MCMC design-based non-parametric regression for rare-event. Application to nested risk computations

Gersende Fort, Emmanuel Gobet, Eric Moulines

MCMC design-based non-parametric regression for rare-event. Application to nested risk computations

Gersende Fort* Emmanuel Gobet† Eric Moulines‡

Abstract

We design and analyze an algorithm for estimating the mean of a function of a conditional expectation, when the outer expectation is related to a rare-event. The outer expectation is evaluated through the average along the path of an ergodic Markov chain generated by a Markov chain Monte Carlo sampler. The inner conditional expectation is computed as a non-parametric regression, using a least-squares method with a general function basis and a design given by the sampled Markov chain. We establish non asymptotic bounds for the L_2-empirical risks associated to this least-squares regression; this generalizes the error bounds usually obtained in the case of i.i.d. observations. Global error bounds are also derived for the nested expectation problem. Numerical results in the context of financial risk computations illustrate the performance of the algorithms.

KEYWORDS: empirical regression scheme, MCMC sampler, rare event

AMS CLASSIFICATION: 65C40, 62G08, 37M25

1 Introduction

Statement of the problem. We consider the problem of estimating the mean of a function of a conditional expectation in a rare-event regime, using Monte Carlo simulations. More precisely, the quantity of interest writes

$$\mathcal{I} := \mathbb{E}\left[f(Y, \mathbb{E}\left[R|Y\right])|Y \in \mathcal{A}\right] \tag{1.1}$$

where R and Y are vector-valued random variables, and $\mathcal{A}$ is a so-called rare subset, i.e. $\mathbb{P}(Y \in \mathcal{A})$ is small. This is a problem of nested Monte Carlo computations with a special

*Email: gersende.fort@telecom-paristech.fr. LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France.

†Email: emmanuel.gobet@polytechnique.edu. Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique and CNRS, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau Cedex, France. Corresponding author. The author's research is part of the Chair *Financial Risks* of the *Risk Foundation*, the *Finance for Energy Market Research Centre* and the ANR project *CAESARS* (ANR-15-CE05-0024).

‡Email: eric.moulines@polytechnique.edu. Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique and CNRS, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau Cedex, France.

emphasis on the distribution tails. In the evaluation of (1.1), which is equivalent to

$$\mathbb{E}\left[f(X, \mathbb{E}\left[R|X\right])\right]$$

where the distribution of X is the conditional distribution of Y given $\{Y \in \mathcal{A}\}$, there are two intertwined issues, which we now explain to emphasize our contributions.

The *outer Monte Carlo stage* samples distributions restricted to $\{Y \in \mathcal{A}\}$. A naive acceptance-rejection on Y fails to be efficient because most of simulations of Y are wasted. Therefore, specific rare-event techniques have to be used. Importance sampling is one of these methods (see e.g. [RK08, BL12]), which can be efficient in small dimension (10 to 100) but fails to deal with larger dimensions. In addition, this approach relies heavily on particular types of models for Y and on suitable information about the problem at hand.

Another option consists in using Markov Chain Monte Carlo (MCMC) methods. Such methods amount to construct a Markov chain $(X^{(m)})_{m\geq 0}$, such that the chain possesses an unique stationary distribution π equal to the conditional distribution of Y given the event $\{Y \in \mathcal{A}\}$. In such case, for π-almost every initial condition $X_0 = x$, the Birkhoff ergodic theorem shows that

$$\lim_{M\to+\infty} \frac{1}{M}\sum_{m=1}^{M} \varphi(X^{(m)}) = \mathbb{E}\left[\varphi(Y)|Y \in \mathcal{A}\right] \quad \text{a.s.}$$

for any (say) bounded function φ. This approach has been developed, analyzed and experimented in [GL15] in quite general and complex situations, demonstrating its efficiency over alternative methods. Therefore, a natural idea for the estimation of (1.1) would be the computation of

$$\frac{1}{M}\sum_{m=1}^{M} f\left(X^{(m)}, \mathbb{E}\left[R\,|\,X^{(m)}\right]\right),$$

emphasizing the need for approximating the quantity $\mathbb{E}\left[R\,|\,X^{(m)}\right]$.

The *inner Monte Carlo stage* is used to approximate these conditional expectations at any $X^{(m)}$ previously sampled. A first idea is to replace $\mathbb{E}\left[R\,|\,X^{(m)}\right]$ by a Crude Monte Carlo sum computed with N draws:

$$\mathbb{E}\left[R\,|\,X^{(m)}\right] \approx \frac{1}{N}\sum_{k=1}^{N} R^{(m,k)}. \tag{1.2}$$

This approach is refereed to as *nested simulation method* in [BDM15] (with the difference that their $X^{(m)}$ are i.i.d. and not given by a Markov chain). This algorithm based on (1.2) is briefly presented and studied in Appendix A. Having a large N reduces the variance of this approximation (and thus ensures convergence as proved in Theorem 5) but it yields a prohibitive computational cost. Furthermore, this naive idea does not take into account cross-informations between the different approximations at the points $\{X^{(m)}, m = 1, \cdots, M\}$. Instead, we follow a non-parametric regression approach for the approximation of the function $\phi_\star$ satisfying $\phi_\star(X) = \mathbb{E}\left[R|X\right]$ (almost-surely): given L basis functions $\phi_1, \cdots, \phi_L$, we regress $\{R^{(m)}, m = 1, \cdots, M\}$ against the variables $\{\phi_1(X^{(m)}), \cdots, \phi_L(X^{(m)}); m = 1, \cdots, M\}$ where $R^{(m)}$ is sampled from the conditional

distribution of R given $\{X = X^{(m)}\}$. Note that this inner Monte Carlo stage only requires a single draw $R^{(m)}$ for each sample $X^{(m)}$ of the outer stage. Our discussion in Subsection 2.4 shows that the regression Monte Carlo method for the inner stage outperforms the crude Monte Carlo method as soon as the regression function can be well approximated by the basis functions (which is especially true when $\phi_\star$ is smooth, with a degree of smoothness qualitatively higher than the dimension d, see details in Subsection 2.4).

The major difference with the standard setting for non-parametric regression [GKKW02] comes from the design $\{X^{(m)}, m = 1, \cdots, M\}$ which is not a i.i.d. sample: the independence fails because $\{X^{(m)}, m = 1, \cdots, M\}$ is a Markov chain path, which is ergodic but not stationary in general.

A precise description of the algorithm is given in Section 2, with a discussion on implementation issues. We also provide some error estimates, in terms of the size M of the sample, and of the function space used for approximating the inner conditional expectation. Proofs are postponed to Section 4. Section 3 gathers some numerical experiments, in the field of financial and actuarial risks. Appendix A presents the analysis of a Monte Carlo scheme for computing (1.1), by using a MCMC scheme for the outer stage and a crude Monte Carlo scheme for the inner stage.

Applications. Numerical evaluation of nested conditional expectations arises in several fields. This pops up naturally in solving dynamic programming equations for stochastic control and optimal stopping problems, see [TR01, LS01, Egl05, LGW06, BKS10]; however, coupling these latter problems with rare-event is usually not required from the problem at hand.

In financial and actuarial management [MFE05], we often retrieve nested conditional expectations, with an additional account for such estimations in the tails (like (1.1)). A major application is the risk management of portfolios written with derivative options [GJ10]: regarding (1.1), R stands for the aggregated cashflows of derivatives at time T', and Y for the underlying asset or financial variables at time $T < T'$. Then $\mathbb{E}[R|Y]$ represents the portfolio value at T given a scenario Y, and the aim is to compute the extreme exposure (Value at Risk, Conditional VaR) of the portfolio. These computations are an essential concern for Solvency Capital Requirement in insurance [DL09].

Literature background and our contributions. In view of the aforementioned applications, it is natural to find most of background results in relation to risk management in finance and insurance. Alternatively to the crude nested Monte Carlo methods (i.e. with an inner and an outer stage, both including sample Monte Carlo averages), several works have tried to speed-up the algorithms, notably by using spatial approximation of the inner conditional expectation: we refer to [HJ09] for kernel estimators, to [LS10] for kriging techniques, to [BDM15] for least-squares regression methods. However, these works do not account for the outside conditional expectation given $Y \in \mathcal{A}$, i.e. the learning design is sampled from the distribution of Y and not from the conditional distribution of Y given $\{Y \in \mathcal{A}\}$. While the latter distribution distorsion is presumably unessential in the computation of (1.1) in the case that $\mathcal{A}$ is not rare, it certainly becomes a major flaw when $\mathbb{P}(Y \in \mathcal{A}) \ll 1$ because the estimator of $\mathbb{E}[R|Y]$ is built using quite irrelevant

data. We mention that the weighted regression method of [BDM15] better accounts for extreme values of Y in the resolution of the least-squares regression, but still, the design remains sampled from the distribution of Y instead of the conditional distribution of Y given $\{Y \in \mathcal{A}\}$ and therefore most of the samples are wasted.

In this work, we use least-squares regression methods to compute the function $\phi_\star$. Our results are derived under weaker conditions than what is usually assumed: contrary to [BDM15], the basis functions $\phi_1, \cdots, \phi_L$ are not necessarily orthonormalized and the design matrix is not necessarily invertible. Therefore we allow general basis functions and we avoid conditions on the underlying distribution. Furthermore, we do not restrict our convergence analysis to $M \to \infty$ (large sample) but we also account for the approximation error (due to the function space). This allows a fine tuning of all parameters to achieve a tolerance on the global error. Finally, as a difference with the usual literature on non-parametric regression [GKKW02, Egl05], the learning sample $(X^{(m)})_{1\leq m\leq M}$ is not an i.i.d. sample of the conditional distribution of Y given $\{Y \in \mathcal{A}\}$: the error analysis is significantly modified. Among the most relevant references in the case of non i.i.d. learning sample, we refer to [BCV01, RM10, DG11]. Namely, in [BCV01], $(X^{(m)})_{1\leq m\leq M}$ is autoregressive or β-mixing: as a difference with our setting, they assume that the learning sample $(X^{(1)}, \ldots, X^{(M)})$ is stationary and that the noise sequence (i.e. $X^{(m)} - \phi_\star(X^{(m)})$, $m \geq 1$) is essentially i.i.d. (and independent of the learning sample). In [RM10], the authors relax the condition on the noise but they impose R to be bounded; the learning sample is still assumed to be stationary and β-mixing. In [DG11] the authors study kernel estimators for $\phi_\star$ (instead of least-squares like we do), under the assumption that the noise is a martingale with uniform exponential moments (we only impose finite variance).

2 Algorithm and convergence results

Let (X, R) be a $\mathbb{R}^d \times \mathbb{R}$-random vector; the distribution of X is the conditional distribution of Y given $\{Y \in \mathcal{A}\}$, with density μ w.r.t. a positive σ-finite measure λ on $\mathbb{R}^d$. For any Borel set A, we denote by $\mathsf{Q}(x, A) := \mathbb{E}\left[\mathbf{1}_A(R) \,|\, X = x\right]$; Q is a Markov kernel, it is the conditional distribution of R given X. Let $\phi_\star$ be the function from $\mathbb{R}^d$ to $\mathbb{R}$, defined by

$$\phi_\star(x) := \int_{\mathbb{R}} r \, \mathsf{Q}(x, \mathrm{d}r). \tag{2.1}$$

It satisfies, $\mu \, \mathrm{d}\lambda$-almost surely, $\phi_\star(X) = \mathbb{E}\left[R|X\right]$ when $X \sim \mu \, \mathrm{d}\lambda$.

For the regression step, choose L measurable functions $\phi_\ell : \mathbb{R}^d \to \mathbb{R}$, $\ell \in \{1, \ldots, L\}$, such that

$$\int \phi_\ell^2(x) \; \mu(x) \, \mathrm{d}\lambda(x) < \infty.$$

Denote by $\mathcal{F}$ the vector space spanned by the functions ϕ_ℓ, $\ell \in \{1, \cdots, L\}$, and by $\underline{\phi}$ the function from $\mathbb{R}^d$ to $\mathbb{R}^L$ collecting the basis functions ϕ_ℓ:

$$\underline{\phi}(x) := \begin{bmatrix} \phi_1(x) \\ \vdots \\ \phi_L(x) \end{bmatrix}.$$

By convention, vectors are column vectors. For a matrix A, A' denotes its transpose. $\langle \cdot ; \cdot \rangle$ denotes the scalar product in $\mathbb{R}^p$, and we will use $|\cdot|$ to denote both the Euclidean norm in $\mathbb{R}^p$ and the absolute value. The identity matrix of size N is denoted by I_N.

