Bernoulli **9**(6), 2003, 1003–1049

A quantization algorithm for solving multi-dimensional discrete-time optimal stopping problems

VLAD BALLY[1] and GILLES PAGÈS[2]

[1]*Laboratoire d'Analyse et Mathématiques Appliquées, UMR 8050, Université de Marne-la-Vallée Cité Descartes, 5 Boulevard Descartes, Champs-sur-Marne, F-77454 Marne-la-Vallée, France. E-mail: vlad.bally@math.univ-mlv.fr*

[2]*Laboratoire de Probabilités et Modèles Aléatoires, UMR 7599, Université Paris 6, case 188, 4 Place Jussieu, F-75252 Paris Cedex 05, France. E-mail: gpa@ccr.jussieu.fr.*

A new grid method for computing the Snell envelope of a function of an $\mathbb{R}^d$-valued simulatable Markov chain $(X_k)_{0\leqslant k\leqslant n}$ is proposed. (This is a typical nonlinear problem that cannot be solved by the standard Monte Carlo method.) Every X_k is replaced by a 'quantized approximation' $\hat{X}_k$ taking its values in a grid Γ_k of size N_k. The n grids and their transition probability matrices form a discrete tree on which a pseudo-Snell envelope is devised by mimicking the regular dynamic programming formula. Using the quantization theory of random vectors, we show the existence of a set of optimal grids, given the total number N of elementary $\mathbb{R}^d$-valued quantizers. A recursive stochastic gradient algorithm, based on simulations of $(X_k)_{0\leqslant k\leqslant n}$, yields these optimal grids and their transition probability matrices. Some a priori error estimates based on the L^p-quantization errors $\|X_k - \hat{X}_k\|_p$ are established. These results are applied to the computation of the Snell envelope of a diffusion approximated by its (Gaussian) Euler scheme. We apply these results to provide a discretization scheme for reflected backward stochastic differential equations. Finally, a numerical experiment is carried out on a two-dimensional American option pricing problem.

Keywords: American option pricing; Markov chains; numerical probability; quantization of random variables; reflected backward stochastic differential equation; Snell envelope

1. Introduction

Since the 1940s, the theory of Markov processes and stochastic calculus have provided a probabilistic interpretation for the solutions of linear partial differential equations (PDEs) based on the Feynman–Kac formula. One of its most striking applications is the emergence of the Monte Carlo method as an alternative to deterministic numerical algorithms for solving linear PDEs. It is widely known that the Monte Carlo method has two advantages: its rate of convergence does not depend upon the dimension d of the state space and is not affected by possible degeneracy of the second-order terms of the equation. For $d \geqslant 4$ the probabilistic approach often remains the only numerical method available.

In the 1990s the theory of backward stochastic differential equations (BSDEs; see

1350–7265 © 2003 ISI/BS

Pardoux and Peng 1992; El Karoui *et al*., 1997a; 1997b; Bally *et al*. 2002a) provided a probabilistic interpretation for nonlinear problems (semi-linear PDEs, PDEs with obstacle etc.). For example, let us focus for a while on the problem of semi-linear PDEs with obstacle (in the weak sense):

$$\max((\partial_t + L)u + f(t, x, u), h(t, x) - u(t, x)) = 0, \quad 0 \leqslant t \leqslant T, u_T = h(T, \cdot), \tag{1}$$

where $f : [0, T] \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, $h : [0, T] \times \mathbb{R}^d \to \mathbb{R}$ are Lipschitz continuous and L is a second-order differential operator defined on twice differentiable functions on $\mathbb{R}^d$ by

$$Lu(x) := \langle b|\nabla u\rangle(x) + \frac{1}{2}\mathrm{tr}(\sigma^* \nabla^2 u\, \sigma)(x)$$

($b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathcal{M}(d \times q)$ are Lipschitz continuous functions, $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$, $(\cdot|\cdot)$ for the inner product). The object involved in the probabilistic interpretation of (1) is the reflected backward stochastic differential equation (RBSDE) associated with the diffusion process $(X_t)_{t\in[0,T]}$ solution of the stochastic differential equation

$$X_t = x + \int_0^t b(X_s)\mathrm{d}s + \int_0^t \sigma(X_s)\mathrm{d}B_s, \tag{2}$$

where $(B_t)_{t\in[0,T]}$ is standard Brownian motion on $\mathbb{R}^q$ (its completed filtration is denoted by $\underline{\mathcal{F}} := (\mathcal{F}_\mathrm{t})_{\mathrm{t}\in[0,\mathrm{T}]}$). The (solution) of the RBSDE is defined as a triplet (Y, Z, K) of square-integrable, $\underline{\mathcal{F}}$-progressively measurable processes satisfying

$$Y_t = h(T, X_T) + \int_t^T f(s, X_s, Y_s)\mathrm{d}s + K_T - K_t - \int_t^T Z_s \mathrm{d}B_s, \tag{3}$$

$$Y_t \geqslant h(t, X_t) \text{ and } \int_0^T (Y_t - h(t, X_t))\,\mathrm{d}K_t = 0, \tag{4}$$

where $(K_t)_{t\in[0,T]}$ has non-decreasing continuous paths and $K_0 := 0$.

