$ by the coefficients $(\tilde{\alpha}_k)_{1 \leq k \leq K_{\text{pcv}}}$ using the empirical norm in (1):

$$(\tilde{\alpha}_k)_{1 \leq k \leq K_{\text{pcv}}} := \underset{\alpha \in \mathcal{A}}{\operatorname{arginf}} \frac{1}{N} \sum_{i=1}^N \left| H_i - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k P_k(Z_i) \right|^2.$$

STEP 2. *Least squares regression.* We compute the coefficients of the least squares regression $(\tilde{\beta}_k)_{1 \leq k \leq K_{\mathcal{F}}}$ verifying:

$$(\tilde{\beta}_k)_{1 \leq k \leq K_{\mathcal{F}}} := \underset{(\beta_k)_k}{\operatorname{arginf}} \frac{1}{N} \sum_{i=1}^N \left| H_i - \sum_{k=1}^{K_{\text{pcv}}} \tilde{\alpha}_k P_k(Z_i) - \sum_{k=1}^{K_{\mathcal{F}}} \beta_k \Phi_k(X_i) \right|^2.$$

Then, \tilde{m}_N is defined by

$$\tilde{m}_N := \sum_{k=1}^{K_{\mathcal{F}}} \tilde{\beta}_k \Phi_k,$$

because the regression function of $H - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k^* P_k(Z)$ is also $m(\cdot)$, due to the fact that $\mathbb{E}[P_k(Z)|X] = 0$.

Hypothesis

To prove our theorems, we assume

(H) For some $L \geq 1$,

- i) $\|P_k\|_\infty \leq 1$ and $\|h\|_\infty \leq L$,
- ii) $\mathcal{A} := \{\alpha \in \mathbb{R}^{K_{\text{pcv}}} : \sum_{i=1}^{K_{\text{pcv}}} |\alpha_i| \leq L\}$.

Clearly, \mathcal{A} is a non empty closed convex set. These are technical assumptions, which allow us to bound uniformly the PCVs and to apply Hoeffding-type concentration inequalities to them.

Main theorems

Theorem 2.2. *Set $\Sigma^2(\alpha^*) = \sup_{x \in \mathbb{R}^{d_x}} \mathbb{V}\text{ar}(H^{\alpha^*} | X = x)$. Then, for any $\rho > 0$,*

$$\mathbb{E} \left[\|\tilde{m}_N - m\|_{L_2(\mu_N)}^2 \right] \leq (1 + \rho^{-1}) L^4 \frac{[c_1 + (c_2 + c_3 \log(N))(K_{\text{pcv}} + 1)]}{N} \\ + (1 + \rho) \frac{K_{\mathcal{F}}}{N} \Sigma^2(\alpha^*) + (1 + \rho) \mathbb{E} \left[\inf_{f \in \mathcal{F}_N} \|f - m\|_{L_2(\mu_N)}^2 \right],$$

where c_1, c_2 et c_3 are universal constants.

This theorem shows that the PCV regression presumably reduces the estimation error through the factor $\Sigma^2(\alpha^*)$ instead of the usual $\Sigma^2 = \Sigma^2(0)$. But on the other hand, the error contains an additional term (related to c_1, c_2, c_3) because of the extra error in the estimation of α^* by $\tilde{\alpha}$.

Hence, in practice, we have to choose a relatively small K_{pcv} with respect to $K_{\mathcal{F}}$. In the numerical tests, we illustrate this issue.

Theorem 2.3. *Consider piecewise constant \mathcal{F}_N -basis functions of the form $\Phi_k = \mathbf{1}_{I_k}$ where $(I_k)_{1 \leq k \leq K_{\mathcal{F}}}$ are disjoint sets. Assume that one of the following assumptions holds, for some constant $c_I \geq 1$:*

- a) $(I_k)_k$ depends on (X_1, \dots, X_N) and their frequencies dominate the uniform distribution, i.e.

$$\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{X_i \in I_k} \geq \frac{1}{c_I K_{\mathcal{F}}}, \quad 1 \leq k \leq K_{\mathcal{F}};$$

- b) $(I_k)_k$ are deterministic and their occurrences w.r.t. the distribution of X dominate the uniform distribution, i.e.

$$\mathbb{P}(X \in I_k) \geq \frac{1}{c_I K_{\mathcal{F}}}, \quad 1 \leq k \leq K_{\mathcal{F}}.$$

Then, for any $\rho > 0$,

$$\mathbb{E} \left[\|\tilde{m}_N - m\|_{L_2(\mu_N)}^2 \right] \leq (1 + \rho^{-1}) L^4 \frac{[c_1 + (c_2 + c_3 \log(N))(K_{\text{pcv}} + 1)]}{N} \\ + (1 + \rho) c_I \frac{K_{\mathcal{F}}}{N} \inf_{\alpha \in \mathcal{A}} \mathbb{E} [\mathbb{V}\text{ar}(H^{\alpha} | X)] \\ + (1 + \rho) \mathbb{E} \left[\inf_{f \in \mathcal{F}_N} \|f - m\|_{L_2(\mu_N)}^2 \right],$$

with the same constants c_1, c_2 et c_3 as before.

In Theorem 2.3, provided an additional hypothesis on Φ_k , we obtain a more precise bound for the overall error, since the second term factor is $\inf_{\alpha \in \mathcal{A}} \mathbb{E}[\text{Var}(H^\alpha|X)]$, which is exactly the quantity minimized in STEP 1 of the PCV algorithm.

3 Proofs

We first introduce notations specific to the proofs.

- i) Denote by $\langle \cdot, \cdot \rangle_\mu$ and $\langle \cdot, \cdot \rangle_{\mu_N}$ the scalar products related to the L_2 -norms $\|\cdot\|_\mu$ et $\|\cdot\|_{\mu_N}$ introduced before: for two functions $g_1, g_2 : \mathbb{R}^{d_z} \times \mathbb{R}^{d_x} \rightarrow \mathbb{R}$,

$$\langle g_1, g_2 \rangle_\mu = \int g_1(z, x)g_2(z, x)\mu(dz, dx), \quad \langle g_1, g_2 \rangle_{\mu_N} = \frac{1}{N} \sum_{i=1}^N g_1(Z_i, X_i)g_2(Z_i, X_i).$$

- ii) *PCV spaces*: set

$$\mathcal{G} := \left\{ \sum_{k=1}^{K_{\text{pcv}}} \alpha_k P_k : \alpha \in \mathbb{R}^{K_{\text{pcv}}} \right\}, \quad \mathcal{G}^{\mathcal{A}} := \left\{ \sum_{k=1}^{K_{\text{pcv}}} \alpha_k P_k : \alpha \in \mathcal{A} \right\},$$

$$T_L \mathcal{G} := \{T_L g : g \in \mathcal{G}\},$$

where T_L is the truncation operator of functions at the level L defined by

$$T_L g(x) = -L \vee g(x) \wedge L.$$

It is important to observe that the assumption **(H)** implies $\mathcal{G}^{\mathcal{A}} \subset T_L \mathcal{G}$.