We adopt the short notation $X^{(1:M)}$ for the sequence $(X^{(1)}, \ldots, X^{(M)})$.

2.1 Algorithm

In Algorithm 1, we provide a description of a Monte Carlo approximation of the unknown quantity (1.1). Note that as a byproduct, this algorithm also provides an approximation $\widehat{\phi}_M$ of the function $\phi_\star$ given by (2.1).

Let P be a Markov transition kernel on $\mathcal{A}$ with unique invariant distribution $\mu \, \mathrm{d}\lambda$.

1 /* Simulation of the design and the observations */
2 $X^{(0)} \sim \xi$, where ξ is a distribution on $\mathcal{A}$;
3 **for** $m = 1$ **to** M **do**
4 $\quad X^{(m)} \sim \mathsf{P}(X^{(m-1)}, \mathrm{d}x)$;
5 $\quad R^{(m)} \sim \mathsf{Q}(X^{(m)}, \mathrm{d}r)$;
6 /* Least-Squares regression */
7 Choose $\hat{\alpha}_M \in \mathbb{R}^L$ solving $\arg\min_{\alpha \in \mathbb{R}^L} \frac{1}{M} \sum_{m=1}^{M} \left| R^{(m)} - \left\langle \alpha ; \underline{\phi}(X^{(m)}) \right\rangle \right|^2$ and set $\widehat{\phi}_M(x) := \left\langle \hat{\alpha}_M ; \underline{\phi}(x) \right\rangle$;
8 /* Final estimator using ergodic average */
9 **Return** $\widehat{\mathcal{I}}_M := \frac{1}{M} \sum_{m=1}^{M} f(X^{(m)}, \widehat{\phi}_M(X^{(m)}))$.

Algorithm 1: Full algorithm with M data, $M \geq L$.

The optimization problem Line 7 Algorithm 1 is equivalent to find a vector $\alpha \in \mathbb{R}^L$ solving

$$\mathbf{A}'\mathbf{A}\alpha = \mathbf{A}'\underline{\mathbf{R}} \tag{2.2}$$

where

$$\underline{\mathbf{R}} := \begin{bmatrix} R^{(1)} \\ \cdots \\ R^{(M)} \end{bmatrix}, \qquad \mathbf{A} := \begin{bmatrix} \phi_1(X^{(1)}) & \cdots & \phi_L(X^{(1)}) \\ \cdots & \cdots & \cdots \\ \phi_1(X^{(M)}) & \cdots & \phi_L(X^{(M)}) \end{bmatrix}. \tag{2.3}$$

There exists at least one solution, and the solution with minimal (Euclidean) norm is given by

$$\hat{\alpha}_M := \left(\mathbf{A}' \, \mathbf{A}\right)^{\#} \mathbf{A}'\underline{\mathbf{R}}, \tag{2.4}$$

where $(\mathbf{A}'\mathbf{A})^{\#}$ denotes the Moore-Penrose pseudo inverse matrix; $(\mathbf{A}'\mathbf{A})^{\#} = (\mathbf{A}'\mathbf{A})^{-1}$ when the rank of $\mathbf{A}$ is L, and in that case, the equation (2.2) possesses an unique solution.

An example of efficient transition kernel P is proposed in [GL15]: this kernel, hereafter denoted by P_{GL}, can be read as a Hastings Metropolis transition kernel targeting $\mu \mathrm{d}\lambda$ and with a proposal kernel with transition density q which is reversible w.r.t. μ, i.e. for all $x, z \in \mathcal{A}$,

$$\mu(x) q(x, z) = q(z, x) \mu(z) \, . \tag{2.5}$$

An algorithmic description for sampling a path of length M of a Markov chain with transition kernel P_{GL} and with initial distribution ξ is given in Algorithm 2.

1 $X^{(0)} \sim \xi$ where ξ is a distribution on $\mathcal{A}$; /* pick one point in $\mathcal{A}$ */
2 /* Simulation of data using reversible Metropolis-Hastings with rejection */
3 **for** $m = 1$ **to** M **do**
4 $\quad \widetilde{X^{(m)}} \leftarrow$ simulation according to distribution $q(X^{(m-1)}, z)\lambda(\mathrm{d}z)$;
5 $\quad$ **if** $\widetilde{X^{(m)}} \in \mathcal{A}$ **then**
6 $\quad\quad X^{(m)} \leftarrow \widetilde{X^{(m)}}$; /* acceptance */
7 $\quad$ **else**
8 $\quad\quad X^{(m)} \leftarrow X^{(m-1)}$; /* rejection */
9 **Return** $X^{(1)}, \cdots, X^{(M)}$.

Algorithm 2: MCMC for rare event: a Markov chain with kernel P_{GL}

When $\mu\,\mathrm{d}\lambda$ is a Gaussian distribution $\mathcal{N}_d(0, \Sigma)$ on $\mathbb{R}^d$ restricted to $\mathcal{A}$, $\widetilde{X} \sim \mathcal{N}_d(\rho x, (1-\rho^2)\Sigma)$ is a candidate with distribution $z \mapsto q(x, z)$ satisfying (2.5); here, $\rho \in [0, 1)$ is a design parameter chosen by the user (see [GL15, Section 4] for a discussion on the choice of ρ). Other proposal kernels q satisfying (2.5) are given in [GL15, Section 3] in the non-Gaussian case.

More generally, building a transition kernel P with invariant distribution $\mu\mathrm{d}\lambda$ is well-known using Hastings Metropolis schemes. Actually, there is no need to impose the condition (2.5) about reversibility of q w.r.t. μ. Indeed, given an arbitrary transition density $q(\cdot, \cdot)$, it is sufficient to replace Lines 5-6 of Algorithm 2 by the following acceptance rule: if $\widetilde{X^{(m)}} \in \mathcal{A}$, accept $\widetilde{X^{(m)}}$ with probability

$$\alpha_{\mathrm{accept}}(X^{(m-1)}, \widetilde{X^{(m)}}) := 1 \wedge \left[\frac{\mu(\widetilde{X^{(m)}}) q(\widetilde{X^{(m)}}, X^{(m-1)})}{\mu(X^{(m-1)}) q(X^{(m-1)}, \widetilde{X^{(m)}})} \right].$$

In the subsequent numerical tests with Gaussian distribution restricted to $\mathcal{A}$: $\mu\mathrm{d}\lambda \propto \mathcal{N}_d(0, \Sigma)\mathbf{1}_{\mathcal{A}}$, we will make use of $\widetilde{X} \sim \mathcal{N}_d(\rho x + (1-\rho)x_{\mathcal{A}}, (1-\rho^2)\Sigma)$ as a candidate for the transition density $z \mapsto q(x, z)$, where $x_{\mathcal{A}}$ is a well-chosen point in $\mathcal{A}$. In that case, we easily check that the acceptance probability is given by

$$\alpha_{\mathrm{accept}}(x, z) = 1 \wedge \exp(x'_{\mathcal{A}} \Sigma^{-1}(x - z)). \tag{2.6}$$

2.2 Convergence results for the estimation of $\phi_\star$

Let $L_2(\mu)$ be the set of measurable functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that $\int \varphi^2 \,\mu\,\mathrm{d}\lambda < \infty$; and define the norm

$$|\varphi|_{L_2(\mu)} := \left(\int \varphi^2 \mu \,\mathrm{d}\lambda \right)^{1/2}. \tag{2.7}$$

Let $\psi_\star$ be the projection of $\phi_\star$ on the linear span of the functions $\phi_1, \cdots, \phi_L$, w.r.t. the norm given by (2.7): $\psi_\star = \langle \alpha_\star; \underline{\phi} \rangle$ where $\alpha_\star \in \mathbb{R}^L$ solves

$$\left(\int \underline{\phi}\, \underline{\phi}' \mu \,\mathrm{d}\lambda \right) \alpha_\star = \int \psi\, \underline{\phi}\, \mu \,\mathrm{d}\lambda.$$

Theorem 1. *Assume that*

(i) the transition kernel P *and the initial distribution* ξ *satisfy: there exists a constant* C_{P} *and a rate sequence* $\{\rho(m), m \geq 1\}$ *such that for any* $m \geq 1$,

$$\left| \xi \mathsf{P}^m[(\psi_\star - \phi_\star)^2] - \int (\psi_\star - \phi_\star)^2 \, \mu \, \mathrm{d}\lambda \right| \leq C_{\mathsf{P}} \, \rho(m). \tag{2.8}$$

(ii) the conditional distribution Q *satisfies*

$$\sigma^2 := \sup_{x \in \mathcal{A}} \left\{ \int r^2 \, \mathsf{Q}(x, \mathrm{d}r) - \left(\int r \, \mathsf{Q}(x, \mathrm{d}r) \right)^2 \right\} < \infty. \tag{2.9}$$

Let $X^{(1:M)}$ *and* $\widehat{\phi}_M$ *be given by Algorithm 1. Then,*

$$\Delta_M := \mathbb{E}\left[\frac{1}{M} \sum_{m=1}^{M} \left(\widehat{\phi}_M(X^{(m)}) - \phi_\star(X^{(m)}) \right)^2 \right] \leq \frac{\sigma^2 L}{M} + |\psi_\star - \phi_\star|^2_{L_2(\mu)} + \frac{C_{\mathsf{P}}}{M} \sum_{m=1}^{M} \rho(m). \tag{2.10}$$

Proof. See Section 4.1. □

Δ_M measures the mean squared error $\widehat{\phi}_M - \phi_\star$ along the design sequence $X^{(1:M)}$. The proof consists in decomposing this error into a variance term and a squared bias term:

a) $\sigma^2 L/M$ in the RHS is the statistical error, decreasing as the size of the design M gets larger and increasing as the size of the approximation space L gets larger.
b) The quantity $|\psi_\star - \phi_\star|^2_{L_2(\mu)}$ is the residual error under the best approximation of $\phi_\star$ by the basis functions $\phi_1, \cdots, \phi_L$ w.r.t. the $L_2(\mu)$-norm: it is naturally expected as the limit of Δ_M when $M \to \infty$.
c) The term with $\{\rho(m), m \geq 1\}$ describes how rapidly the Markov chain $\{X^{(m)}, m \geq 1\}$ converges to its stationary distribution $\mu \mathrm{d}\lambda$.

This theorem extends known results in the case of i.i.d. design $X^{(1:M)}$, which is the major novelty of our contribution. The i.i.d. case is a special case of this general setting: it is retrieved by setting $\mathsf{P}(x, \mathrm{d}z) = \mu(z)\mathrm{d}\lambda(z)$. Note that in that case, the assumption (i) is satisfied with $C_{\mathsf{P}} = 0$, and the upper bound in (2.10) coincides with classic results (see e.g. [GKKW02, Theorem 11.1]). The theorem covers the situation when the outer Monte Carlo stage relies on a Markov chain Monte Carlo sampler; we will discuss below how to check the assumption (i) in practice.

The assumptions on the basis functions $\phi_1, \cdots, \phi_L$ are weaker than what is usually assumed in the literature on nested simulation. Namely, as a difference with [BDM15, Assumption A2] in the i.i.d. case, Theorem 1 holds even when the functions $\phi_1, \cdots, \phi_L$ are not orthonormal in $L_2(\mu)$, and it holds without assuming that almost-surely, the rank of the matrix $\mathbf{A}$ is L.

The assumption (ii) says that the conditional variance of R given X is uniformly bounded. This condition could be weakened and replaced by an ergodic condition on the Markov kernel P implying that

$$\tilde{\sigma}^2_L := \sup_{M \geq L} \mathbb{E}\left[\left| \mathbf{A}(\mathbf{A}'\mathbf{A})^{\#} \mathbf{A}' \left(\underline{\mathbf{R}} - \mathbb{E}\left[\underline{\mathbf{R}} | X^{(1:M)} \right] \right) \right|^2 \right] < \infty;$$

$\mathbf{A}$ and $\underline{\mathbf{R}}$ are given by (2.3) and depend on $X^{(1:M)}$. In that case, the upper bound (2.10) holds with $\sigma^2 L$ replaced by $\tilde{\sigma}^2_L$ (see the inequality (4.2) in the proof of Theorem 1).