We wish to solve a BSDE (i.e. (3), but we also require Y to remain larger than the obstacle $h(t, X_t)$). Then we need a non-decreasing process K to 'push' Y upwards. K is required to be minimal: it pushes in the critical situation $Y_t = h(t, X_t)$ only. Z is the strategy to be used so that Y_t starts with Y_0 at time $t = 0$ and reaches $h(T, X_T)$ at time T *in a non-anticipative way*, although $h(T, X_T)$ depends on the entire information up to T.

The following theorem is due to El Karoui *et al*. (1997a).

Theorem 1. *Assume that the following assumptions hold for some real constant* $\gamma_0 > 0$*:*

$$\forall x, x' \in \mathbb{R}^d, \qquad |\sigma(x) - \sigma(x')| \vee |b(X) - b(x')| \leqslant \gamma_0|x - x'|, \tag{5}$$

$\forall t, t' \in \mathbb{R}_+, x, x' \in \mathbb{R}^d, y, y' \in \mathbb{R},$

$$|f(t, x, y) - f(t', x', y')| \leqslant \gamma_0(|t - t'| + |x - x'| + |y - y'|), \tag{6}$$

$$\forall t, t' \in \mathbb{R}_+, x, x' \in \mathbb{R}^d, \qquad |h(t, x) - h(t', x')| \leqslant \gamma_0(|x - x'| + |t - t'|). \tag{7}$$

Then the RBSDE (3) *has a unique solution* (Y, Z, K)*. Furthermore, the process Y admits the representation*

$$Y_t = u(t, X_t),$$

where u denotes the unique solution (in the viscosity sense) of (1).

Another approach is developed in Bally *et al*. (2002a): the function u solves in a variational sense the PDE with obstacle h, and u is the minimal solution for the corresponding variational inequality. Then, using the connections between variational inequalities and optimal stopping theory (see Bensoussan and Lions 1982) leads to the representation of the above process Y as a Snell envelope:

Proposition 1. *If* $(Y_t)_{t\in[0,T]}$ *solves the RBSDE* (3)*, then*

$$Y_t = \operatorname{ess\,sup}_{\tau\in\mathcal{T}_t}\mathbb{E}\left(\int_t^\tau f(s, X_s, Y_s)\mathrm{d}s + h(\tau, X_\tau)|\mathcal{F}_t\right), \tag{8}$$

where $\mathcal{T}_t$ *is the set of* $[t, T]$*-valued* $\underline{\mathcal{F}}$*-stopping times.*

When the function f *does not depend upon* Y_t, equation (8) becomes an alternative definition of $(Y_t)_{t\in[0,T]}$ which then appears as the usual Snell envelope of a regular optimal stopping problem associated with the Brownian diffusion $(X_t)_{t\in[0,T]}$. Then f represents the *instantaneous gain* and h the *final gain*. Of course optimal stopping theory for diffusions is a classical topic in probability theory, and its numerical aspects have been investigated since the very beginnings of numerical probability, motivated by a wide range of applications in engineering; see, for example Kushner (1977) on elliptic diffusions, and Bensoussan and Lions (1982). However, mathematical finance made it still more strategic in the 1980s: pricing an American option is in some way an almost generic optimal stopping problem (with $f \equiv 0$, $h \geqslant 0$ and X a non-negative martingale).

In brief, a (vanilla) American option is a contract that gives the right to receive once and only once $h(t, X_t)$ currency units, at a time t chosen between time 0 and the maturity $T > 0$ of the contract. The possibly multidimensional (non-negative) process $(X_t)_{t\in[0,T]}$ is called the *underlying asset* or *risky asset* process. If one assumes for the sake of simplicity that the interest rate is 0, classical arguments in the modelling of financial markets show that the price Y_t of such a contract is given at every time t by (8) (setting $f \equiv 0$), with respect to a so-called *risk-neutral* probability which makes the diffusion $(X_t)_{t\in[0,T]}$ a martingale. Among all possible models for the asset dynamics, the geometric Brownian motion on $\mathbb{R}^q$,

$$\mathrm{d}X_t = \sigma X_t\,\mathrm{d}B_t, \qquad X_0 = x_0 \in \mathbb{R}_+^d, \qquad \sigma \in \mathcal{M}(d, q),$$

is widely used. When $q = d$, it is known as the *Black–Scholes model*. Another question of interest is whether, at any time t, there is some optimal stopping strategy for exercising this right in the future. An answer is provided by the (lowest) optimal stopping time given by $\tau_*^t := \inf\{s \geqslant t | Y_s = h(s, X_s)\}$ in the sense that τ_*^t satisfies $Y_{\tau_*^t} = \mathbb{E}(h_{\tau_*^t}|\mathcal{F}_t)$.