- iii) Define \tilde{m}_N^* by

$$\tilde{m}_N^* = \sum_{k=1}^{K_{\mathcal{F}}} \tilde{\beta}_k^* \Phi_k$$

where

$$(\alpha_k^*)_{1 \leq k \leq K_{\text{pcv}}} = \underset{\alpha \in \mathcal{A}}{\text{arginf}} \mathbb{E} \left[\left| H - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k P_k(Z) \right|^2 \right],$$

$$(\tilde{\beta}_k^*)_{1 \leq k \leq K_{\mathcal{F}}} = \underset{(\beta_k)_k}{\text{arginf}} \frac{1}{N} \sum_{i=1}^N \left| H_i - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k^* P_k(Z_i) - \sum_{k=1}^{K_{\mathcal{F}}} \beta_k \Phi_k(X_i) \right|^2.$$

We now present some tools useful in the theory of non-parametric regression, see [9, Chapter 9] for full details.

Definition 3.1. [Covering numbers $\mathcal{N}_p(\varepsilon, \mathcal{G}, z_1^M)$] Let \mathcal{G} be a set of functions from \mathbb{R}^d to \mathbb{R} and let $z_1^M = (z_1, \dots, z_M)$ be M points in \mathbb{R}^d : denote by ν_M the corresponding empirical measure and by $\|\cdot\|_{L_p(\nu_M)}$ the related L_p -norm ($p \geq 1$).

- (i) For $\varepsilon > 0$, $(g_j : \mathbb{R}^d \rightarrow \mathbb{R})_{1 \leq j \leq n}$ is called an ε -cover of \mathcal{G} w.r.t $\|\cdot\|_{L_p(\nu_M)}$, if for every $g \in \mathcal{G}$, there is a $j \in \{1, \dots, n\}$ such that $\|g - g_j\|_{L_p(\nu_M)} < \varepsilon$.
- (ii) We denote by $\mathcal{N}_p(\varepsilon, \mathcal{G}, z_1^M)$ the minimal size n of ε -covers $(g_j)_{1 \leq j \leq n}$ of \mathcal{G} w.r.t $\|\cdot\|_{L_p(\nu_M)}$.

The combination of Lemma 9.2, Theorem 9.4, Theorem 9.5 and Equation (10.23) in [9] leads to the following ready-to-use estimates.

Proposition 3.2. *For any z_1^M points in \mathbb{R}^d and any $\varepsilon \in]0, L/2]$, we have*

$$\begin{aligned} \mathcal{N}_1(\varepsilon, T_L \mathcal{G}, z_1^M) &\leq 3 \left(\frac{2e(2L)}{\varepsilon} \log \left(\frac{3e(2L)}{\varepsilon} \right) \right)^{K_{\text{pcv}}+1} \leq 3 \left(\frac{9L}{\varepsilon} \right)^{2(K_{\text{pcv}}+1)}, \\ \mathcal{N}_2(\varepsilon, T_L \mathcal{G}, z_1^M) &\leq 3 \left(\frac{2e(2L)^2}{\varepsilon^2} \log \left(\frac{3e(2L)^2}{\varepsilon^2} \right) \right)^{K_{\text{pcv}}+1} \leq 3 \left(\frac{18L^2}{\varepsilon^2} \right)^{2(K_{\text{pcv}}+1)}. \end{aligned}$$

Actually, the second series of inequalities are easily derived from $\log(x) \leq x/e$ for any $x > 0$.

3.1 Proof of Theorem 2.2

We have :

$$\begin{aligned} \mathbb{E} \left[\|\tilde{m}_N - m\|_{L_2(\mu_N)}^2 \right] &\leq (1 + \rho^{-1}) \mathbb{E} \left[\|\tilde{m}_N - \tilde{m}_N^*\|_{L_2(\mu_N)}^2 \right] \\ &\quad + (1 + \rho) \mathbb{E} \left[\|\tilde{m}_N^* - m\|_{L_2(\mu_N)}^2 \right]. \end{aligned} \quad (3)$$

Step 1: Bound for $\mathbb{E}[\|\tilde{m}_N - \tilde{m}_N^*\|_{L_2(\mu_N)}^2]$. \tilde{m}_N and \tilde{m}_N^* are respectively the orthogonal projections onto $\text{Span}(\Phi_k : 1 \leq k \leq K_{\mathcal{F}})$ with respect to the scalar product $\langle \cdot, \cdot \rangle_{\mu_N}$ of $h - \sum_{k=1}^{K_{\text{pcv}}} \tilde{\alpha}_k P_k$ and $h - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k^* P_k$. By the contraction property of projection, we get:

$$\begin{aligned} \mathbb{E} \left[\|\tilde{m}_N - \tilde{m}_N^*\|_{L_2(\mu_N)}^2 \right] &\leq \mathbb{E} \left[\left\| \sum_{k=1}^{K_{\text{pcv}}} \tilde{\alpha}_k P_k - \sum_{k=1}^{K_{\text{pcv}}} \alpha_k^* P_k \right\|_{L_2(\mu_N)}^2 \right] \\ &= \mathbb{E} \left[\|r_N - r\|_{L_2(\mu_N)}^2 \right] \end{aligned} \quad (4)$$

where we have set $r := \sum_{k=1}^{K_{\text{pcv}}} \alpha_k^* P_k$ and $r_N := \sum_{k=1}^{K_{\text{pcv}}} \tilde{\alpha}_k P_k$.