We conclude this section by conditions on P and $\mathcal{A}$ implying the ergodicity assumption (2.8) with a geometric rate sequence $\rho(m) = \kappa^m$ for some $\kappa \in (0,1)$. Sufficient conditions for sub-geometric rate sequences can be found e.g. in [FM03a, DFMS04].

Proposition 2 ([MT93, Theorem 15.0.1] and [FM03b, Proposition 2])**.** *Assume that* P *is phi-irreducible and there exists a measurable function* $V : \mathcal{A} \to [1, +\infty)$ *such that*

(i) there exist $\delta \in (0,1)$ *and* $b < \infty$ *such that for any* $x \in \mathcal{A}$, $\mathsf{P}V(x) \leq \delta V(x) + b$,

(ii) there exists $\upsilon_\star \in (b/(1-\delta), +\infty)$*, such that the level set* $\mathcal{C}_\star := \{V \leq \upsilon_\star\}$ *is 1-small: there exist* $\epsilon > 0$ *and a probability distribution* ν *on* $\mathcal{A}$ *(with* $\nu(\mathcal{C}_\star) = 1$*) such that for any* $x \in \mathcal{C}_\star$, $\mathsf{P}(x, \mathrm{d}z) \geq \epsilon\nu(\mathrm{d}z)$.

Then there exist $\kappa \in (0,1)$ *and a finite constant* C_1 *such that for any measurable function* $g : \mathcal{A} \to \mathbb{R}$*, any* $m \geq 1$ *and any* $x \in \mathcal{A}$*,*

$$\left| \mathsf{P}^m g(x) - \int g \, \mu \, \mathrm{d}\lambda \right| \leq C_1 \left(\sup_{\mathcal{A}} \frac{|g|}{V} \right) \kappa^m V(x).$$

In addition, there exists a finite constant C_2 *such that for any measurable function* $g : \mathcal{A} \to \mathbb{R}$ *and any* $M \geq 1$*,*

$$\mathbb{E}\left[\left| \sum_{m=1}^{M} \{ g(X^{(m)}) - \int g \, \mu \, \mathrm{d}\lambda \} \right|^2 \right] \leq C_2 \left(\sup_{\mathcal{A}} \frac{|g|}{\sqrt{V}} \right)^2 \mathbb{E}\left[V(X^{(0)}) \right] M.$$

An explicit expression of the constant C_2 is given in [FM03b, Proposition 2]. When $\mathsf{P} = \mathsf{P}_{\mathrm{GL}}$ as described in Algorithm 2, we have the following corollary.

Corollary 3. *Assume the following conditions:*

(i) For all $x \in \mathcal{A}$*:* $\mu(z) > 0 \Longrightarrow q(x,z) > 0$.

(ii) There exists $\delta_1 \in (0,1)$ *such that* $\sup_{x \in \mathcal{A}} \int_{\mathcal{A}^c} q(x,z) \mathrm{d}\lambda(z) \leq \delta_1$.

(iii) There exist $\delta_2 \in (\delta_1, 1)$*, a measurable function* $V : \mathcal{A} \to [1, +\infty)$ *and a set* $\mathcal{B} \subset \mathcal{A}$ *such that*

$$b := \sup_{x \in \mathcal{B}} \int_{\mathcal{A}} V(z) \, q(x,z) \mathrm{d}\lambda(z) < \infty, \qquad \sup_{x \in \mathcal{B}^c} V^{-1}(x) \int_{\mathcal{A}} V(z) \, q(x,z) \mathrm{d}\lambda(z) \leq \delta_2 - \delta_1.$$

(iv) For some $\upsilon_\star > b/(1-\delta_2)$*, the level set* $\mathcal{C}_\star := \{V \leq \upsilon_\star\}$ *is such that*

$$\inf_{(x,z) \in \mathcal{C}_\star^2} \left(\frac{q(x,z) \mathbf{1}_{\mu(z) \neq 0}}{\mu(z)} \right) > 0, \qquad \int_{\mathcal{C}_\star} \mu \mathrm{d}\lambda > 0.$$

Then the assumptions of Proposition 2 are satisfied for the kernel $\mathsf{P} = \mathsf{P}_{\mathrm{GL}}$.

Proof. See Section 4.2. □

When μ is a Gaussian density $\mathcal{N}_d(0, I_d)$ on $\mathbb{R}^d$ restricted to $\mathcal{A}$ and the proposal density $q(x,y)$ is a Gaussian random variable with mean ρx and covariance $\sqrt{1-\rho^2} I_d$ (with $\rho \in (0,1)$), it is easily seen that the conditions (i), (iii) and (iv) of Corollary 3 are satisfied (choose e.g. $V(x) = \exp(s\,|x|)$, with $s > 0$). The condition (ii) is problem specific since it depends on the geometry of $\mathcal{A}$.

2.3 Convergence results for the estimation of $\mathcal{I}$

When the problem (1.1) is of the form

$$\mathbb{E}\left[f\left(Y, \mathbb{E}\left[R|Y\right]\right)|Y \in \mathcal{A}\right]$$

for a globally Lipschitz function f (in the second variable), we have the following control on the Monte Carlo error $\widehat{\mathcal{I}}_M - \mathcal{I}$ from Algorithm 1.

Theorem 4. *Assume*

(i) $f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ *is globally Lipschitz in the second variable: there exists a finite constant* C_f *such that for any* $(r_1, r_2, y) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$,

$$|f(y, r_1) - f(y, r_2)| \leq C_f \, |r_1 - r_2| \, .$$

(ii) There exists a finite constant C *such that for any* M

$$\mathbb{E}\left[\left(M^{-1}\sum_{m=1}^{M} f\left(X^{(m)}, \phi_\star(X^{(m)})\right) - \int f(x, \phi_\star(x))\mu(x)\,\mathrm{d}\lambda(x)\right)^2\right] \leq \frac{C}{M}.$$

Then

$$\left(\mathbb{E}\left[\left|\widehat{\mathcal{I}}_M - \mathcal{I}\right|^2\right]\right)^{1/2} \leq C_f\sqrt{\Delta_M} + \sqrt{\frac{C}{M}},$$

where $\mathcal{I}$, $\widehat{\mathcal{I}}_M$ *and* Δ_M *are resp. given by (1.1), Algorithm 1 and (2.10).*

Proof. See Section 4.3. □

Sufficient conditions for the assumption (ii) to hold are given in Proposition 2 when $\{X^{(m)}, m \geq 1\}$ is a Markov chain. When the draws $\{X^{(m)}, m \geq 1\}$ are i.i.d. with distribution $\mu\mathrm{d}\lambda$, the condition (ii) is verified with $C = \mathbb{V}\mathrm{ar}\left(f(X, \phi_\star(X))\right)$ with $X \sim \mu\mathrm{d}\lambda$.

2.4 Asymptotic optimal tuning of parameters

In this paragraph, we discuss how to tune the parameters of the algorithm (i.e. M, $\phi_1, \cdots, \phi_L$ and L), given a Markov kernel P. To simplify the discussion, we assume from now on that

(Hyp$_{(2.8)}$) the constant C_{P} of (2.8) can be chosen independently of $\psi_\star$; furthermore the series $(\rho(m))_{m\geq 1}$ defined in (2.8) is convergent.

The above condition on ρ is quite little demanding: see Proposition 2 where the convergence is geometric. Regarding the condition on C_{P}, although not trivial, this assumption seems reasonable since $\psi_\star$ is the best approximation of $\phi_\star$ on the function basis w.r.t. the target measure $\mu\mathrm{d}\lambda$: it means that first, $|\psi_\star - \phi_\star|_{L_2(\mu)} \leq |\phi_\star|_{L_2(\mu)}$; and second, $\psi_\star - \phi_\star$ is expected to converge to 0 in $L_2(\mu)$ as the number L of basis functions increases. Besides, in the context of Proposition 2, the control of C_{P} would follow from the control of $\sup_{\mathcal{A}} \frac{|\psi_\star - \phi_\star|}{V}$, which is a delicate task because of the lack of knowledge on $\psi_\star$.

A direct consequence of **(Hyp$_{(2.8)}$)** is that the last term in (2.10) is such that

$$\frac{C_{\mathsf{P}}}{M}\sum_{m=1}^{M}\rho(m) = O\left(\frac{1}{M}\right),$$

uniformly in the function basis. In other words, the mean empirical squared error Δ_M is bounded by $\mathrm{Cst} \times \left(\frac{L}{M} + |\psi_\star - \phi_\star|^2_{L_2(\mu)} \right)$, as in the case of i.i.d. design (see [GKKW02, Theorem 11.1]).

There are many choices of function basis [GKKW02], but due to the lack of knowledge on the target measure and in the perspective of discussing convergence rates, it is relevant to adopt local approximation techniques, like piecewise polynomial partitioning estimates (i.e. local polynomials defined on a tensored grid); for a detailed presentation, see [GT16, Section 4.4.]. Assume that the conditional expectation $\phi_\star$ is smooth on $\mathcal{A}$, namely $\phi_\star$ is p_0 continuously differentiable, with bounded derivatives, and the p_0-th derivatives is p_1-Hölder continuous. Set $p := p_0 + p_1$. If $\mathcal{A}$ is bounded, it is well-known [GKKW02, Corollary 11.1 for $d = 1$] that taking local polynomials of order p_0 on a tensored grid with edge length equal to $\mathrm{Cst} \times M^{-\frac{1}{2p+d}}$ ensures that both the statistical error L/M and the approximation error $|\psi_\star - \phi_\star|^2_{L_2(\mu)}$ have the same magnitude and we get

$$\Delta_M = O\left(M^{-\frac{2p}{2p+d}} \right). \tag{2.11}$$

If $\mathcal{A}$ is not anymore bounded, under the additional assumption that $\mu \mathrm{d}\lambda$ has tails with exponential decay, it is enough to consider similar local polynomials but on a tensored grid truncated at distance $\mathrm{Cst} \times \log(M)$; this choice maintains the validity of the estimate (2.11), up to logarithmic factors [GT16, Section 4.4.], which we omit to write for the sake of simplicity.

Regarding the complexity Cost (computational cost), the simulation cost (for $X^{(1:M)}, R^{(1:M)}$) is proportional to M, the computation of $\widehat{\phi}_M$ needs $\mathrm{Cst} \times M$ operations (taking advantage of the tensored grid), as well as the final evaluation of $\widehat{\mathcal{I}}_M$. Thus we have $\mathsf{Cost} \sim \mathrm{Cst} \times M$, with another constant. Finally, in view of Theorem 4, we derive

$$\mathsf{Error}_{\text{Regression Alg. 1}} = O\left(\mathsf{Cost}^{-\frac{p}{2p+d}} \right).$$

This is similar to the rate we would obtain in a i.i.d. setting. For very smooth $\phi_\star$ ($p \to +\infty$), we retrieve asymptotically the order $\frac{1}{2}$ of convergence.

This global error may be compared to the situation where the inner conditional expectation is computed using a crude Monte Carlo method (using N samples of $R^{(m,k)}$ for each of the M $X^{(m)}$'s); this scheme is described and analyzed in Appendix A. Its computational cost is $\mathrm{Cst} \times MN$ and its global error is $O(1/\sqrt{N} + 1/\sqrt{M})$ if f is Lipschitz (resp. $O(1/N + 1/\sqrt{M})$ if f is smoother); thus we have (by taking $M = N$ resp. $N = \sqrt{M}$)

$$\mathsf{Error}_{\text{Crude MC Alg. 3}} = O\left(\mathsf{Cost}^{-\frac{1}{4}} \right) \text{ if } f \text{ Lipschitz} \quad \left(\text{resp. } O\left(\mathsf{Cost}^{-\frac{1}{3}} \right) \text{ if } f \text{ smoother} \right).$$

In the standard case of Lipschitz f, the regression-based Algorithm 1 converges faster than Algorithm 3 under the condition $p \geq d/2$. In low dimension, this condition is easy to satisfy but it becomes problematic as the dimension increases, this is the usual curse of dimensionality.