Historically, the underlying asset of the first massively traded American option contracts was one-dimensional (a single stock). However, many American options, mostly traded ‘over the counter’, have a much more complex structure depending on a whole basket of d

underlying risky assets $X_t := (X_t^1, \dots, X_t^d)$. If one thinks, for example, of indices (Dow Jones, DAX, CAC40, etc.), d is usually greater than 2, 3 or 4.

The usual numerical approach to solving (low-dimensional) optimal stopping problems is essentially analytic: it consists in solving the variational inequality (1) using standard techniques of numerical analysis (finite differences, finite elements, finite volumes, etc.). In spite of the loss of probabilistic interpretation, especially when implicit schemes are used, these methods are unrivalled in one or two dimensions in terms of their rate of convergence. This holds similarly for RBSDEs.

However, the autonomous development of mathematical finance, mainly influenced by its probabilistic background, gave rise to algorithms directly derived from discretizations of the Snell envelope (8) (when $f \equiv 0$). The most famous method is undoubtedly the binomial tree made very popular in the financial world by its simplicity of implementation and interpretation. Some very accurate rates of convergence are available for such models in the case of the vanilla American put option ($h(t, x) := \max(K - x, 0)$) (see Lamberton 1998; 2002; Lamberton and Rogers 2000; Bally and Saussereau 2002).

As the dimension increases, analytic methods become insufficient and the probabilistic interpretation becomes the key to any numerical approach. Although this paper is concerned with the more general case of the (h, f)-Snell envelopes embodying the numerical approximation of solutions of RBSDEs, let us persist with regular optimal stopping for a while. All probabilistic approaches roughly follow the same three steps:

1. *Time discretization*. One approximates the $\underline{\mathcal{F}}$-adapted diffusion process $(X_t)_{t\in[0,T]}$ at times $t_k = kT/n$, $k = 0, \dots, n$, by an $(\tilde{\mathcal{F}}_k)_{0\leqslant k\leqslant n}$*-Markov chain* $\tilde{X} = (\tilde{X}_k)_{0\leqslant k\leqslant n}$, where $\tilde{\mathcal{F}}_k = \mathcal{F}_{t_k}$. The chain $\tilde{X}$ is assumed to be easily simulatable on a computer (index k then stands for absolute time t_k). Then one approximates the continuous-time $\mathcal{F}$-Snell envelope Y of X by the discrete time $\underline{\tilde{\mathcal{F}}}$-Snell envelope of $\tilde{X}$ defined by

$$\tilde{U}_k := \operatorname{ess\,sup}\{\mathbb{E}(h(\theta T/n, \tilde{X}_\theta)|\tilde{\mathcal{F}}_k),\ \theta \in \Theta_k\},$$

where Θ_k denotes the set of $\{k, \dots, n\}$-valued stopping times. The approximating process $\tilde{X}$ will often be chosen to be the Euler scheme $(\overline{X}_k)_{0\leqslant k\leqslant n}$ (with Gaussian increments) of the diffusion, but other choices are possible (Milshtein scheme, etc). In one-dimension a binomial tree can be considered as a weak approximation. When samples $(X_0, X_{t_1}, \dots, X_{t_n})$ of the diffusion are simulatable, for instance because $X_t = \varphi(t, B_t)$ as in the Black–Scholes model, the best choice is of course to consider $\tilde{X}_k = X_{t_k}$. In that case, its $(\tilde{\mathcal{F}}_k)_{0\leqslant k\leqslant n}$-Snell envelope will be simply denoted by $(U_k)_{0\leqslant k\leqslant n}$. Bally and Pagès (2003) carry out a detailed analysis of the resulting L^p-error: if assumptions (5)–(7) hold and $\tilde{X}_k = \overline{X}_k$ or X_{t_k}, then $\|Y_{kT/n} - \tilde{U}_k\|_p = O(1/\sqrt{n})$; if, furthermore, $\tilde{X}_k = X_{t_k}$ and h is *semi-convex* (see (16) below for a definition), then $\|Y_{kT/n} - U_k\|_p = O(1/n)$.

2. *Dynamic programming principle*. The discrete-time Snell envelope associated with the obstacle $(h(t_k, \tilde{X}_k))_{0\leqslant k\leqslant n}$ satisfies (see Neveu 1971) the following backward dynamic programming principle:

$$\tilde{U}_n := h(t_n, \tilde{X}_n), \qquad \tilde{U}_k := \max(h(t_k, \tilde{X}_k), \mathbb{E}(\tilde{U}_{k+1}|\tilde{\mathcal{F}}_k)) = \max(h(t_k, \tilde{X}_k), \mathbb{E}(\tilde{U}_{k+1}|\tilde{X}_k)). \tag{9}$$

The main feature of this formula is that it involves at each step *the computation of conditional expectations*: this is the probabilistic counterpart for nonlinearity which makes the regular Monte Carlo method fail.

3. *Computation of conditional expectations*. Numerical methods for the massive computation of conditional expectations can roughly be divided into three families: spatial discretization of $\tilde{X}_k$; regression of (truncated) expansions on a basis of $L^2(\tilde{X}_k)$ (see Longstaff and Schwartz, 2001); and representation formulae based on Malliavin calculus (as in Fournié *et al*. 1999). The first two approaches are both finite-dimensional. An important drawback of the last two methods – especially in higher dimensions – is that they directly depend on the obstacle process $(h(t_k, \cdot))$. A spatial discretization method is 'obstacle free' in the sense that it produces a discrete semi-group independently of any obstacle process and then works for any such obstacle process.