We now estimate $\mathbb{E}[\|r_N - r\|_{L_2(\mu)}^2]$ before deducing a bound on $\mathbb{E}[\|r_N - r\|_{L_2(\mu_N)}^2]$. Actually, in view of (1), r is the $L_2(\mu)$ -projection of h onto the non-empty closed convex set $\mathcal{G}^{\mathcal{A}}$; since r_N is also in $\mathcal{G}^{\mathcal{A}}$, we have

$$\int |r_N(z) - r(z)|^2 \mu(\mathbf{d}z, \mathbf{d}x) \leq \int |r_N(z) - h(z)|^2 \mu(\mathbf{d}z, \mathbf{d}x) - \mathbb{E}[|r(Z) - H|^2].$$

Because of $\frac{1}{N} \sum_{i=1}^N |r_N(Z_i) - H_i|^2 - \frac{1}{N} \sum_{i=1}^N |r(Z_i) - H_i|^2 \leq 0$, we deduce:

$$\begin{aligned} & \int |r_N(z) - r(z)|^2 \mu(\mathbf{d}z, \mathbf{d}x) \\ & \leq T_{1,N} := \int |r_N(z) - h(z)|^2 \mu(\mathbf{d}z, \mathbf{d}x) - \mathbb{E}[|r(Z) - H|^2] \\ & \quad - 2 \left(\frac{1}{N} \sum_{i=1}^N |r_N(Z_i) - H_i|^2 - \frac{1}{N} \sum_{i=1}^N |r(Z_i) - H_i|^2 \right). \end{aligned}$$

Let us estimate $\mathbb{E}(T_{1,N})$ by controlling $\mathbb{P}(T_{1,N} > t)$, $t > 0$. Since $r_N \in \mathcal{G}^{\mathcal{A}}$, we have

$$\begin{aligned} \mathbb{P}(T_{1,N} > t) & \leq \mathbb{P} \left(\exists f \in \mathcal{G}^{\mathcal{A}} : \mathbb{E}[|f(Z) - H|^2] - \mathbb{E}[|r(Z) - H|^2] \right. \\ & \quad \left. - \left[\frac{1}{N} \sum_{i=1}^N |f(Z_i) - H_i|^2 - \frac{1}{N} \sum_{i=1}^N |r(Z_i) - H_i|^2 \right] \right. \\ & \quad \left. > \frac{1}{2} \left[\frac{t}{2} + \frac{t}{2} + \mathbb{E}[|f(Z) - H|^2] - \mathbb{E}[|r(Z) - H|^2] \right] \right). \end{aligned}$$

By using Theorem A.2 in Appendix and Proposition 3.2 with $\mathcal{G}^{\mathcal{A}} \subset T_L \mathcal{G}$, and restricting to $t \geq \frac{320L^4}{N}$, we get

$$\begin{aligned} \mathbb{P}(T_{1,N} > t) & \leq 14 \sup_{z_1^N} \mathcal{N}_1 \left(\frac{t}{640L^3}, \mathcal{G}^{\mathcal{A}}, z_1^N \right) \exp \left(-\frac{N}{5136L^4} t \right) \\ & \leq 14 \sup_{z_1^N} \mathcal{N}_1 \left(\frac{L}{2N}, T_L \mathcal{G}, z_1^N \right) \exp \left(-\frac{N}{5136L^4} t \right) \\ & \leq 42(18N)^{2(K_{\text{pcv}}+1)} \exp \left(-\frac{N}{5136L^4} t \right). \end{aligned}$$

We deduce, for any $\varepsilon \geq \frac{320L^4}{N}$,

$$\mathbb{E}(T_{1,N}) \leq \varepsilon + \int_{\varepsilon}^{\infty} 42(18N)^{2(K_{\text{pcv}}+1)} \exp \left(-\frac{N}{5136L^4} t \right) dt$$

$$\leq \varepsilon + 42(18N)^{2(K_{\text{pcv}}+1)} \frac{5136L^4}{N} \exp\left(-\frac{N}{5136L^4}\varepsilon\right).$$

The expression above is minimal for $\varepsilon = \frac{5136L^4}{N} \log(42(18N)^{2(K_{\text{pcv}}+1)})$ (note that $\varepsilon \geq \frac{320L^4}{N}$). By plugging this value, we find :

$$\mathbb{E}\left[\|r_N - r\|_{L_2(\mu)}^2\right] \leq \mathbb{E}(T_{1,N}) \leq \frac{5136L^4}{N} (\log(42e) + 2(K_{\text{pcv}} + 1) \log(18N)). \quad (5)$$

To deduce a bound for $\mathbb{E}\left[\|r_N - r\|_{L_2(\mu_N)}^2\right]$, we set

$$T_{2,N} := 2(\max\{\|r_N - r\|_{L_2(\mu_N)} - 2\|r_N - r\|_{L_2(\mu)}, 0\})^2$$

and we write the decomposition

$$\begin{aligned} \|r_N - r\|_{L_2(\mu_N)}^2 &\leq (\max\{\|r_N - r\|_{L_2(\mu_N)} - 2\|r_N - r\|_{L_2(\mu)}, 0\} + 2\|r_N - r\|_{L_2(\mu)})^2 \\ &\leq T_{2,N} + 8\|r_N - r\|_{L_2(\mu)}^2. \end{aligned} \quad (6)$$

Let $u > \frac{144L^2}{N}$. Using $r_N \in \mathcal{G}^A \subset T_L\mathcal{G}$, Lemma A.1 in Appendix and Proposition 3.2, we obtain

$$\begin{aligned} \mathbb{P}(T_{2,N} > u) &\leq \mathbb{P}(\exists f \in T_L\mathcal{G} : \|f - r\|_{L_2(\mu_N)} - 2\|f - r\|_{L_2(\mu)} > \sqrt{u/2}) \\ &\leq 3 \sup_{z_1^{2N}} \mathcal{N}_2\left(\frac{\sqrt{u}}{24}, T_L\mathcal{G}, z_1^{2N}\right) \exp\left(-\frac{N}{2304L^2}u\right) \\ &\leq 3 \sup_{z_1^{2N}} \mathcal{N}_2\left(\frac{L}{2\sqrt{N}}, T_L\mathcal{G}, z_1^{2N}\right) \exp\left(-\frac{N}{2304L^2}u\right) \\ &\leq 9(72N)^{2(K_{\text{pcv}}+1)} \exp\left(-\frac{N}{2304L^2}u\right). \end{aligned}$$

Thus, similarly to the evaluation of $\mathbb{E}(T_{1,N})$, we obtain, for any $\varepsilon > \frac{144L^2}{N}$,

$$\mathbb{E}(T_{2,N}) \leq \varepsilon + 9(72N)^{2(K_{\text{pcv}}+1)} \frac{2304L^2}{N} \exp\left(-\frac{N}{2304L^2}\varepsilon\right).$$