3 Application: Put options in a rare event regime

The goal is to approximate the quantity

$$\mathcal{I} := \mathbb{E}\left[\left(\mathbb{E}\left[(K - h(S_{T'}))_+ \,|S_T\right] - p_\star\right)_+ |S_T \in \mathcal{S}\right] \tag{3.1}$$

for various choices of h, where $\{S_t, t \geq 0\}$ is a d-dimensional geometric Brownian motion, $T < T'$ and $\{S_T \in \mathcal{S}\}$ is a rare event.

3.1 A toy example in dimension 1

We start with a toy example: in dimension $d = 1$, when $h(y) = y$ and $\mathcal{S} = \{s \in \mathbb{R}_+ : s \leq s_\star\}$ so that

$$\mathcal{I} = \mathbb{E}\left[\left(\mathbb{E}\left[(K - S_{T'})_+ \,|S_T\right] - p_\star\right)_+ |S_T \leq s_\star\right].$$

$(K - S_{T'})_+$ is the Put payoff written on one stock with price $(S_t)_{t\geq 0}$, with strike K and maturity T': this is a standard financial product used by asset managers to insure their portfolio against the decrease of stock price. We take the point of view of the seller of the contract, who is mostly concerned by large values of the Put price, i.e. he aims at valuing the excess of the Put price at time $T \in (0, T')$ beyond the threshold $p_\star > 0$, for stock value S_T smaller than $s_\star > 0$. We assume that $\{S_t, t \geq 0\}$ evolves like a geometric Brownian motion, with volatility $\sigma > 0$ and zero drift. For the sake of simplicity, we assume that the interest rate is 0; extension to non-zero interest rate is obvious.

Upon noting that $S_T = \xi(Y)$ and $S_{T'} = \xi(Y)\exp(-\frac{1}{2}\sigma^2\tau + \sigma\sqrt{\tau}Z)$ where Y, Z are independent standard gaussian variables and

$$\xi(y) := S_0 \exp\left(-\frac{1}{2}\sigma^2 T + \sigma\sqrt{T}y\right), \qquad \tau := T' - T$$

we have

$$\mathcal{I} = \mathbb{E}\left[\left(\mathbb{E}\left[\left(K - \xi(Y)\exp(-\frac{1}{2}\sigma^2\tau + \sigma\sqrt{\tau}Z)\right)_+ \Big| Y\right] - p_\star\right)_+ \Big| Y \leq y_\star\right],$$

where

$$y_\star := \frac{1}{\sigma\sqrt{T}} \ln\left(\frac{s_\star}{S_0}\right) + \frac{1}{2}\sigma\sqrt{T}.$$

Therefore, the problem (3.1) is of the form (1.1) with

$$R = (K - \xi(Y)\exp(-\frac{1}{2}\sigma^2\tau + \sigma\sqrt{\tau}Z))_+, \quad f(y, r) = (r - p_\star)_+, \quad \mathcal{A} = \{y \in \mathbb{R} : y \leq y_\star\},$$

and $[Y, Z]' \sim \mathcal{N}_2(0, I_2)$. In this example, $\mathbb{P}(Y \in \mathcal{A})$ and $\mathbb{E}[R|Y]$ are explicit. We have indeed $\mathbb{P}(Y \in \mathcal{A}) = \Phi(y_\star)$ where Φ denotes the cumulative distribution function (cdf) of a standard Gaussian distribution. Furthermore, $\mathbb{E}[R|Y] = \Phi_\star(\xi(Y))$ where

$$\Phi_\star(s) := K\,\Phi(d_+(s)) - s\,\Phi(d_-(s)), \quad \text{with } d_\pm(s) := \frac{1}{\sigma\sqrt{\tau}}\ln(K/s) \pm \frac{1}{2}\sigma\sqrt{\tau};$$

note that $\phi_\star = \Phi_\star \circ \xi$. The parameter values for the numerical tests are given in Table 1.

Table 1: Parameter values for the $1d$-example

T	T'	S_0	K	σ	$s_\star$	$p_\star$
1	2	100	100	30%	30	10

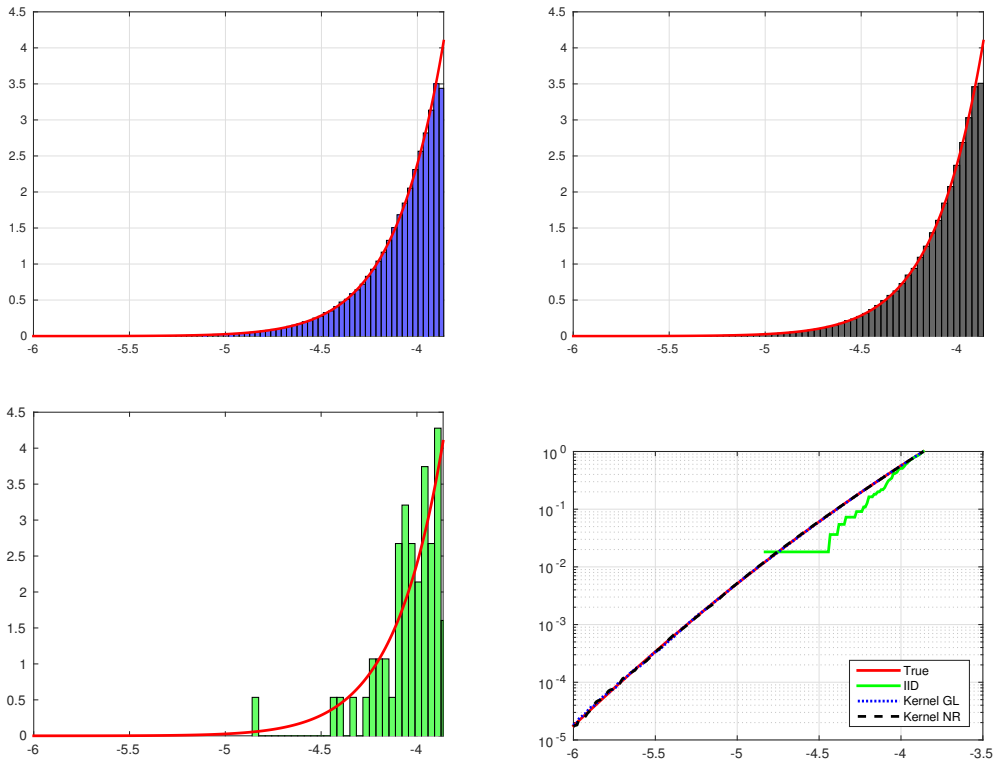

Figure 1: Normalized histograms of the M points from the Markov chains GL (top left), NR (top right) and from the i.i.d. sampler with rejection (bottom left). (bottom right) Restricted to $[-6, y_\star]$, the cdf of Y given $\{Y \in \mathcal{A}\}$, two MCMC approximations (with P_{GL} and P_{NR}) and an i.i.d. approximation.

We first illustrate the behavior of the kernel P_{GL} described by Algorithm 2. Since Y is a standard Gaussian random variable, we design P_{GL} as a Hastings-Metropolis sampler, with invariant distribution $\mu \mathrm{d}\lambda$ equal to a standard $\mathcal{N}(0,1)$ restricted to $\mathcal{A}$ and with proposal distribution $q(x,\cdot)\,\mathrm{d}\lambda \equiv \mathcal{N}(\rho x, 1-\rho^2)$. Observe that this proposal kernel is reversible w.r.t. μ, see (2.5). Note that the condition (ii) in Corollary 3 gets into

$$\sup_{y \le y_\star} \Phi\left(\frac{\rho y - y_\star}{\sqrt{1-\rho^2}}\right) < 1$$

which holds true since $\rho > 0$. In the following, the performance of the kernel P_{GL} is compared to that of the kernel P_{NR} defined as a Hastings-Metropolis kernel with proposal $q(x,\cdot)\mathrm{d}\lambda \equiv \mathcal{N}((1-\rho)y_\star + \rho x, 1-\rho^2)$ and with invariant distribution a standard Gaussian random variable restricted to $\mathcal{A}$. As a main difference with P_{GL}, this proposal transition

density q is not reversible w.r.t. μ (whence the notation P_{NR} for the kernel); therefore, the acceptance-rejection ratio of the new point z is given by (see Equality (2.6))

$$\left(1 \wedge \exp\left(y_\star\left(x - z\right)\right)\right) \mathbf{1}_{z \leq y_\star}.$$

On Figure 1(bottom right), the true cdf of Y given $\{Y \in \mathcal{A}\}$ (which is a density on $(-\infty, y_\star]$) is displayed on $[-6, y_\star]$ together with three empirical cdfs $x \mapsto M^{-1} \sum_{m=1}^{M} \mathbf{1}_{\{X^{(m)} \leq x\}}$: the first one is computed from i.i.d. samples with distribution $\mathcal{N}(0,1)$ and the second one (resp. the third one) is computed from a Markov chain path $X^{(1:M)}$ of length M with kernel P_{GL} (resp. P_{NR}) and started at $X^{(0)} = y_\star$. The two kernels provide a similar approximation of the true cdf. Here $M = 1e6$, and $\rho = 0.85$ for both kernels. We also display the normalized histograms of the points $X^{(m)}$ sampled respectively from P_{GL} (top left), P_{NR} (top right) and the crude rejection algorithm with Gaussian proposal (bottom left). In the latter plot, the histogram is built with only around 50-60 points which correspond to the accepted points among $M = 1e6$ proposal points.

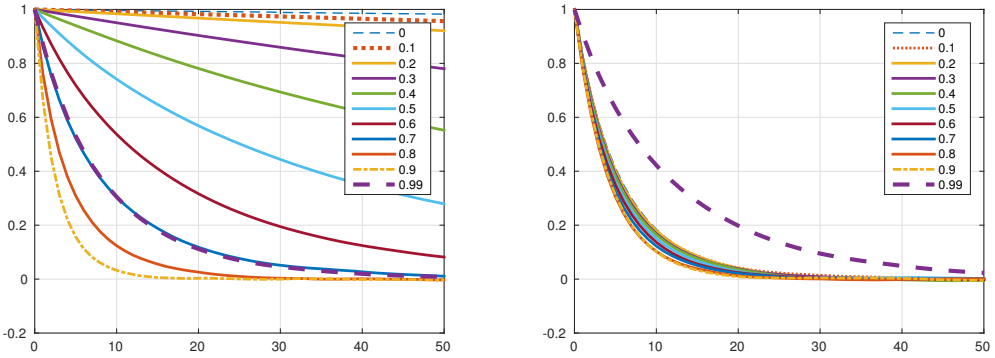

Figure 2: For different values of ρ, estimation of the autocorrelation function (over 100 independent runs) of the chain P_{GL} (left) and P_{NR} (right). Each curve is computed using $1e6$ sampled points.

To assess the speed of convergence of the samplers P_{GL} and P_{NR} to their stationary distributions, we additionally plot in Figure 2 the autocorrelation function for both chains. For P_{GL} the choice of ρ is quite significant, as observed in [GL15]; values of ρ around 0.9 give usually good results. For P_{NR}, in this example the choice of ρ is less significant because we are able to define a proposal which takes advantage of the knowledge on the rare set. A comparison of acceptance rates is provided below (see Figure 3(left)).

We also illustrate the behavior of these two MCMC samplers for the estimation of the rare event probability $\mathbb{P}(Y \in \mathcal{A})$. Following the approach of [GL15], we use the decomposition

$$\mathbb{P}(Y \in \mathcal{A}) = \prod_{j=1}^{J} \mathbb{P}\left(Y \leq w_j | Y \leq w_{j-1}\right) \approx \hat{\pi} := \prod_{j=1}^{J} \left(\frac{1}{M} \sum_{m=1}^{M} \mathbf{1}_{X^{(m,j)} \leq w_j} \right)$$

where $w_0 = +\infty > w_1 > \cdots > w_J = y_\star$, and $\{X^{(m,j)}, m \geq 0\}$ is a Markov chain with kernel $\mathsf{P}_{\mathrm{GL}}^{(j)}$ or $\mathsf{P}_{\mathrm{NR}}^{(j)}$ having a standard Gaussian restricted to $(-\infty, w_{j-1}]$ as invariant distribution. The J intermediate levels are chosen such that $\mathbb{P}(Y \leq w_j | Y \leq w_{j-1}) \approx 0.1$.