The *quantization tree method* that we propose and analyse in this paper belongs to the first family (spatial discretization). A specific characteristic of the method is that it does not explode, even in higher dimensions. The initial idea is simple and shared by many grid methods (see Broadie and Glasserman 1997; Chevance 1997; Longstaff and Schwartz 2001). First, at each time step k (i.e. t_k) one projects $\tilde{X}_k$ onto a fixed grid $\Gamma_k := \{x_1^k, \ldots, x_{N_k}^k\}$ following *a closest neighbour rule*, that is to say, one sets

$$\hat{X}_k := \sum_{1 \leqslant i \leqslant N_k} x_i^k \mathbf{1}_{\{\tilde{X}_k \in C_i^k\}},$$

where $(C_i^k)_{1 \leqslant i \leqslant N_k}$ is a Borel partition of $\mathbb{R}^d$ such that $C_i^k \subset \{u|\, |u - x_i^k| = \min_{1 \leqslant \ell \leqslant N_k} |u - x_\ell^k|\}$.

As a second step, one 'mimics' the above dynamic programming principle (9) on the tree formed by the grids $\Gamma_0, \ldots, \Gamma_n$. The process $\tilde{X}$ being simulatable, it is possible to compute by simulation $\pi_{ij}^k := \mathbb{P}(\tilde{X}_{k+1} \in C_j^{k+1} | \tilde{X}_k \in C_i^k)$. Although $(\hat{X}_k)_{0 \leqslant k \leqslant n}$ is not a Markov chain, one may still define a pseudo-Snell envelope $(\hat{U}_k)_{0 \leqslant k \leqslant n}$ by setting

$$\hat{U}_n := h(\hat{X}_n), \quad \text{and} \quad \hat{U}_k := \max(h(\hat{X}_k), \mathbb{E}(\hat{U}_{k+1} | \hat{X}_k)), \qquad 0 \leqslant k \leqslant n-1.$$

Since $\mathbb{E}(\hat{U}_{k+1} | \hat{X}_k = x_i^k) = \sum_{j=1}^{N_{k+1}} \pi_{ij}^k \hat{u}_{k+1}(x_j^{k+1})$, a backward induction shows that $\hat{U}_k := \hat{u}_k(\hat{X}_k)$, where $\hat{u}_k$ satisfies the backward dynamic programming formula

$$\hat{u}_n(x_i^n) = h(x_i^n), \qquad 1 \leqslant i \leqslant N_n,$$

$$\hat{u}_k(x_i^k) = \max(h(x_i^k), \sum_{j=1}^{N_{k+1}} \hat{u}_{k+1}(x_j^{k+1}) \pi_{ij}^k), \qquad 1 \leqslant i \leqslant N_k,\ 0 \leqslant k \leqslant n-1, \tag{10}$$

which we will call the *quantization tree algorithm*. One may reasonably expect the error $\tilde{U}_k - \hat{U}_k$ to be small. Indeed, we are able to prove that, *for some specific choices of grid* Γ_k, and under some appropriate assumptions on the diffusion coefficients, for every $p \geqslant 1$,

$$\|\hat{U}_k - \tilde{U}_k\|_p \leqslant C_p \sum_{k=1}^{n} \|\hat{X}_k - \tilde{X}_k\|_p = O\left(\frac{n^{1+1/d}}{N^{1/d}}\right).$$

Practical processing of the quantization tree algorithm (10) raises the following questions:

A. How do we specify the grids $\Gamma_k := \{x_1^k, \ldots, x_{N_k}^k\}$? This means two things: first, how do we choose in an optimal way the sizes N_k of the grids, given that $N_0 + N_1 + \ldots + N_n = N$; and second, how do we choose the points x_i^k to keep the L^p-quantization errors $\|\tilde{X}_k - \hat{X}_k\|_p$ minimal?
B. How do we compute the weights π_{ij}^k?
C. How do we evaluate the error $\|\tilde{U}_k - Y_{kT/n}\|_p$ (and $\|U_k - Y_{kT/n}\|_p$)?
D. What is the complexity of the quantization tree algorithm?

One crucial feature of the problem must be emphasized at this stage: whatever the methods selected to obtain optimal grids and weights, phases A, B and C are 'one-shot': once the grids are settled, their weights and resulting L^p-quantization errors estimated, the computation of the pseudo-Snell envelope of $(h(t_k, \hat{X}_k))_{0 \leqslant k \leqslant n}$ using the quantization tree algorithm (10) is almost instantaneous on any computer. The numerical experiments carried out in Section 6.2 (pricing of American options) indicate that, in fact, the quantization tree grid optimization phase outlined below in Section 3.3 entails a very reasonable cost (less than 15 minutes on a on 1 GHz PC computer), given the fact that, once completed, one can instantly price any American pay-off option in that model. Furthermore, in many applications, one may rely on the quantization of universal objects such as standard Brownian motion. In this latter case, the optimization of the quantization amounts by scaling to that of a normal q-dimensional vector $\mathcal{N}(0; I_q)$ processed once and stored on a CD-ROM.