For the choice $\varepsilon = \frac{2304L^2}{N} \log(9(72N)^{2(K_{\text{pcv}}+1)}) > \frac{144L^2}{N}$, we get:

$$\mathbb{E}(T_{2,N}) \leq \frac{2304L^2}{N} (\log(9e) + 2(K_{\text{pcv}} + 1) \log(72N)). \quad (7)$$

Plugging (5-6-7) into (4) and using $L^2 \leq L^4$, we obtain

$$\mathbb{E}\left[\|\tilde{m}_N - \tilde{m}_N^*\|_{L_2(\mu_N)}^2\right] \leq L^4 \frac{c_1 + (c_2 + c_3 \log(N))(K_{\text{pcv}} + 1)}{N}. \quad (8)$$

Step 2: Bound for $\mathbb{E}[\|\tilde{m}_N^* - m\|_{L_2(\mu_N)}^2]$. To simplify, we introduce the notation $\mathbb{E}^*[\cdot] := \mathbb{E}[\cdot | X_1, \dots, X_N]$.

Consider a complete orthonormal basis (f_1, \dots, f_K) ($K \leq K_{\mathcal{F}}$) for \mathcal{F}_N w.r.t. the empirical scalar product $\langle \cdot, \cdot \rangle_{\mu_N}$:

$$\langle f_k, f_l \rangle_{\mu_N} = \frac{1}{N} \sum_{i=1}^N f_k(X_i) f_l(X_i) = \delta_{k,l}. \quad (9)$$

Observe that

$$\begin{aligned} \mathbb{E}^*[\|\tilde{m}_N^* - m\|_{L_2(\mu_N)}^2] &= \mathbb{E}^* \left[\frac{1}{N} \sum_{i=1}^N |\tilde{m}_N^*(X_i) - \mathbb{E}^*[\tilde{m}_N^*(X_i)]|^2 \right] \\ &\quad + \mathbb{E}^* \left[\frac{1}{N} \sum_{i=1}^N |\mathbb{E}^*[\tilde{m}_N^*(X_i)] - m(X_i)|^2 \right] \end{aligned}$$

because $\mathbb{E}^* \left[\frac{1}{N} \sum_{i=1}^N (\tilde{m}_N^*(X_i) - \mathbb{E}^*[\tilde{m}_N^*(X_i)])(\mathbb{E}^*[\tilde{m}_N^*(X_i)] - m(X_i)) \right] = 0$.

This proves

$$\mathbb{E} \left[\|\tilde{m}_N^* - m\|_{L_2(\mu_N)}^2 \right] = \mathbb{E} \left[\|\tilde{m}_N^* - \mathbb{E}^*[\tilde{m}_N^*]\|_{L_2(\mu_N)}^2 \right] + \mathbb{E} \left[\|\mathbb{E}^*[\tilde{m}_N^*] - m\|_{L_2(\mu_N)}^2 \right]. \quad (10)$$

Similarly to the proof of [9, Theorem 11.1, pp.185–187], we have

$$\mathbb{E} \left[\|\mathbb{E}^*[\tilde{m}_N^*] - m\|_{L_2(\mu_N)}^2 \right] = \mathbb{E} \left[\inf_{f \in \mathcal{F}_N} \|f - m\|_{L_2(\mu_N)}^2 \right], \quad (11)$$

$$\mathbb{E}^* \left[\|\tilde{m}_N^* - \mathbb{E}^*[\tilde{m}_N^*]\|_{\mu_N}^2 \right] = \frac{1}{N^2} \sum_{i=1}^N \text{Var}(H_i^{\alpha^*} | X_i) \sum_{k=1}^K (f_k(X_i))^2. \quad (12)$$

Introducing $\Sigma^2(\alpha^*)$ for the uniform bound regarding to the conditional variance of H^{α^*} and using $\|f_k\|_{L_2(\mu_N)} = 1$, we obtain

$$\mathbb{E}^* \left[\|\tilde{m}_N^* - \mathbb{E}^*[\tilde{m}_N^*]\|_{L_2(\mu_N)}^2 \right] \leq \frac{K}{N} \Sigma^2(\alpha^*).$$

Thus, since $K \leq K_{\mathcal{F}}$, it follows

$$\mathbb{E} \left[\|\tilde{m}_N^* - \mathbb{E}^*[\tilde{m}_N^*]\|_{L_2(\mu_N)}^2 \right] \leq \frac{K_{\mathcal{F}}}{N} \Sigma^2(\alpha^*). \quad (13)$$

We complete the proof by combining (3-8-10-11-13).

3.2 Proof of Theorem 2.3

Case a) The orthonormal basis of \mathcal{F}_N w.r.t. $\langle \cdot, \cdot \rangle_{\mu_N}$ (see (9)) is readily given by

$$f_k = \frac{\mathbf{1}_{I_k}}{\sqrt{\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{X_i \in I_k}}} \leq \sqrt{c_I K_{\mathcal{F}}} \mathbf{1}_{I_k}, \quad 1 \leq k \leq K = K_{\mathcal{F}}.$$

Therefore, we have $\sum_{k=1}^K (f_k(X_i))^2 \leq c_I K_{\mathcal{F}}$ and from (12), we deduce

$$\mathbb{E}[\|\tilde{m}_N^* - \mathbb{E}^*[\tilde{m}_N^*]\|_{L_2(\mu_N)}^2] \leq \frac{1}{N} \mathbb{E}[\text{Var}(H^{\alpha^*} | X)] c_I K_{\mathcal{F}}. \quad (14)$$

Gathering the identity (2) and the inequalities (3-8-10-11-14) leads to the theorem result.