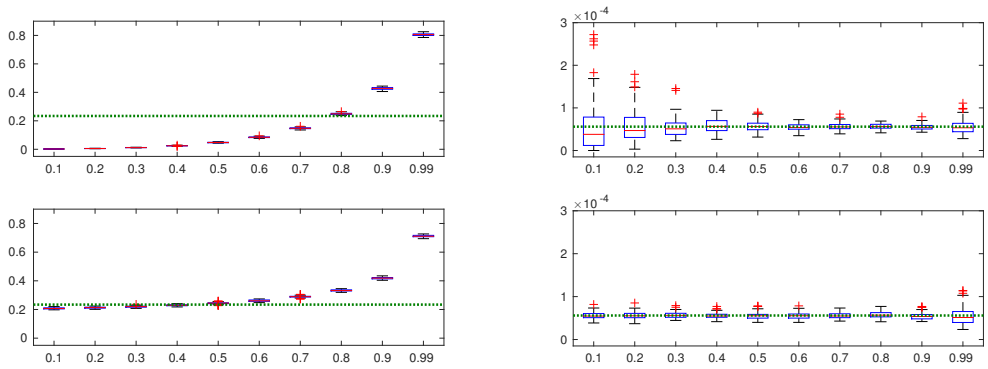

Figure 3: Comparison of the MCMC sampler P_{GL} (top) and P_{NR} (bottom), for different values of $\rho \in \{0.1, \cdots, 0.9, 0.99\}$. (left) Mean acceptance rate when computing $\mathbb{P}(Y \leq y_\star | Y \leq w_{J-1})$ after M iterations of the chain; (right) Estimation of $\mathbb{P}(Y \in \mathcal{A})$ by combining splitting and MCMC.

Figure 3(right) displays the boxplot of 100 independent realizations of the estimator $\hat{\pi}$ for different values of $\rho \in \{0.1, \cdots, 0.9\}$; the horizontal dotted line indicates the true value $\mathbb{P}(Y \in \mathcal{A}) = 5.6e{-}5$. Here $J = 5$, $(w_1, \cdots, w_4) = (0, -1.6, -2.5, -3.2)$ and $M = 1e4$. Figure 3(left) displays the boxplot of 100 mean acceptance rates $M^{-1} \sum_{m=1}^{M} \mathbf{1}_{\{X^{(m,J)} = \widetilde{X}^{(m,J)}\}}$ computed along 100 independent chains $\{X^{(m,J)}, m \leq M\}$, for different values of ρ; the horizontal dotted line is set to 0.234 which is usually chosen as the target rate when fixing some design parameters in a Hastings-Metropolis algorithm (see e.g. [Ros08]). We observe that the use of non-reversible proposal kernel P_{NR} yields more accurate results than P_{GL}; this is intuitively easy to understand since P_{GL} better accounts for the point $y_\star$ around which one should sample.

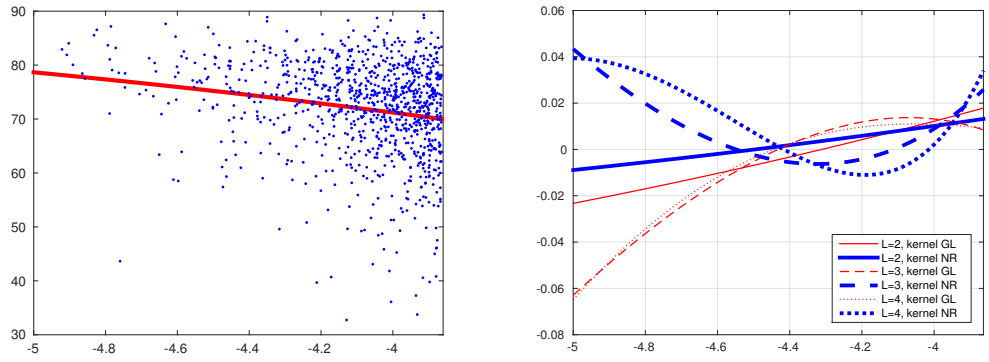

Figure 4: (left) 1000 sampled points $(X^{(m)}, R^{(m)})$ (using the sampler P_{GL}), together with $\phi_\star$; (right) A realization of the error function $x \mapsto \widehat{\phi}_M(x) - \phi_\star(x)$ on $[-5, y_\star]$, for different values of $L \in \{2, 3, 4\}$ and two different kernels when sampling $X^{(1:M)}$.

We now run Algorithm 1 for the estimation of the conditional expectation $x \mapsto \phi_\star(x)$ on $(-\infty, y_\star]$. The algorithm is run with $M = 1e6$, successively with $\mathsf{P} = \mathsf{P}_{\mathrm{GL}}$ and $\mathsf{P} = \mathsf{P}_{\mathrm{NR}}$

both with $\rho = 0.85$; the L basis functions are $\{x \mapsto \phi_\ell(x) = (\xi(x))^{\ell-1}, l = 1, \ldots, L\}$ and we consider successively $L \in \{2, 3, 4\}$. On Figure 4(right), the error function $x \mapsto \widehat{\phi}_M(x) - \phi_\star(x)$ is displayed for different values of L when computing $\widehat{\phi}_M$. It is displayed on the interval $[-5, y_\star]$, which is an interval with probability larger than $1 - 5e{-}3$ under the distribution of Y given $\{Y \in \mathcal{A}\}$ (see Figure 1). Note that the errors may be quite large for x close to -5; however these values are very unlikely (see Figure 1), and therefore these large errors are not representative of the global quadratic error. On Figure 4(left), we display 1000 sampled points of $(X^{(m)}, R^{(m)})$. These points are taken from the sampler P_{GL}, every 20 iterations, in order to obtain quite uncorrelated design points. Observe that the regression function $\phi_\star$ looks like affine, which explains why the results with $L = 2$ only are quite accurate.

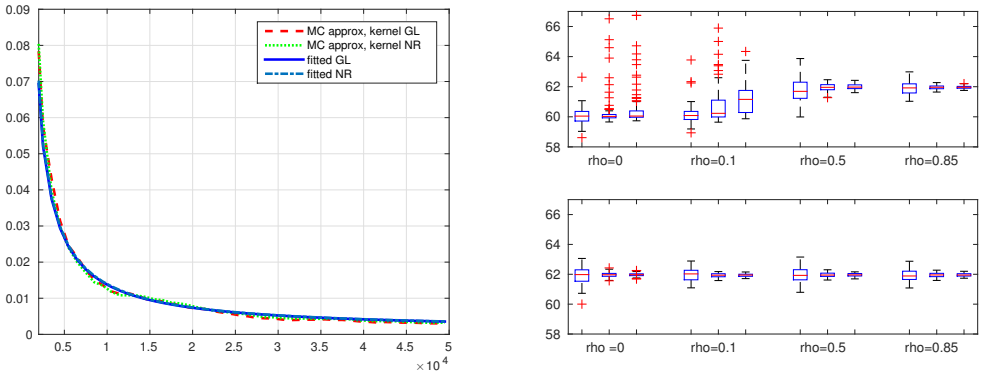

Figure 5: (left) Monte Carlo approximations of $M \mapsto \Delta_M$, and fitted curves of the form $M \mapsto \alpha + \beta/M$. (right) For different values of ρ, and for three different values of M, boxplot of 100 independent estimates $\widehat{\mathcal{I}}_M$ when $X^{(1:M)}$ is sampled from a chain with kernel P_{GL} (top) and P_{NR} (bottom).

We finally illustrate Algorithm 1 for the estimation of $\mathcal{I}$ (see (3.1)). On Figure 5(right), the boxplot of 100 independent outputs $\widehat{\mathcal{I}}_M$ of Algorithm 1 is displayed when run with $\mathsf{P} = \mathsf{P}_{\mathrm{GL}}$ (top) and $\mathsf{P} = \mathsf{P}_{\mathrm{NR}}$ (bottom); different values of ρ and M are considered, namely $\rho \in \{0, 0.1, 0.5, 0.85\}$ and $M \in \{5e2, 5e3, 1e4\}$; the regression step is performed with $L = 2$ basis functions. Figure 5(right) illustrates well the benefit of using MCMC sampler for the current regression problems: when $\mathsf{P} = \mathsf{P}_{\mathrm{GL}}$, compare the distribution for $\rho = 0$ (i.i.d. samples) and $\rho = 0.85$: observe the bias when $\rho = 0$ which does not disappear even when $M = 1e4$ and note that the variance is very significantly reduced (when $M = 5e2, 5e3, 1e4$ respectively, the standard deviation is reduced by a factor 1.11, 6.58 and 11.96).

Figure 5(left) is an empirical verification of the statement of Theorem 1. 100 independent runs of Algorithm 1 are performed, and for different values of M, the quantities $M^{-1} \sum_{m=1}^{M} \left(\widehat{\phi}_M(X^{(m)}) - \phi_\star(X^{(m)}) \right)^2$ are collected; here $\widehat{\phi}_M$ is computed with $L = 2$ basis functions. The mean value over these 100 points is displayed as a function of M; it is a Monte Carlo approximation of Δ_M (see (2.10)). We compare two implementations of Algorithm 1: first, $\mathsf{P} = \mathsf{P}_{\mathrm{GL}}$ with $\rho = 0.85$ and then $\mathsf{P} = \mathsf{P}_{\mathrm{NR}}$ with $\rho = 0.85$. Theorem 1 establishes that Δ_M is upper bounded by a quantity of the form $\alpha + \beta/M$; such a curve is fitted by a mean square technique (we obtain $\alpha = 0.001$ for both kernels, which is in

adequation with the theorem since this term does not depend on the Monte Carlo stages). The fitted curves are shown on Figure 5(left) and they demonstrate a good match between the theory and the numerical studies.

3.2 Correlated geometric Brownian motions in dimension 2

We adapt the one-dimensional example, taking a Put on the geometric average of two correlated assets $\{S_t = (S_{t,1}, S_{t,2}), t \geq 0\}$. In this example, $d = 2$, $h(s_1, s_2) = \sqrt{s_1 s_2}$ and $\mathcal{S} = \{(s_1, s_2) \in \mathbb{R}_+ \times \mathbb{R}_+ : s_1 \leq s_\star, s_2 \leq s_\star\}$. We denote by σ_1, σ_2 and ϱ, respectively each volatility and the correlation; the drift of $\{S_t, t \geq 0\}$ is zero. Set

$$\Gamma := \begin{bmatrix} 1 & \varrho \\ \varrho & 1 \end{bmatrix}, \qquad \xi(y_1, y_2) := \begin{bmatrix} S_{0,1} \exp\left(-\frac{1}{2}\sigma_1^2 T + \sqrt{T}\sigma_1 y_1\right) \\ S_{0,2} \exp\left(-\frac{1}{2}\sigma_2^2 T + \sqrt{T}\sigma_2 y_2\right) \end{bmatrix}.$$

We have $S_T = \xi(Y)$ where $Y \sim \mathcal{N}_2(0, \Gamma)$. Furthermore, it is easy to verify that $\{\sqrt{S_{t,1} S_{t,t}}, t \geq 0\}$ is still a geometric Brownian motion, with volatility σ' and drift μ' given by

$$\sigma' := \frac{1}{2}\sqrt{\sigma_1^2 + \sigma_2^2 + 2\varrho\sigma_1\sigma_2}, \qquad \mu' := -\frac{1}{8}(\sigma_1^2 + \sigma_2^2 - 2\varrho\sigma_1\sigma_2).$$

Hence, the problem (3.1) is of the form (1.1) with

$$\begin{aligned} f(y, r) &:= (r - p_\star)_+, \\ \mathcal{A} &:= \{y \in \mathbb{R}^2 : \xi(y) \in (-\infty, s_\star] \times (-\infty, s_\star]\} \\ &= \{y \in \mathbb{R}^2, y_i \leq y_{\star,i}\}, \qquad \text{where } y_\star := \left[\frac{1}{\sigma_i\sqrt{T}} \ln(s_\star / S_{0,i}) + \frac{1}{2}\sigma_i\sqrt{T}\right]_{i=1,2}, \\ R &:= \left(K - \Psi(Y) \ \exp\left\{(\mu' - \frac{1}{2}(\sigma')^2)(T' - T) + \sqrt{T' - T}\sigma' Z\right\}\right)_+, \end{aligned}$$

where $Z \sim \mathcal{N}(0, 1)$ is independent of Y, and $\Psi(y) := \sqrt{(\xi(y))_1 \, (\xi(y))_2}$.