Let us turn now to the optimization phases (A, B and C). Actually this is a very old story: since the early 1950s people working in signal processing and information theory have been concerned with the compression of the information contained in a continuous 'signal' $(\tilde{X}_k)$ using a finite number of 'codebooks' (the points x_i^k) in an optimal way (see Section 3.1). Several deterministic algorithms have been designed for this purpose, essentially one-dimensional signals. Among them, let us mention Lloyd's Method I (see Kieffer 1982). Meanwhile, a sound mathematical theory of quantization of probability distributions has been developed (for a recent monograph, see Graf and Luschgy 2000). In the 1980s, with the emergence of artificial neural networks, some new algorithmic aspects of quantization in higher dimensions were investigated, mainly the *competitive learning vector quantization* algorithm (and its variants) which appeared as a degenerate setting for the *Kohonen self-organizing maps* (see Fort and Pagès 1995; Bouton and Pagès 1997; and the references therein). This stochastic algorithmic approach will be adapted below to Markov dynamics. It is based on massive simulations of independent copies of the random vector to be quantized. Pagès (1997) appears to have made the first attempt to apply optimal multidimensional quantization to numerical probability.

As mentioned above, after Kushner's pioneering work, the pricing of American options led to renewed interest in numerical aspects of optimal stopping: we might mention, among many others, the analysis of the convergence of the Snell envelope in an abstract approximation framework (see Lamberton and Pagès 1990) or the rate of convergence of the premium of a regular American put priced in a binomial tree toward its Black–Scholes counterpart (see Lamberton 1998; and the references therein).

In higher dimensions, several numerical methods have been designed and analysed in the literature of the past ten years to solve optimal stopping problems like those naturally arising in finance or, more generally, to process massive computations of conditional expectations. In the class of grid methods, one may cite the algorithm devised by Broadie and Glasserman (1997) for pricing multiasset American options, and the discretization scheme for BSDEs proposed by Chevance (1997) – the latter is one-dimensional but easy to extend to higher dimensions. In both approaches the spatial discretization of discrete-time approximation $(\tilde{X}_k)_k$ consists of N independent copies of $(\tilde{X}_k)_{0\leqslant k\leqslant n}$ which form a grid of size N at every time step $k = 1, \ldots, n$. In Broadie and Glasserman (1997), the transition between these grids is based on the likelihood ratios between $\tilde{X}_k$ and $\tilde{X}_{k+1}$. A convergence theorem is established without rate of convergence. In Chevance (1997), $(\tilde{X}_k)_{0\leqslant k\leqslant n}$ is in fact the Euler scheme of a diffusion and at every time step the grid is uniformly weighted by $1/N$. The transition weights are based on an empirical frequency approach based on a Monte Carlo simulation as well. Some a priori L^1-error bounds are proposed for functions h having finite variations. Longstaff and Schwartz (2001) develop a regression method based on truncated expansions in $L^2(\tilde{X}_k)$. They introduce a dual dynamic programming principle for the lowest stopping time τ^* and then compute $\mathbb{E}(h(\tau^*, X_{\tau^*}))$ by a Monte Carlo simulation. The 'Malliavin calculus' method was introduced by Fournié *et al*. (1999) and then developed in Fournié *et al*. (2001) and Lions and Régnier (2002). These papers point out the importance of a localization procedure for variance reduction purposes. Optimal localization is investigated in one-dimension in Kohatsu-Higa and Petterson (2002) and extended to d dimensions in Bouchard-Denize and Touzi (2002).

For a weak convergence approach to RSBDE discretization, see Ma *et al*. (2002) as well as Briand *et al.* (2001; 2002) (the latter two are less directly focused on numerical aspects).

Before outlining the structure of the paper, we list some notation:

- For any Lipschitz continuous function $\varphi : \mathbb{R}^d \to \mathbb{R}$, we denote by $[\varphi]_{\text{Lip}}$ its Lipschitz coefficient $[\varphi]_{\text{Lip}} := \sup_{x\neq y}|\frac{\varphi(x)-\varphi(y)}{x-y}|$.
- The set $\mathcal{M}(d \times q, \mathbb{R})$ of matrices with d rows, q columns and real-valued entries will be endowed with the norm $\|M\| := \sqrt{\text{tr}(MM^*)}$, where M^* denotes the transpose of M.
- For every finite set A, we denote by $|A|$ its cardinality.
- $\delta_{x,y}$ denotes the usual Kronecker delta.
- For every $x \in \mathbb{R}$, $[x] := \max\{n \in \mathbb{Z} | n \leqslant x\}$ and $\lceil x \rceil := \min\{n \in \mathbb{Z} | n \geqslant x\}$.