Case b) Concerning the orthogonalisation of $(\Phi_k)_{1 \leq k \leq K_{\mathcal{F}}}$, only $K \leq K_{\mathcal{F}}$ sets contain at least one data. More precisely, for such a set I_k for which $p_{k,N} := \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{X_i \in I_k} > 0$, we can set

$$f_k = \frac{\mathbf{1}_{I_k}}{\sqrt{p_{k,N}}}.$$

Hence, the last factor in (12) is equal to $\sum_{k=1}^K (f_k(X_i))^2 = \sum_{k=1}^{K_{\mathcal{F}}} \frac{\mathbf{1}_{X_i \in I_k}}{p_{k,N}} \mathbf{1}_{p_{k,N} > 0}$ using the convention $0/0 = 0$. Then, by a symmetry argument within (X_1, \dots, X_N) , we have

$$\begin{aligned} \mathbb{E}[\|\tilde{m}_N^* - \mathbb{E}^*[\tilde{m}_N^*]\|_{L_2(\mu_N)}^2] &= \frac{1}{N} \mathbb{E} \left[\text{Var}(H_1^{\alpha^*} | X_1) \sum_{k=1}^{K_{\mathcal{F}}} \frac{\mathbf{1}_{X_1 \in I_k}}{p_{k,N}} \mathbf{1}_{p_{k,N} > 0} \right] \\ &= \sum_{k=1}^{K_{\mathcal{F}}} \mathbb{E} \left\{ \text{Var}(H_1^{\alpha^*} | X_1) \mathbf{1}_{X_1 \in I_k} \mathbb{E} \left[\frac{1}{1 + \sum_{i=2}^N \mathbf{1}_{X_i \in I_k}} \right] \right\}. \end{aligned}$$

Observe that $\sum_{i=2}^N \mathbf{1}_{X_i \in I_k}$ is binomially distributed with parameters $(N-1, \mathbb{P}(X \in I_k))$: from [9, Lemma 4.1] we know that

$$\mathbb{E} \left[\frac{1}{1 + \sum_{i=2}^N \mathbf{1}_{X_i \in I_k}} \right] \leq \frac{1}{N \mathbb{P}(X \in I_k)} \leq \frac{c_I K_{\mathcal{F}}}{N}.$$

Hence, $\mathbb{E}[\|\tilde{m}_N^* - \mathbb{E}^*[\tilde{m}_N^*]\|_{L_2(\mu_N)}^2] \leq c_I \frac{K_{\mathcal{F}}}{N} \inf_{\alpha \in \mathbb{R}^{K_{\text{pcv}}}} \mathbb{E}[\text{Var}(H^{\alpha} | X)]$. We conclude as for Case a).

4 Numerical experiments

Consider two independent Brownian motions W and B . We experiment the method in two cases related to the dimension of X : $d_x = 1$ and $d_x = 2$.

4.1 Dimension 1

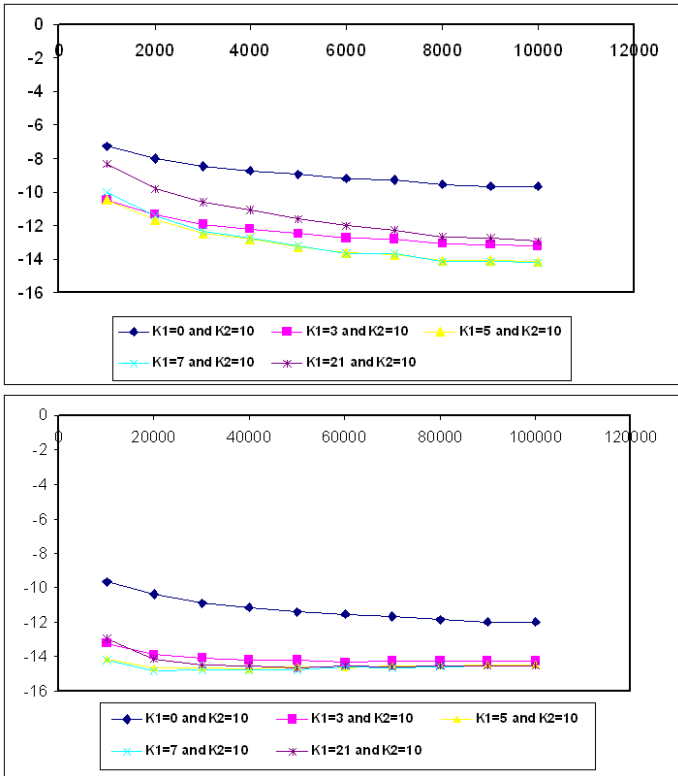


Figure 1. Empirical error (in a log scale) as a function of N . Parameters: $d_x = 1$, $K_1 = K_{pcv}$, $K_2 = K_{\mathcal{F}}$.

Our goal is to estimate $m(x) = \mathbb{E}[h(W_2)|W_1 = x]$ where $h(x) = e^{-\frac{x^2}{2}}$; due to model assumption, m is explicit. For the PCVs, we take the hat functions, see example b) in Paragraph 2.3. Regarding \mathcal{F}_N , we choose $\Phi_k(x) = x^k, 0 \leq k \leq 9$.

The Figure 1 shows the global empirical error³ $\mathbb{E} \left[\|\tilde{m}_N - m\|_{L_2(\mu_N)}^2 \right]$ for $K_{\mathcal{F}} = 10$ and various values of K_{pcv} , in a range of few simulations ($N \leq 10000$) and in a range of many simulations ($10000 \leq N \leq 100000$). The variables on the plots are $K_1 = K_{\text{pcv}}$ and $K_2 = K_{\mathcal{F}}$.

We observe that the PCV method, compared to a simple regression, improves the convergence of the global error. However, the PCV efficiency for small N deteriorates when K_{pcv} becomes large ($K_{\text{pcv}} = 21$). This is explained by the fact that the statistical error regarding the estimation of α^* (term with c_1, c_2, c_3 in Theorem 2.2) is significant for small N and large K_{pcv} . Thus, for a small number of simulations, using $K_{\text{pcv}} = 3$ or 5 is optimal. The standard method with $N = 100000$ yields an error equivalent to that using PCV with $N = 2000$: hence, the PCV yields a improvement factor of 50 regarding to the simulation effort. In the range of large N , the PCV method reaches the approximation error while the standard method still requires more simulations.

4.2 Dimension 2

Consider the estimation of $m(x) = \mathbb{E}[h(W_2, B_2) | W_1 = x, B_1 = x]$ where $h(x_1, x_2) = e^{-\frac{x_1^2 + x_2^2 + \rho x_1 x_2}{2}}$ with $\rho = 0.5$. We use the hat functions in dimension 2, see example c) in Paragraph 2.3. The Figure 2 represents the global empirical error $\mathbb{E} \left[\|\tilde{m}_N - m\|_{L_2(\mu_N)}^2 \right]$ when $K_{\mathcal{F}} = 13 \times 13$ and with different values of K_{pcv} . We observe the same features as in dimension 1. Here, the improvement factor is about 25 comparing the PCV with $K_{\text{pcv}} = 5 \times 5$ at $N = 2000$ and the standard method at $N = 50000$.