For the outer Monte Carlo stage, P_{GL} is defined as the Hastings-Metropolis kernel with proposal distribution $q(x, \cdot)\mathrm{d}\lambda \equiv \mathcal{N}_2(\rho x, (1 - \rho^2)\Gamma)$ (with $\rho \in (0, 1)$) and with invariant distribution, a bi-dimensional Gaussian distribution $\mathcal{N}_2(0, \Gamma)$ restricted to the set $\mathcal{A}$. We compare this Markov kernel to the kernel P_{NR} with non reversible proposal, defined as a Hastings-Metropolis with proposal distribution $\mathcal{N}_2(\rho x + (1 - \rho)y_\star, (1 - \rho^2)\Gamma)$ and with invariant distribution, a bi-dimensional Gaussian distribution $\mathcal{N}_2(0, \Gamma)$ restricted to the set $\mathcal{A}$. The acceptance-rejection ratio for this algorithm is given by (2.6) with $x_{\mathcal{A}} \leftarrow y_\star$ and $\Sigma \leftarrow \Gamma$.

In this example, the inner conditional expectation is explicit: $\phi_\star(x) = \Phi_\star(\Psi(x))$ with

$$\begin{aligned} \Phi_\star(u) &:= K \ \Phi(d_+(u)) - u \, e^{\mu'(T' - T)} \ \Phi(d_-(u)), \qquad u > 0, \\ d_\pm(u) &:= \frac{1}{\sigma'\sqrt{T' - T}} \ln(K / (u \, e^{\mu'(T' - T)})) \pm \frac{1}{2}\sigma'\sqrt{T' - T}. \end{aligned}$$

For the basis functions, we take

$$\begin{aligned} &\varphi_1(x) = 1, \qquad \varphi_2(x) = \sqrt{(\xi(x))_1}, \qquad \varphi_3(x) = \sqrt{(\xi(x))_2}, \\ &\varphi_4(x) = (\xi(x))_1, \quad \varphi_5(x) = (\xi(x))_2, \qquad \varphi_6(x) = \sqrt{(\xi(x))_1 \, (\xi(x))_2}. \end{aligned}$$

Table 2: Parameter values for the $2d$-example

T	T'	$S_{0,1}$	$S_{0,2}$	K	σ_1	σ_2	ϱ	$s_\star$	$p_\star$
1	2	100	100	100	25%	35%	50%	50	5

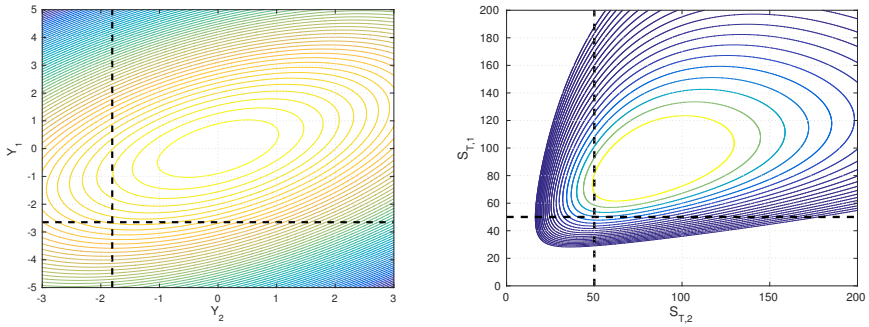

Figure 6: (left) Level curves of $\mathcal{N}_2(0,\Gamma)$ and the rare set in the lower left area delimited by the two hyperplanes. (right) Level curves of the density function of $(S_{T,1}, S_{T,2})$ and the rare set in the lower left area delimited by the two hyperplanes.

The parameter values for the numerical tests are given in Table 2.

Figure 6 depicts the rare event $\mathcal{A}$: on the left (resp. on the right), some level curves of the distribution of $\mathcal{N}_2(0,\Gamma)$ (resp. distribution of $(S_{T,1}, S_{T,2})$) are displayed, together with the rare event in the bottom left corner.

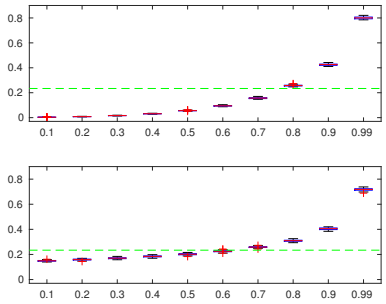

Figure 7: Boxplot over 100 independent runs, of the mean acceptance rate after $M = 1e4$ iterations for the kernel $\mathsf{P} = \mathsf{P}_{\mathrm{GL}}$ (top) and the kernel $\mathsf{P} = \mathsf{P}_{\mathrm{NR}}$ (bottom). Different values of ρ are considered.

We run two Markov chains resp. with kernel P_{GL} and P_{NR} and compute the mean acceptance-rejection rate after $M = 1e4$ iterations. For different values of ρ, this experiment is repeated 100 times, independently; Figure 7 reports the boxplot of these mean acceptance rates. It shows that a rate close to 0.234 is reached with $\rho = 0.8$ for $\mathsf{P} = \mathsf{P}_{\mathrm{GL}}$ and $\rho = 0.7$ for $\mathsf{P} = \mathsf{P}_{\mathrm{NR}}$. In all the experiments below involving these kernels, we will use these values of the design parameter ρ.

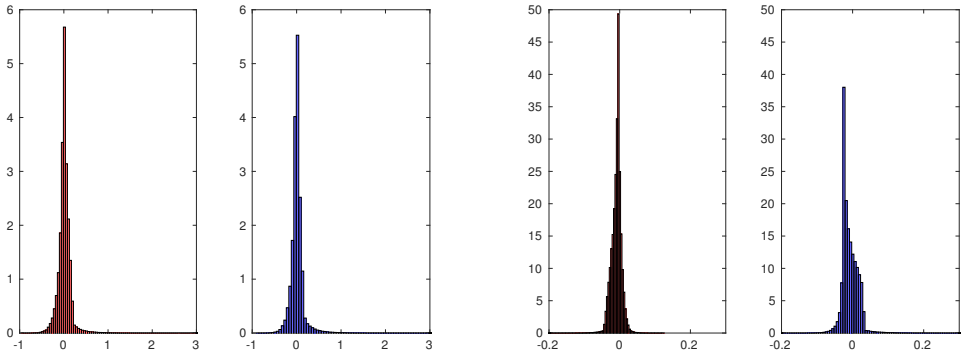

Figure 8: (left) Normalized histograms of the error $\{\widehat{\phi}_M(X^{(m)}) - \phi_\star(X^{(m)}), m = 1, \cdots, M\}$, when $L = 3$, with design points sampled with P_{GL} (left) and P_{NR} (right). (right): the same case with $L = 6$.

On Figure 8(left), the normalized histogram of the errors $\{\widehat{\phi}_M(X^{(m)}) - \phi_\star(X^{(m)}), m = 1, \cdots, M\}$ is displayed when $L = 3$ and the samples $X^{(1:M)}$ are sampled from $\mathsf{P} = \mathsf{P}_{\mathrm{GL}}$ (left) or $\mathsf{P} = \mathsf{P}_{\mathrm{NR}}$ (right). Figure 8(right) shows the case $L = 6$. Here, $M = 1e6$. This clearly shows an improvement by choosing more basis functions. Especially, the $6th$ basis function brings much accuracy, as expected, since the regression function $\phi_\star$ depends directly on it.

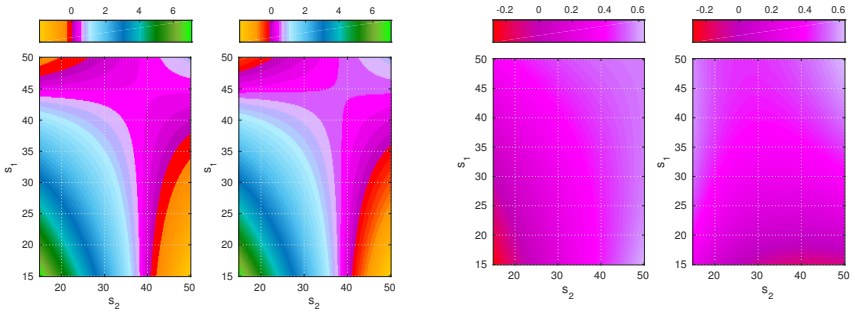

Figure 9: (left) Error function $s \mapsto \widehat{\phi}_M(\xi^{-1}(s)) - \phi_\star(\xi^{-1}(s))$, with $L = 3$, with design points sampled with P_{GL} (left) and P_{NR} (right). (right): the same case with $L = 6$.

On Figure 9(left), the errors $s \mapsto \widehat{\phi}_M(\xi^{-1}(s)) - \phi_\star(\xi^{-1}(s))$ are displayed on $[15, s_\star] \times [15, s_\star]$ when $L = 3$ and the outer samples $X^{(1:M)}$ used in the computation of $\widehat{\phi}_M$ are sampled from $\mathsf{P} = \mathsf{P}_{\mathrm{GL}}$ (left) and $\mathsf{P} = \mathsf{P}_{\mathrm{NR}}$ (right). Figure 9(right) shows the case $L = 6$. Here, $M = 1e6$. This is complementary to Figure 8 since it shows the prediction error everywhere in the space, and not only along the design points.

On Figure 10(left), a Monte Carlo approximation of Δ_M (see (2.10)) computed from 100 independent estimators $\widehat{\phi}_M$ is displayed as a function of M for M in the range $[3e3, 5e4]$; where $\widehat{\phi}_M$ is computed with $L = 6$. We also fit a curve of the form $M \mapsto \alpha + \beta/M$ to illustrate the sharpness of the upper bound in (2.10). On Figure 10(right), the boxplot

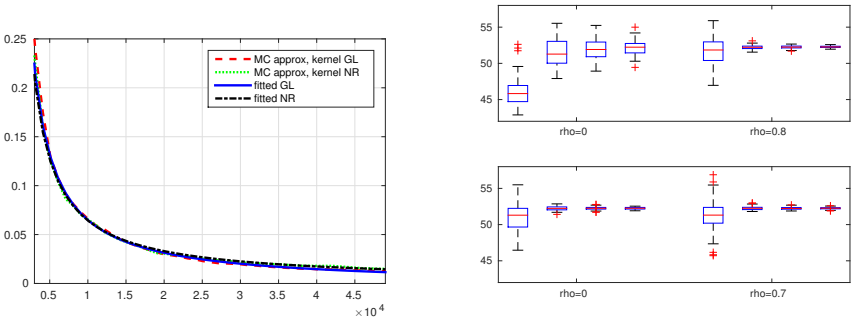

Figure 10: (left) A Monte Carlo approximation of $M \mapsto \Delta_M$, and a fitted curve of the form $M \mapsto \alpha + \beta/M$. (right) For different values of ρ_{GL} and ρ_{NR}, and for four different values of M -namely $M \in \{1e2, 5e3, 1e4, 2e4\}$-, boxplot of 100 independent estimates $\widehat{\mathcal{I}}_M$.

of 100 independent outputs $\widehat{\mathcal{I}}_M$ of Algorithm 1 is displayed, for $M \in \{1e2, 5e3, 1e4, 2e4\}$ and different values of ρ_{GL} (resp ρ_{NR}) - the design parameter in P_{GL} (resp. P_{NR}). Here again, we observe the advantage of using MCMC samplers to reduce the variance in this regression problem coupled with rare event regime: when $M = 5e3, 1e4, 2e4$ respectively, the standard deviation is reduced by a factor 6.89, 7.27 and 7.74.