Section 2 is devoted to the computation of the Snell envelope of a discrete-time $\mathbb{R}^d$-valued homogeneous Markov chain using a quantization tree. We start with an example, the discretization of an RBSDE, in Section 2.1, to introduce the (h, f)-Snell envelope. In Section 2.2 we propose the (backward) quantization tree algorithm and derive some a priori L^p-error bounds using the L^p-quantization error. Section 3 is devoted to optimal quantization from a theoretical point of view. Then the extension of the *competitive learning vector quantization algorithm* to Markov chains is presented to process the numerical optimization of the grids and the computation of their transition weights. Section 4 briefly recalls some error bounds concerning the Monte Carlo estimation of the transition weights (for a fixed, possibly not optimal, quantization tree and in the linear case $f \equiv 0$). These are established in Bally and Pagès (2003). In Section 5 a first comparison with the finite-element method is carried out. In Section 6, the above results are applied to the discretization of RBSDEs, with some a priori error bounds, when the diffusion is uniformly elliptic. We conclude with a numerical illustration: the pricing of American style exchange options.

2. Quantization of the Snell envelope of a Markov chain

Before dealing with the general case, let us look more precisely at the case of the RBSDE presented in the Introduction.

2.1. Time discretization of an RBSDE by an (h, f)-Snell envelope

We follow the notation introduced in the Introduction for RBSDEs. It is natural to derive a discretization scheme for the solution of an RBSDE from the representation formula (8) following the approach described in the Introduction. Let $t_k := kT/n$, $k = 0, \dots, n$, denote the discretization epochs. One just needs to add a discretization term for the integral $\int_t^T f(s, X_s, Y_s)\,\mathrm{d}s$.

One first considers the homogeneous Markov chain $(X_{t_k})_{0\leqslant k\leqslant n}$. Its transition is given by $P_{T/n}(x, \mathrm{d}y)$, where $P_t(x, \mathrm{d}y)$ denotes the transition of the diffusion X. The discrete-time $(\mathcal{F}_{t_k})_{0\leqslant k\leqslant n}$-$(h, f)$-*Snell envelope* $(U_k)_{0\leqslant k\leqslant n}$ of $(X_{t_k})_{0\leqslant k\leqslant n}$ is defined by

$$U_k := \operatorname{ess\,sup}\left\{\mathbb{E}\left(h\left(\theta\frac{T}{n}, X_{\theta T/n}\right) + \frac{T}{n}\sum_{i=k+1}^{\theta} f(t_i, X_{t_i}, U_i)|\mathcal{F}_{t_k}\right), \theta\in\Theta_k\right\}, \tag{11}$$

where Θ_k denotes the set of $\{k, \dots, n\}$-valued $(\mathcal{F}_{t_k})_{0\leqslant k\leqslant n}$-stopping times (the index k stands for absolute discrete time in this formulation). When samples $(X_{t_1}, \dots, X_{t_n})$ can easily be simulated, $(U_k)_{0\leqslant k\leqslant n}$ becomes the quantity of interest. When one is dealing with the Snell envelope of the Euler scheme, (U_n) remains a tool in the error analysis.

The Gaussian Euler scheme with general step $\Delta > 0$ (here $\Delta = T/n$) is recursively defined by $\overline{X}_0^\Delta := X_0$ and,

$$\forall k\in\mathbb{N},\qquad \overline{X}_{k+1}^\Delta := \overline{X}_k^\Delta + \Delta b(\overline{X}_k^\Delta) + \sigma(\overline{X}_k^\Delta)\sqrt{\Delta}\,\varepsilon_{k+1}, \tag{12}$$

where $\varepsilon_k := (B_{k\Delta} - B_{(k-1)\Delta})/\sqrt{\Delta}$, $k\geqslant 1$, are i.i.d. $\mathcal{N}(0, I_q)$-distributed. The sequence $(\overline{X}_k^\Delta)_{0\leqslant k\leqslant n}$ is a homogeneous Markov chain with transition given on bounded Borel functions by

$$P^\Delta(f)(x) = \int_{\mathbb{R}^q} f(x + \Delta b(x) + \sqrt{\Delta}\sigma(x).u)\mathrm{e}^{-|u|^2/2}\frac{\mathrm{d}u}{(2\pi)^{q/2}}. \tag{13}$$

The discrete-time $(\mathcal{F}_{t_k})_{0\leqslant k\leqslant n}$-$(h, f)$-Snell envelope $(\overline{U}_k)_{0\leqslant k\leqslant n}$ of $(\overline{X}_k^\Delta)_{0\leqslant k\leqslant n}$ is

$$\overline{U}_k := \operatorname{ess\,sup}\left\{\mathbb{E}\left(h(\theta\Delta, \overline{X}_\theta) + \frac{T}{n}\sum_{i=k+1}^{\theta} f(t_i, \overline{X}_i, \overline{U}_i)/\mathcal{F}_{t_k}\right), \theta\in\Theta_k\right\}. \tag{14}$$

One crucial fact for our purpose is that both transitions of interest $P_\Delta(x, \mathrm{d}y)$ and $P^\Delta(x, \mathrm{d}y)$ are *Lipschitz* in the sense of the following definition.