5 Conclusion and perspectives

The PCV method significantly accelerates the convergence of the estimation error to 0, regardless of the selected approximation space. It provides a higher accuracy of regression-based Monte-Carlo algorithms, especially for few simulations; thus, it can be used for efficiently reducing the computational time. However, in view of theoretical and numerical results, a special attention has to be paid to the choice of the PCVs and their number K_{pcv} , which has to be small compared to the dimension $K_{\mathcal{F}}$ of the approximation space.

An adaptive selection procedure of PCVs would be worth being designed, which is left to further research. Moreover, we will investigate how to relax the bound-

³ the outside expectation is computed through 100 runs of the algorithm.

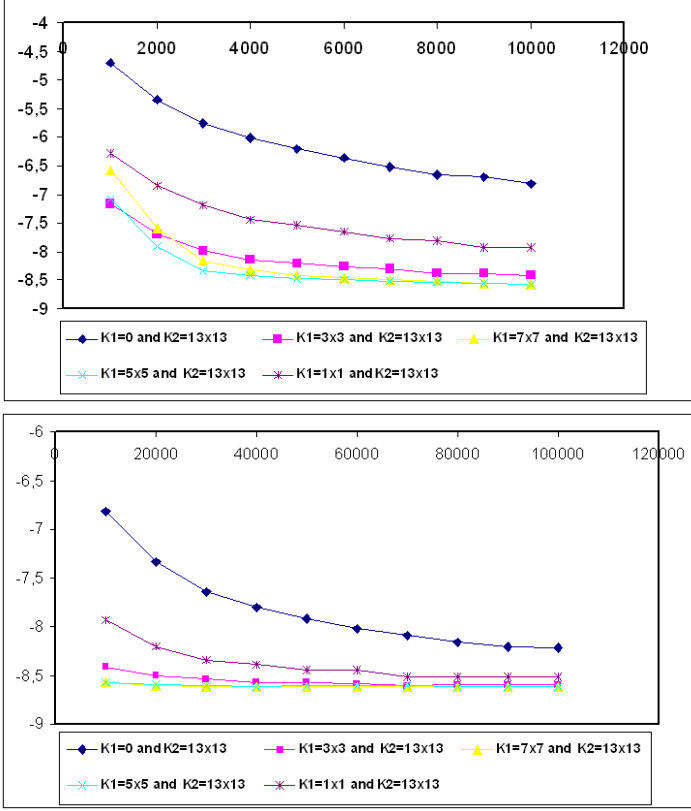


Figure 2. Empirical error (in a log scale) as a function of N . Parameters: $d_x = 2$, $K_1 = K_{\text{pcv}}$, $K_2 = K_{\mathcal{F}}$.

edness assumptions on P_k . It would allow to consider polynomials for the PCVs, see [3] for promising related numerical experiments.

A Appendix

Lemma A.1. *Let \mathcal{F} be a countable set of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ bounded by L and let $Z_1^{2N} = (Z_1, \dots, Z_N, Z_{N+1}, \dots, Z_{2N})$ be a sample of i.i.d random variables. Let the empirical measure μ_N be defined w.r.t. $Z_1^N = (Z_1, \dots, Z_N)$. For $\varepsilon > 0$, we have*

$$\mathbb{P}(\exists f \in \mathcal{F} : \|f\|_{L_2(\mu_N)} - 2\|f\|_{L_2(\mu)} > \varepsilon)$$

$$\leq 3\mathbb{E} \left[\mathcal{N}_2 \left(\frac{\sqrt{2}}{24} \varepsilon, \mathcal{F}, Z_1^{2N} \right) \right] \exp \left(-\frac{N\varepsilon^2}{288L^2} \right).$$

Proof: we can restrict to the case $N \geq \frac{4L^2}{\varepsilon^2}$ since otherwise, the above inequality is obvious.

We follow the four steps of the proof of [9, Theorem 11.2]. Only STEP 1 is different, we detail it here. Define:

$$\|f\|_{L_2(\mu_N)}'^2 = \frac{1}{N} \sum_{i=N+1}^{2N} |f(Z_i)|^2.$$

Let f^* be a function in \mathcal{F} such as

$$\|f^*\|_{L_2(\mu_N)} - 2\|f^*\|_{L_2(\mu)} > \varepsilon$$

if there exists any such function, otherwise let f^* be another arbitrary fixed function in \mathcal{F} ; note that f^* depends on Z_1^N . Then, basic computations and Chebyshev inequality yield

$$\begin{aligned} & \mathbb{P} \left(2\|f^*\|_{L_2(\mu)} + \frac{\varepsilon}{2} > \|f^*\|_{L_2(\mu_N)}' \mid Z_1^N \right) \\ & \geq \mathbb{P} \left(4\|f^*\|_{L_2(\mu)}^2 + \frac{\varepsilon^2}{4} > \|f^*\|_{L_2(\mu_N)}'^2 \mid Z_1^N \right) \\ & = 1 - \mathbb{P} \left(3\|f^*\|_{L_2(\mu)}^2 + \frac{\varepsilon^2}{4} \leq \|f^*\|_{L_2(\mu_N)}'^2 - \|f^*\|_{L_2(\mu)}^2 \mid Z_1^N \right) \\ & \geq 1 - \frac{\text{Var} \left(\frac{1}{N} \sum_{i=N+1}^{2N} |f^*(Z_i)|^2 \mid Z_1^N \right)}{\left(3\|f^*\|_{L_2(\mu)}^2 + \frac{\varepsilon^2}{4} \right)^2} \\ & \geq 1 - \frac{\frac{1}{N}L^2\|f^*\|_{L_2(\mu)}^2}{\left(3\|f^*\|_{L_2(\mu)}^2 + \frac{\varepsilon^2}{4} \right)^2} \geq 1 - \frac{L^2}{3N} \frac{4}{\varepsilon^2} \geq \frac{2}{3} \end{aligned}$$

using the restriction on N . For such N , it follows that

$$\begin{aligned} & \mathbb{P} \left(\exists f \in \mathcal{F} : \|f\|_{L_2(\mu_N)} - \|f\|_{L_2(\mu_N)}' > \frac{\varepsilon}{4} \right) \\ & \geq \mathbb{P} \left(\|f^*\|_{L_2(\mu_N)} - \|f^*\|_{L_2(\mu_N)}' > \frac{\varepsilon}{4} \right) \\ & \geq \mathbb{P} \left(\|f^*\|_{L_2(\mu_N)} + \frac{3\varepsilon}{4} - \|f^*\|_{L_2(\mu_N)}' > \varepsilon, 2\|f^*\|_{L_2(\mu)} + \frac{3\varepsilon}{4} > \|f^*\|_{L_2(\mu_N)}' \right) \end{aligned}$$