4 Proofs of the results of Section 2.2

4.1 Proof of Theorem 1

By construction of the random variables $\underline{\mathbf{R}}$ and $X^{(1:M)}$ (see Algorithm 1), for any bounded and positive measurable functions $g_1, \cdots, g_M$, it holds

$$\mathbb{E}\left[\prod_{m=1}^{M} g_m(R^{(m)})|X^{(1:M)}\right] = \prod_{m=1}^{M} \mathbb{E}\left[g_m(R^{(m)})|X^{(m)}\right] = \prod_{m=1}^{M} \int g_m(r)\, \mathsf{Q}(X^{(m)}, \mathrm{d}r). \quad (4.1)$$

Lemma 1. *If $\mathbf{A}'\mathbf{A}\alpha = \mathbf{A}'\mathbf{A}\tilde{\alpha}$ then $\mathbf{A}\alpha = \mathbf{A}\tilde{\alpha}$. In other words, any coefficient solution α of the least-squares problem (2.2) yields the same values for the approximated regression function along the design $X^{(1:M)}$.*

Proof. Denote by r the rank of $\mathbf{A}$ and write $\mathbf{A} = UDV'$ for the singular value decomposition of $\mathbf{A}$. It holds

$$\mathbf{A}'\mathbf{A}\alpha = \mathbf{A}'\mathbf{A}\tilde{\alpha} \Longleftrightarrow D'DV'\alpha = D'DV'\tilde{\alpha}$$

by using $V'V = I_L$ and $U'U = I_M$. This implies that the first r components of $V'\alpha$ and $V'\tilde{\alpha}$ are equal and thus $DV'\alpha = DV'\tilde{\alpha}$. This concludes the proof. □

The next result justifies a possible interchange between least-squares projection and conditional expectation.

Lemma 2. *Set* $\widehat{\phi}_M = \langle \hat{\alpha}_M; \underline{\phi} \rangle$ *where* $\hat{\alpha}_M \in \mathbb{R}^L$ *is any solution to* $\mathbf{A}'\mathbf{A}\alpha = \mathbf{A}'\underline{\mathbf{R}}$. *Then the function* $x \mapsto \mathbb{E}[\widehat{\phi}_M(x)|X^{(1:M)}]$ *is a solution to the least-squares problem*

$$\min_{\varphi \in \mathcal{F}} \frac{1}{M} \sum_{m=1}^{M} \left(\phi_\star(X^{(m)}) - \varphi(X^{(m)}) \right)^2,$$

where $\mathcal{F} := \{\varphi = \langle \alpha; \underline{\phi} \rangle, \alpha \in \mathbb{R}^L\}$.

Proof. It is sufficient to prove that

$$\min_{\varphi \in \mathcal{F}} \frac{1}{M} \sum_{m=1}^{M} \left(\phi_\star(X^{(m)}) - \varphi(X^{(m)}) \right)^2 = \frac{1}{M} \left| \underline{\phi}_\star - \mathbf{A}\mathbb{E}\left[\hat{\alpha}_M | X^{(1:M)} \right] \right|^2$$

where $\underline{\phi}_\star := \left(\phi_\star(X^{(1)}), \cdots, \phi_\star(X^{(M)}) \right)'$. The solution of the above least-squares problem is of the form $x \mapsto \langle \hat{\alpha}_{M,\star}; \underline{\phi}(x) \rangle$ where $\hat{\alpha}_{M,\star} \in \mathbb{R}^L$ satisfies $\mathbf{A}'\mathbf{A}\,\hat{\alpha}_{M,\star} = \mathbf{A}'\underline{\phi}_\star$. By (4.1) and the definition of $\hat{\alpha}_M$, this yields

$$\mathbf{A}'\underline{\phi}_\star = \mathbf{A}' \begin{bmatrix} \mathbb{E}\left[R^{(1)} | X^{(1)} \right] \\ \cdots \\ \mathbb{E}\left[R^{(M)} | X^{(M)} \right] \end{bmatrix} = \mathbb{E}\left[\mathbf{A}'\underline{\mathbf{R}} | X^{(1:M)} \right] = \mathbf{A}'\mathbf{A}\mathbb{E}\left[\hat{\alpha}_M | X^{(1:M)} \right].$$

We then conclude by Lemma 1 that $\mathbf{A}\hat{\alpha}_{M,\star} = \mathbf{A}\mathbb{E}\left[\hat{\alpha}_M | X^{(1:M)} \right]$. We are done. □

Proof. (of Theorem 1) Using Lemma 2 and the Pythagoras theorem, it holds

$$\frac{1}{M} \sum_{m=1}^{M} \left(\widehat{\phi}_M(X^{(m)}) - \phi_\star(X^{(m)}) \right)^2 = \mathcal{T}_1 + \mathcal{T}_2$$

with

$$\mathcal{T}_1 := \frac{1}{M} \sum_{m=1}^{M} \left(\widehat{\phi}_M(X^{(m)}) - \mathbb{E}\left[\widehat{\phi}_M(X^{(m)}) | X^{(1:M)} \right] \right)^2 = \frac{1}{M} \left| \mathbf{A} \left(\hat{\alpha}_M - \mathbb{E}\left[\hat{\alpha}_M | X^{(1:M)} \right] \right) \right|^2,$$

$$\mathcal{T}_2 := \frac{1}{M} \sum_{m=1}^{M} \left(\mathbb{E}\left[\widehat{\phi}_M(X^{(m)}) | X^{(1:M)} \right] - \phi_\star(X^{(m)}) \right)^2.$$

By Lemma 1, we can take $\hat{\alpha}_M = (\mathbf{A}'\mathbf{A})^{\#}\mathbf{A}'\underline{\mathbf{R}}$, which is the coefficient with minimal norm among the solutions of least-squares problem of Algorithm 1. Let us consider $\mathcal{T}_1$. Set $\mathbf{B} := \mathbf{A}\left(\mathbf{A}'\mathbf{A}\right)^{\#}\mathbf{A}'$, a $M \times M$ matrix. By (2.4) and Lemma 2

$$M\mathcal{T}_1 = |\mathbf{B}\Upsilon|^2 = \mathrm{Trace}(\mathbf{B}\Upsilon\Upsilon'\mathbf{B}), \qquad \text{with } \Upsilon := \begin{bmatrix} R^{(1)} - \phi_\star(X^{(1)}) \\ \cdots \\ R^{(M)} - \phi_\star(X^{(M)}) \end{bmatrix}$$

so $M\,\mathbb{E}\left[\mathcal{T}_1 | X^{(1:M)} \right]$ is equal to $\mathrm{Trace}\left(\mathbf{B}\,\mathbb{E}\left[\Upsilon\Upsilon' \middle| X^{(1:M)} \right] \mathbf{B} \right)$. Under (4.1) and (2.9), the matrix $\mathbb{E}\left[\Upsilon\Upsilon' \middle| X^{(1:M)} \right]$ is diagonal with diagonal entries upper bounded by σ^2. Therefore,

$$M\mathbb{E}\left[\mathcal{T}_1 | X^{(1:M)} \right] \leq \sigma^2 \mathrm{Trace}\left(\mathbf{B}^2 \right) = \sigma^2 \mathrm{rank}(\mathbf{A}) \leq \sigma^2 L. \tag{4.2}$$

This concludes the control of $\mathbb{E}[\mathcal{T}_1]$.
Using again Lemma 2,

$$\mathcal{T}_2 = \min_{\varphi\in\mathcal{F}} \frac{1}{M}\sum_{m=1}^{M}\left(\varphi(X^{(m)}) - \phi_\star(X^{(m)})\right)^2 \le \frac{1}{M}\sum_{m=1}^{M}\left(\psi_\star(X^{(m)}) - \phi_\star(X^{(m)})\right)^2.$$

Hence,

$$\begin{aligned}\mathbb{E}[\mathcal{T}_2] &\le \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left(\psi_\star(X^{(m)}) - \phi_\star(X^{(m)})\right)^2\right] \\ &= |\psi_\star - \phi_\star|^2_{L_2(\mu)} + \frac{1}{M}\sum_{m=1}^{M}\left\{\mathbb{E}\left[\left(\psi_\star(X^{(m)}) - \phi_\star(X^{(m)})\right)^2\right] - \int(\psi_\star - \phi_\star)^2\,\mu\,\mathrm{d}\lambda\right\}.\end{aligned}$$

By (2.8), the RHS is upper bounded by $|\psi - \phi_\star|^2_{L_2(\mu)} + C_{\mathsf{P}}\sum_{m=1}^{M}\rho(m)/M$. This concludes the proof of (2.10). □

4.2 Proof of Corollary 3

P_{GL} is a Hastings-Metropolis kernel; hence, for any $x \in \mathcal{A}$ and any measurable set A in $\mathcal{A}$,

$$\mathsf{P}_{\mathrm{GL}}(x, A) = \int_{A\cap\mathcal{A}} q(x,z)\mathrm{d}\lambda(z) + \delta_x(A)\int_{\mathcal{A}^c} q(x,z)\mathrm{d}\lambda(z). \tag{4.3}$$

Irreducibility. Let A be a measurable subset of $\mathcal{A}$ such that $\int_A \mu\mathrm{d}\lambda > 0$ (which implies that $\int_A \mathrm{d}\lambda > 0$). Then,

$$\mathsf{P}_{\mathrm{GL}}(x, A) \ge \int_{A\cap\mathcal{A}\cap\{z:\mu(z)>0\}} q(x,z)\mathrm{d}\lambda(z)$$

and the RHS is positive since, owing to assumption (i), $q(x,z) > 0$ for all $x \in \mathcal{A}, z \in A\cap\mathcal{A}\cap\{z : \mu(z) > 0\}$. This implies that P_{GL} is phi-irreducible with $\mu\,\mathrm{d}\lambda$ as irreducibility measure.

Drift assumption. By assumption (ii) and from (4.3), we have

$$\mathsf{P}_{\mathrm{GL}}(x, A) \le \delta_1\,\delta_x(A) + \int_{A\cap\mathcal{A}} q(x,z)\mathrm{d}\lambda(z),$$

which implies by (iii) that

$$\begin{aligned}\mathsf{P}_{\mathrm{GL}}V(x) &\le \delta_1 V(x) + \int_{\mathcal{A}} V(z)\,q(x,z)\mathrm{d}\lambda(z) \\ &\le \delta_1 V(x) + \mathbf{1}_{\mathcal{B}}(x)\sup_{x\in\mathcal{B}}\int_{\mathcal{A}} V(z)\,q(x,z)\mathrm{d}\lambda(z) + \mathbf{1}_{\mathcal{B}^c}(x)(\delta_2 - \delta_1)V(x) \\ &\le \delta_2 V(x) + \sup_{x\in\mathcal{B}}\int_{\mathcal{A}} V(z)\,q(x,z)\mathrm{d}\lambda(z).\end{aligned}$$

Small set assumption. Let $\mathcal{C}_\star$ be given by the assumption (iv). We have $\int_{\mathcal{C}_\star} \mu \mathrm{d}\lambda > 0$; thus define the probability measure $\mathrm{d}\nu := \mathbf{1}_{\mathcal{C}_\star}\mu \mathrm{d}\lambda / \int_{\mathcal{C}_\star} \mu \mathrm{d}\lambda$. Then for any $x \in \mathcal{C}_\star$ and any measurable subset A of $\mathcal{C}_\star$, it readily follows from (4.3) that

$$\begin{aligned}
\mathsf{P}_{\mathrm{GL}}(x, A) &\geq \int_{A\cap\mathcal{A}} q(x,z)\mathbf{1}_{\mu(z)\neq 0}\mathrm{d}\lambda(z) \\
&\geq \inf_{(x,z)\in\mathcal{C}_\star^2} \left(\frac{q(x,z)\mathbf{1}_{\mu(z)\neq 0}}{\mu(z)}\right) \int_{A\cap\mathcal{A}} \mu(z)\mathrm{d}\lambda(z) \\
&= \inf_{(x,z)\in\mathcal{C}_\star^2} \left(\frac{q(x,z)\mathbf{1}_{\mu(z)\neq 0}}{\mu(z)}\right) \left(\int_{\mathcal{C}_\star} \mu \mathrm{d}\lambda\right) \nu(A\cap\mathcal{A}).
\end{aligned}$$

Thanks to the lower bounds of (iv), we complete the proof.