Definition 1. *A transition* $(P(x, \mathrm{d}y))_{x\in\mathbb{R}^d}$ *is* K*-Lipschitz if,*

$$\forall g : \mathbb{R}^d \to \mathbb{R}, \textit{ Lipschitz continuous}, \qquad [Pg]_{\mathrm{Lip}} \leqslant K[g]_{\mathrm{Lip}}. \tag{15}$$

Proposition 2. *Assume that the drift* $b : \mathbb{R}^d \to \mathbb{R}^d$ *and the diffusion coefficient* $\sigma : \mathbb{R}^d \to \mathcal{M}(d \times q)$ *of the diffusion* X *are Lipschitz continuous. Set* $\Delta := T/n$.

(a) *Euler scheme. Then* P^Δ *is Lipschitz with ratio*

$$K_\Delta^{\mathrm{Euler}} = \sqrt{1 + \Delta\gamma_0(2 + \gamma_0(1 + \Delta))} = 1 + \Delta\gamma_0(1 + \gamma_0/2) + O(\Delta^2).$$

If, furthermore, b *and* σ *satisfy the so-called 'asymptotic flatness' assumption,*

$$\exists a > 0, \forall x, y \in \mathbb{R}^d, \qquad \frac{1}{2}\|\sigma(x) - \sigma(y)\|^2 + (x - y|b(x) - b(y)) \leqslant -a|x - y|^2,$$

then $K_\Delta^{\mathrm{Euler}} \leqslant \sqrt{1 - 2a\Delta + \Delta^2\gamma_0^2} = 1 - a\Delta + O(\Delta^2)$.

(b) *Diffusion. The transition* P_Δ *is Lipschitz with ratio*

$$K_\Delta^{\mathrm{diff}} = \exp(\gamma_0(1 + \gamma_0/2)\Delta).$$

If furthermore, the asymptotic flatness assumption holds, then

$$K_\Delta^{\mathrm{diff}} = \exp(-a\Delta).$$

Proof. (a) Let $g : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz continuous function. Then, for every $x, y \in \mathbb{R}^d$,

$$\begin{aligned}
|P^\Delta(g)(x) - P^\Delta(g)(y)|^2 &\leqslant \mathbb{E}(|g(x + \Delta b(x) + \sqrt{\Delta}\sigma(x)\varepsilon_1) - g(y + \Delta b(y) + \sqrt{\Delta}\sigma(y)\varepsilon_1)|^2) \\
&\leqslant [g]_{\mathrm{Lip}}^2\mathbb{E}(|x + \Delta b(x) + \sqrt{\Delta}\sigma(x)\varepsilon_1 - (y + \Delta b(y) + \sqrt{\Delta}\,\sigma(y)\varepsilon_1)|^2) \\
&\leqslant [g]_{\mathrm{Lip}}^2(|x - y|^2 + \Delta\|\sigma(x) - \sigma(y)\|^2 + 2\Delta(x - y|b(x) - b(y)) \\
&\quad + \Delta^2|b(x) - b(y)|^2) \\
&\leqslant [g]_{\mathrm{Lip}}^2(1 + \Delta\gamma_0^2 + 2\Delta\gamma_0 + \Delta^2\gamma_0^2)|x - y|^2.
\end{aligned}$$

The 'asymptotically flat' case is established similarly.

(b) Itô's formula implies (with obvious notation) that

$$\begin{aligned}|X_t^x - X_t^y|^2 = |x-y|^2 + 2\int_0^t \Big((X_s^x - X_s^y|b(X_s^x) - b(X_s^y)) + \frac{1}{2}\mathrm{tr}(\sigma(X_s^x) \\ - \sigma(X_s^y))(\sigma(X_s^x) - \sigma(X_s^y))^*\Big)\mathrm{d}s \\ + \int_0^t (X_s^x - X_s^y|(\sigma(X_s^x) - \sigma(X_s^y))\mathrm{d}B_s) \qquad \text{(true martingale)}\end{aligned}$$

$$\mathbb{E}|X_t^x - X_t^y|^2 \leqslant |x-y|^2 + \gamma_0(2+\gamma_0)\int_0^t \mathbb{E}|X_s^x - X_s^y|^2\mathrm{d}s.$$

Gronwall's lemma finally leads to $\mathbb{E}|X_t^x - X_t^y|^2 \leqslant |x-y|^2 \exp(\gamma_0(2+\gamma_0)t)$.

In the 'asymptotically flat' case, one verifies that $t \mapsto \mathbb{E}|X_t^x - X_t^y|^2$ is differentiable and satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}|X_t^x - X_t^y|^2 \leqslant -a\,\mathbb{E}|X_t^x - X_t^y|^2,$$

whence the result claimed. □

Remark 1. The result of the above proposition still holds if one considers an Euler scheme where the increments ε_k are simply square-integrable, centred and normalized.

Remark 2. The simplest 'asymptotically flat' transitions are the Euler scheme of the Ornstein–Uhlenbeck process $\mathrm{d}Y_t := -\frac{1}{2}Y_t\,\mathrm{d}t + \sigma\mathrm{d}B_t$ for which the property holds with $a = \frac{1}{4}$.