$$\begin{aligned}
&\geq \mathbb{P} \left(\|f^*\|_{L_2(\mu_N)} - 2\|f^*\|_{L_2(\mu)} > \varepsilon, 2\|f^*\|_{L_2(\mu)} + \frac{3\varepsilon}{4} > \|f^*\|'_{L_2(\mu_N)} \right) \\
&\geq \mathbb{E} \left(\mathbf{1}_{\{\|f^*\|_{L_2(\mu_N)} - 2\|f^*\|_{L_2(\mu)} > \varepsilon\}} \mathbb{P}(2\|f^*\|_{L_2(\mu)} + \frac{\varepsilon}{2} > \|f\|'_{L_2(\mu_N)} \mid Z_1^N) \right) \\
&\geq \frac{2}{3} \mathbb{P} (\|f^*\|_{L_2(\mu_N)} - 2\|f^*\|_{L_2(\mu)} > \varepsilon) \\
&= \frac{2}{3} \mathbb{P} (\exists f \in \mathcal{F} : \|f\|_{L_2(\mu_N)} - 2\|f\|_{L_2(\mu)} > \varepsilon).
\end{aligned}$$

In conclusion, for $N \geq \frac{4L^2}{\varepsilon^2}$, we get

$$\begin{aligned}
&\mathbb{P} (\exists f \in \mathcal{F} : \|f\|_{L_2(\mu_N)} - 2\|f\|_{L_2(\mu)} > \varepsilon) \\
&\leq \frac{3}{2} \mathbb{P} \left(\exists f \in \mathcal{F} : \|f\|_{L_2(\mu_N)} - \|f\|'_{L_2(\mu_N)} > \frac{\varepsilon}{4} \right).
\end{aligned}$$

STEPS 2, 3 and 4 of the proof of [9, Theorem 11.2] lead to:

$$\begin{aligned}
&\mathbb{P} \left(\exists f \in \mathcal{F} : \|f\|_{L_2(\mu_N)} - \|f\|'_{L_2(\mu_N)} > \frac{\varepsilon}{4} \right) \\
&\leq 2\mathbb{E} \left[\mathcal{N}_2 \left(\frac{\sqrt{2}}{24} \varepsilon, \mathcal{F}, Z_1^{2N} \right) \right] \exp \left(-\frac{N\varepsilon^2}{288L^2} \right).
\end{aligned}$$

The lemma is proved.

Theorem A.2. Let $r := \sum_{k=1}^{K_{\text{pev}}} \alpha_k^* P_k$ be the minimizer of $\mathbb{E}(|H - f(Z)|^2)$ over $f \in \mathcal{G}^A$, see (1). For $\alpha, \beta > 0$ and $0 < \varepsilon \leq 1/2$, we have

$$\begin{aligned}
&\mathbb{P} \left\{ \exists f \in \mathcal{G}^A : \mathbb{E}[|f(Z) - H|^2] - \mathbb{E}[|r(Z) - H|^2] \right. \\
&\quad - \left(\frac{1}{N} \sum_{i=1}^N |f(Z_i) - H_i|^2 - \frac{1}{N} \sum_{i=1}^N |r(Z_i) - H_i|^2 \right) \\
&\quad \left. > \varepsilon (\alpha + \beta + \mathbb{E}[|f(Z) - H|^2] - \mathbb{E}[|r(Z) - H|^2]) \right\} \\
&\leq 14 \sup_{z_1^N} \mathcal{N}_1 \left(\frac{\beta\varepsilon}{160L^3}, \mathcal{G}^A, z_1^N \right) \exp \left(-\frac{\varepsilon^2(1-\varepsilon)\alpha N}{214(1+\varepsilon)L^4} \right).
\end{aligned}$$

Proof: this is an adaptation of [9, Theorem 11.4] where we take the above definition for r instead of their $r(z) = \mathbb{E}(H|Z = z)$, and where we correct a small error in their proof (which only modifies a constant).

Let us follow step by step the proof of the quoted reference, by detailing only the differences related to our choice of r and \mathcal{G}^A .

In STEP 1, the single difference might be the inequality at the sixth line of [9, p. 211], i.e.

$$\text{Var}(g_f(Z)) \leq 16L^2 \mathbb{E}[g_f(Z)] \quad (15)$$

for any given $f \in \mathcal{G}^A$ and where we put $g_f(z) = |f(z) - h(z)|^2 - |r(z) - h(z)|^2$. Actually (15) holds true in our setting, using firstly the boundedness of f, h, r and secondly the projection property on the convex set \mathcal{G}^A :

$$\begin{aligned} \text{Var}(g_f(Z)) &\leq \mathbb{E}(g_f^2(Z)) \leq (4L)^2 \mathbb{E}[(f(Z) - r(Z))^2] \\ &\leq 16L^2 \left(\mathbb{E}[(f(Z) - h(Z))^2] - \mathbb{E}[(r(Z) - h(Z))^2] \right) \\ &= 16L^2 \mathbb{E}[g_f(Z)]. \end{aligned}$$

Regarding to STEP 2, the computations are the same as in the quoted reference, where the inequality (15) is again used at [9, line 13, p. 213]. Observe that a square is missing on g_f at the sixth line of [9, p. 213], i.e. one should read $4\mathbb{E} \left[\mathcal{N}_1 \left(\frac{(\alpha+\beta)\varepsilon}{5}, \{g_f^2 : f \in \mathcal{G}^A\}, Z_1^N \right) \right] \exp \left(-\frac{3\varepsilon^2(\alpha+\beta)N}{40(16L^4)} \right)$.