4.3 Proof of Theorem 4

We write $\widehat{\mathcal{I}}_M - \mathcal{I} = \mathcal{T}_1 + \mathcal{T}_2$ with

$$\begin{aligned}
\mathcal{T}_1 &:= \frac{1}{M}\sum_{m=1}^{M} f\left(X^{(m)}, \widehat{\phi}_M(X^{(m)})\right) - \frac{1}{M}\sum_{m=1}^{M} f\left(X^{(m)}, \phi_\star(X^{(m)})\right), \\
\mathcal{T}_2 &:= \frac{1}{M}\sum_{m=1}^{M} f\left(X^{(m)}, \phi_\star(X^{(m)})\right) - \int f(x, \phi_\star(x))\mu(x)\,\mathrm{d}\lambda(x).
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{T}_1|^2\right] &\leq \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\left|f\left(X^{(m)}, \widehat{\phi}_M(X^{(m)})\right) - f\left(X^{(m)}, \phi_\star(X^{(m)})\right)\right|^2\right] \\
&\leq C_f^2\,\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\left|\widehat{\phi}_M(X^{(m)}) - \phi_\star(X^{(m)})\right|^2\right] = C_f^2\Delta_M.
\end{aligned}$$

The second term is controlled by Assumption (ii). We then conclude by the Minkowski inequality.

A Algorithm where the inner stage uses a crude Monte Carlo method and the outer stage uses MCMC sampling

Here, the regression function $\phi_\star$ is approximated by an empirical mean using N (conditionally) independent samples $R^{(m,k)}$, as in (1.2). We keep the same notations as in Section 2.

A.1 Algorithm

```
/* Simulation of the design and the observations                    */
X^(0) ~ ξ, where ξ is a distribution on A ;
for m = 1 to M do
    X^(m) ~ P(X^(m-1), dx) ;
    for k = 1 to N do
        R^(m,k) ~ Q(X^(m), dr);
/* Conditional expectation by crude Monte Carlo                     */
Compute R̄_N^(m) = (1/N) Σ_{k=1}^N R^(m,k);
/* Final estimator using ergodic average                            */
Return Ĩ_M := (1/M) Σ_{m=1}^M f(X^(m), R̄_N^(m)).
```

Algorithm 3: Full algorithm with M outer samples, and N inner samples for each outer one.

A.2 Convergence results for the estimation of $\widetilde{\mathcal{I}}_M$

We extend Theorem 1 to this new setting. Actually when the function f in (1.1) is smoother than Lipschitz continuous, we can prove that the impact of N on the quadratic error is $1/N$ instead of the usual $1/\sqrt{N}$. This kind of improvement has been derived by [GJ10] in the i.i.d. setting (for both the inner and outer stages).

Theorem 5. *Assume that*

(i) the (second and) fourth conditional moments of Q *are bounded: for* $p = 2$ *and* $p = 4$, *we have*

$$\sigma_p := \left(\sup_{x \in \mathcal{A}} \int \left| r - \int r \, \mathsf{Q}(x, \mathrm{d}r) \right|^p \mathsf{Q}(x, \mathrm{d}r) \right)^{1/p} < \infty.$$

(ii) There exists a finite constant C *such that for any* M

$$\mathbb{E}\left[\left(M^{-1} \sum_{m=1}^{M} f\left(X^{(m)}, \phi_\star(X^{(m)}) \right) - \int f(x, \phi_\star(x)) \mu(x) \, \mathrm{d}\lambda(x) \right)^2 \right] \leq \frac{C}{M}.$$

If $f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ *is globally* C_f*-Lipschitz in the second variable, then*

$$\left(\mathbb{E}\left[\left| \widetilde{\mathcal{I}}_M - \mathcal{I} \right|^2 \right] \right)^{1/2} \leq \frac{C_f \sigma_2}{\sqrt{N}} + \sqrt{\frac{C}{M}},$$

where $\mathcal{I}$ *and* $\widetilde{\mathcal{I}}_M$ *are resp. given by (1.1) and Algorithm 3.*

If f *is continuously differentiable in the second variable, with a derivative in the second variable which is bounded and globally* $C_{\partial_r f}$*-Lipschitz, then*

$$\left(\mathbb{E}\left[\left| \widetilde{\mathcal{I}}_M - \mathcal{I} \right|^2 \right] \right)^{1/2} \leq \frac{C_{\partial_r f}}{2} \frac{1}{N} \sqrt{3\sigma_2^4 + \frac{\sigma_4^4}{N}} + \frac{\sigma_2}{\sqrt{NM}} \sup_x |\partial_r f(x, \phi_\star(x))| + \sqrt{\frac{C}{M}}.$$

A.3 Proof of Theorem 5

▷ *1st case: f Lipschitz.* We write $\widetilde{\mathcal{I}}_M - \mathcal{I} = \mathcal{T}_1 + \mathcal{T}_2$ with

$$\mathcal{T}_1 := \frac{1}{M}\sum_{m=1}^{M} f\left(X^{(m)}, \overline{R}_N^{(m)}\right) - \frac{1}{M}\sum_{m=1}^{M} f\left(X^{(m)}, \phi_\star(X^{(m)})\right),$$
$$\mathcal{T}_2 := \frac{1}{M}\sum_{m=1}^{M} f\left(X^{(m)}, \phi_\star(X^{(m)})\right) - \int f(x, \phi_\star(x))\mu(x)\,\mathrm{d}\lambda(x).$$

For the first term, since f is globally Lipschitz with constant C_f, we have

$$\mathbb{E}\left[|\mathcal{T}_1|^2\right] \le C_f^2\,\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\left|\overline{R}_N^{(m)} - \phi_\star(X^{(m)})\right|^2\right].$$

Since $(R^{(m,k)} : 1 \le k \le N)$ are independent conditionally on $X^{(1:M)}$, $\mathbb{E}\left[\overline{R}_N^{(m)} \mid X^{(1:M)}\right] = \phi_\star(X^{(m)})$ and $\mathrm{Var}\left[\overline{R}_N^{(m)} \mid X^{(1:M)}\right] \le \sigma_2^2/N$. Thus,

$$\mathbb{E}\left[|\mathcal{T}_1|^2\right] \le \frac{C_f^2}{N}\sigma_2^2.$$

The second term is controlled by Assumption (ii). We are done.

▷ *2nd case: f smooth.* Set $\mathcal{T}_1 = \mathcal{T}_{1,a} + \mathcal{T}_{1,b}$ with

$$\mathcal{T}_{1,a} := \frac{1}{M}\sum_{m=1}^{M}\left(f(X^{(m)}, \overline{R}_N^{(m)}) - f(X^{(m)}, \phi_\star(X^{(m)})) - \partial_r f(X^{(m)}, \phi_\star(X^{(m)}))(\overline{R}_N^{(m)} - \phi_\star(X^{(m)}))\right),$$
$$\mathcal{T}_{1,b} := \frac{1}{M}\sum_{m=1}^{M}\partial_r f(X^{(m)}, \phi_\star(X^{(m)}))(\overline{R}_N^{(m)} - \phi_\star(X^{(m)})).$$

A Taylor expansion gives

$$|\mathcal{T}_{1,a}| \le \frac{1}{2}C_{\partial_r f}\,\frac{1}{M}\sum_{m=1}^{M}\left|\overline{R}_N^{(m)} - \phi_\star(X^{(m)})\right|^2,$$
$$\mathbb{E}\left[|\mathcal{T}_{1,a}|^2\right] \le \left(\frac{1}{2}C_{\partial_r f}\right)^2\,\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left|\overline{R}_N^{(m)} - \phi_\star(X^{(m)})\right|^4\right].$$

Invoking that $(R^{(m,k)} : 1 \le k \le N)$ are independent conditionally on $X^{(1:M)}$ leads to

$$\mathbb{E}\left[\left|\overline{R}_N^{(m)} - \phi_\star(X^{(m)})\right|^4 \mid X^{(1:M)}\right] \le 3\sigma_2^4\frac{N-1}{N^3} + \sigma_4^4\frac{1}{N^3}.$$

Moreover, upon noting that for $m \ne m'$,

$$\mathbb{E}\left[\left(\partial_r f(X^{(m)}, \phi_\star(X^{(m)}))(\overline{R}_N^{(m)} - \phi_\star(X^{(m)}))\right)\left(\partial_r f(X^{(m')}, \phi_\star(X^{(m')}))(\overline{R}_N^{(m')} - \phi_\star(X^{(m')}))\right)\right] = 0,$$

we have

$$\mathbb{E}\left[|\mathcal{T}_{1,b}|^2\right] = \mathbb{E}\left[\frac{1}{M^2}\sum_{m=1}^{M}\left|\partial_r f(X^{(m)}, \phi_\star(X^{(m)}))\right|^2\mathbb{E}\left[\left|\overline{R}_N^{(m)} - \phi_\star(X^{(m)})\right|^2 \mid X^{(1:M)}\right]\right]$$
$$\le \frac{\sup_x |\partial_r f(x, \phi_\star(x))|^2}{M}\frac{\sigma_2^2}{N}.$$

This concludes the proof. □

References

[BCV01] Y. Baraud, F. Comte and G. Viennet. Adaptive estimation in autoregression or β-mixing regression via model selection. *The Annals of Statistics*, 29(3):839–875, 2001.

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

[BG13] T. Ben Zineb and E. Gobet. Preliminary control variates to improve empirical regression methods. *Monte-Carlo methods and Applications*, 19(4):331–354, 2013.

[BKS10] D. Belomestny, A. Kolodko, and J. Schoenmakers. Regression methods for stochastic control problems and their convergence analysis. *SIAM Journal on Control and Optimization*, 48(5):3562–3588, 2010.

[BL12] J. Blanchet and H. Lam. State-dependent Importance Sampling for Rare Event Simulation: An Overview and Recent advances. *Surveys in Operations Research and Management Sciences*, 17:38–59, 2012.

[DG11] S. Delattre and S. Gaïffas. Nonparametric regression with martingale increment errors. *Stochastic Processes and their Applications*, 121:2899–2924, 2011.

[DL09] L. Devineau and S. Loisel. Construction d'un algorithme d'accélération de la méthode des «simulations dans les simulations» pour le calcul du capital économique solvabilité ii. *Bulletin Français d'Actuariat*, 10(17):188–221, 2009.

[DFMS04] R. Douc, G. Fort, E. Moulines and P. Soulier. Practical drift conditions for subgeometric rates of convergence. *Ann. Appl. Probab.*, 14(3) :1353-1377, 2004.

[Egl05] D. Egloff. Monte-Carlo algorithms for optimal stopping and statistical learning. *Ann. Appl. Probab.*, 15:1396–1432, 2005.

[FM03a] G. Fort and E. Moulines. Polynomial ergodicity of Markov transition kernels. *Stochastic Processes Appl.*, 103(1):57-99, 2003.

[FM03b] G. Fort and E. Moulines. Convergence of the Monte Carlo Expectation Maximization for Curved Exponential Families. *Ann. Statist.*, 31(4):1220-1259, 2003.

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

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

[GL15] E. Gobet and G. Liu. Rare event simulation using reversible shaking transformations. *SIAM Scientific Computing*, 37(5):A2295–A2316, 2015.

[GT16] E.Gobet and P. Turkedjiev. Linear regression MDP scheme for discrete backward stochastic differential equations under general conditions. *Mathematics of Computation*, 299(85):1359–1391, 2016.

[HJ09] L.J. Hong and S. Juneja. Estimating the mean of a non-linear function of conditional expectation. In *Simulation Conference (WSC), Proceedings of the 2009 Winter*, pages 1223–1236. IEEE, 2009.

[LGW06] 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(5):889–916, 2006.

[LS01] F. Longstaff and E.S. Schwartz. Valuing American options by simulation: A simple least squares approach. *The Review of Financial Studies*, 14:113–147, 2001.

[LS10] M. Liu and J. Staum. Stochastic kriging for efficient nested simulation of expected shortfall. *The Journal of Risk*, 12(3):3, 2010.

[MFE05] A.J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management.* Princeton series in finance. Princeton University Press, 2005.

[MT93] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, 1993.

[Ros08] J.S. Rosenthal *Optimal Proposal Distributions and Adaptive MCMC.* In *Handbook of MCMC*, Chapman&Hall, 2008.

[RK08] R.Y. Rubinstein and D.P. Kroese. *Simulation and the Monte-Carlo method.* Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2008.

[RM10] Q. Ren and M. Mojirsheibani. A Note on nonparametric regression with β-mixing sequences. *Communications in Statistics - Theory and Methods*, 39(12):2280–2287, 2010.

[TR01] J.N. Tsitsiklis and B. Van Roy. Regression methods for pricing complex American-style options. *IEEE Transactions on Neural Networks*, 12(4):694–703, 2001.