Next, computations of STEPS 3, 4, 5 AND 6 are the same. As a conclusion (see [9, p. 219]), we obtain that the probability in the Theorem A.2 statement is bounded, for $n > \frac{128L^2}{\varepsilon^2(\alpha+\beta)}$, by

$$\begin{aligned} p &:= \frac{32}{7} \sup_{z_1^N} \mathcal{N}_1 \left(\frac{\varepsilon\beta}{20L}, \mathcal{G}^A, z_1^N \right) \exp \left(-\frac{\varepsilon^2(1-\varepsilon)\alpha N}{140L^2(1+\varepsilon)} \right) \\ &+ \frac{64}{7} \sup_{z_1^N} \mathcal{N}_1 \left(\frac{(\alpha+\beta)\varepsilon}{5}, \{g_f^2 : f \in \mathcal{G}^A\}, z_1^N \right) \exp \left(-\frac{3\varepsilon^2(\alpha+\beta)N}{640L^4} \right). \end{aligned}$$

At this stage, we should manage the extra exponent 2 on g_f which was missing in the quoted reference. This follows from simple arguments. Let $(f_j)_{1 \leq j \leq n}$ be a δ -covering of \mathcal{G}^A with respect to the L_1 -empirical norm associated to z_1^N and without loss of generality, assume that $|f_j(z)| \leq L$: then for any $f \in \mathcal{G}^A$, there is $j \in \{1, \dots, n\}$ such that

$$\begin{aligned} &\frac{1}{N} \sum_{i=1}^N |g_f^2(z_i) - g_{f_j}^2(z_i)| \\ &= \frac{1}{N} \sum_{i=1}^N |g_f(z_i) + g_{f_j}(z_i)| |f(z_i) + f_j(z_i) - 2h(z_i)| |f(z_i) - f_j(z_i)| \end{aligned}$$

$$\leq 32L^3 \frac{1}{N} \sum_{i=1}^N |f(z_i) - f_j(z_i)| \leq 32L^3 \delta,$$

which proves that $\mathcal{N}_1 \left(32L^3 \delta, \{g_f^2 : f \in \mathcal{G}^{\mathcal{A}}\}, z_1^N \right) \leq \mathcal{N}_1 \left(\delta, \mathcal{G}^{\mathcal{A}}, z_1^N \right)$ for any $\delta > 0$. With this inequality at hand, using $L \geq 1$ and $\varepsilon \geq 0$, we obtain

$$\begin{aligned} p &\leq \frac{32}{7} \sup_{z_1^N} \mathcal{N}_1 \left(\frac{\varepsilon\beta}{20L}, \mathcal{G}^{\mathcal{A}}, z_1^N \right) \exp \left(-\frac{\varepsilon^2(1-\varepsilon)\alpha N}{140L^4(1+\varepsilon)} \right) \\ &\quad + \frac{64}{7} \sup_{z_1^N} \mathcal{N}_1 \left(\frac{(\alpha+\beta)\varepsilon}{160L^3}, \mathcal{G}^{\mathcal{A}}, z_1^N \right) \exp \left(-\frac{3\varepsilon^2(1-\varepsilon)\alpha N}{640L^4(1+\varepsilon)} \right) \\ &\leq \frac{96}{7} \sup_{z_1^N} \mathcal{N}_1 \left(\frac{\varepsilon\beta}{160L^3}, \mathcal{G}^{\mathcal{A}}, z_1^N \right) \exp \left(-\frac{3\varepsilon^2(1-\varepsilon)\alpha N}{640L^2(1+\varepsilon)} \right) \end{aligned}$$

and this readily leads to our statement. For $N \leq \frac{128L^2}{\varepsilon^2(\alpha+\beta)}$, the announced inequality is obvious.

Bibliography

- [1] R. Aid, L. Campi, N. Langrené and H. Pham, A probabilistic numerical method for optimal multiple switching problem and application to investments in electricity generation, *Preprint available at <http://hal.archives-ouvertes.fr/ccsd-00747229/>* (2012).
- [2] C. Barrera, F. Bergeret, C. Dossal, E. Gobet, A. Meziou, R. Munos and D. Reboul, Numerical methods for the pricing of Swing options: a stochastic control approach, *Methodology and Computing in Applied Probability* **8** (2006), 517–540.
- [3] T. Ben Zineb, *Efficient numerical methods for pricing GMWB*, Ph.D. thesis, Ecole Polytechnique, 2012.
- [4] B. Bouchard and X. Warin, *Monte-Carlo valuation of American options: facts and new algorithms to improve existing methods.*, Carmona, René A. (ed.) et al., Numerical methods in finance. Selected papers based on the presentations at the workshop, Bordeaux, France, June 2010. Berlin: Springer. Springer Proceedings in Mathematics 12, 215-255, 2012.
- [5] C. Cuchiero, M. Keller-Ressel and J. Teichmann, Polynomial processes and their applications to mathematical finance, *Finance Stoch.* **16** (2012), 711–740.
- [6] D. Egloff, Monte Carlo algorithms for optimal stopping and statistical learning, *Annals of Applied Probability* **15** (2005), 1396–1432.

- [7] E. Gobet, J.P. Lemor and X. Warin, A regression-based Monte Carlo method to solve backward stochastic differential equations., *Annals of Applied Probability* **15** (2005), 2172–2202.
- [8] E. Gobet and P. Turkedjiev, Approximation of discrete BSDE using least-squares regression, *Preprint available at <http://hal.archives-ouvertes.fr/hal-00642685>* (2011).
- [9] L. Györfi, M. Kohler, A. Krzyzak and H. Walk, *A distribution-free theory of non-parametric regression*, Springer Series in Statistics, 2002.
- [10] L. Györfi and M. Wegkamp, Quantization for nonparametric regression, *IEEE Trans. Inform. Theory* **54** (2008), 867–874.
- [11] J.P. Lemor, E. Gobet and X. Warin, Rate of convergence of an empirical regression method for solving generalized backward stochastic differential equations, *Bernoulli* **12** (2006), 889–916.
- [12] F. Longstaff and E.S. Schwartz, Valuing American Options by Simulation: A Simple Least Square Approach, *The Review of Financial Studies* **14** (2001), 113–147.
- [13] A. Nemirovski, *Topics in non-parametric statistics*, Lectures on probability theory and statistics (Saint-Flour, 1998), Lecture Notes in Math. 1738, Springer, Berlin, 2000, pp. 85–277.
- [14] W. Schoutens, *Stochastic processes and orthogonal polynomials*, Lecture Notes in Statistics. 146. New York, NY: Springer, 2000.
- [15] A.B. Tsybakov, *Introduction to nonparametric estimation*, Springer Series in Statistics, Springer, New York, 2009.

Author information

Tarik Ben Zineb, Centre de Mathématiques Appliquées, Ecole Polytechnique and CNRS, 91128 Palaiseau Cedex, France.

E-mail: tarik.benzineb@gmail.com

Emmanuel Gobet, Centre de Mathématiques Appliquées, Ecole Polytechnique and CNRS, 91128 Palaiseau Cedex, France.

E-mail: emmanuel.gobet@polytechnique.edu