Bally and Pagès (2003) carry out an analysis of the L^p-error induced by considering the discrete-time (h, f)-Snell envelopes $(U_k)_{0\leqslant k\leqslant n}$ and $(\overline{U}_k)_{0\leqslant k\leqslant n}$ instead of the process $(Y_t)_{t\in[0,T]}$. The main result is summed up in the proposition below.

Proposition 3. *Assume that* (5)–(7) *hold and that* $X_0 = x \in \mathbb{R}^d$.

(*a*) *Lipschitz continuous setting. Let* $p \geqslant 1$. *Then*

$$\forall k \in \{0, \ldots, n\}, \qquad \|Y_{t_k} - \overline{U}_k\|_p + \|Y_{t_k} - U_k\|_p \leqslant C_p \mathrm{e}^{C_p T}(1+|x|)\frac{1}{\sqrt{n}},$$

where C_p *is a positive real constant depending upon* p, b, σ, f *and* h *(by means of* γ_0*).*

(*b*) *Semi-convex setting. Assume, furthermore, that* f *is* $\mathcal{C}_{\mathrm{b}}^{1,2,2}$ *(the set of* $\mathcal{C}^{1,2,2}$ *functions* f *whose existing partial derivatives are all bounded.) and that* h *is semi-convex in following sense:*

$$\forall t \in [0, T], \forall x, y \in \mathbb{R}^d, \qquad h(t, y) - h(t, x) \geqslant (\delta_h(t, x)|y - x) - \rho|x-y|^2, \qquad (16)$$

where δ_h *is a bounded function on* $[0, T] \times \mathbb{R}^d$ *and* $\rho \geqslant 0$. *Then, the* (h, f)*-Snell envelope of the discretized diffusion* $(X_{t_k})_{0\leqslant k\leqslant n}$ *satisfies, for every* $p \geqslant 1$,

$$\forall k \in \{0, \ldots, n\}, \qquad \|Y_{t_k} - U_k\|_p \leqslant C_p \mathrm{e}^{C_p T}(1+|x|)\frac{1}{n}.$$

(The real constant C_p *is a a priori different from that in the Lipschitz continuous case.)*

Remark 3. The semi-convexity assumption is a generalization of convexity which embodies smooth enough functions. This notion seems to have been introduced in Caverhill and Webber (1990) for pricing one-dimensional American options. See also Lamberton (2002) for recent developments in finance.

Remark 4. If $h(t, \cdot)$ is convex for every $t \in [0, T]$ with a bounded spatial derivative $\delta_h(t, \cdot)$ (in the distribution sense), then h is semi-convex with $\rho = 0$. Thus, it embodies most American style pay-off functions used in mathematical finance for options pricing like those involving the positive part of linear combinations or extrema of the underlying traded asset).

Remark 5. If $h(t, \cdot)$ is $\mathcal{C}^1$ for every $t \in \mathbb{R}_+$ and $\partial h(t, x)/\partial x$ *is* ρ-Lipschitz in x, uniformly in t, then h is semi-convex (with $\delta_h(t, x) := \partial h(t, x)/\partial x$).

2.2. Quantization of the (h, f)-Snell envelope of a Lipschitz Markov chain

Let $(X_k)_{k\in\mathbb{N}}$ be a homogeneous $\mathbb{R}^d$-valued $(\mathcal{F}_k)_{k\in\mathbb{N}}$-Markov chain with transition $P(x, \mathrm{d}y)$. Motivated by the Section 2.1, one is interested in computing the following (h, f)-Snell envelope $(U_k)_{0\leqslant k\leqslant n}$ related to a finite horizon n and to some functions $h := (h_k)_{0\leqslant k\leqslant n}$ and $f := (f_k)_{0\leqslant k\leqslant n}$ defined on $\{0, \ldots, n\} \times \mathbb{R}^d$ and $\{0, \ldots, n\} \times \mathbb{R}^d \times \mathbb{R}^d$, respectively:

$$U_k := \operatorname{ess\,sup}\left\{\mathbb{E}\left(h_\theta(X_\theta) + \sum_{i=k+1}^{\theta} f_i(X_i, U_i)|\mathcal{F}_k\right), \theta\{k, \ldots, n\}\text{-valued } \mathcal{F}_l\text{-stopping time}\right\}. \tag{17}$$

In fact, the (h, f)-Snell envelope is simply connected with the regular Snell envelope appearing in optimal stopping theory: one verifies that $V_k := U_k + \sum_{i=1}^k f_i(X_i, U_i)$ is the standard Snell envelope of the $\mathcal{F}_k$-adapted sequence $Z_k := h_k(X_k) + \sum_{i=1}^k f_i(X_i, U_i)$. Hence, following Neveu (1971), for example, V is the Snell envelope of Z, that is, it satisfies the backward dynamic programming principle

$$V_n = Z_n \quad \text{and} \quad V_k = \max(Z_k, \mathbb{E}(V_{k+1}|\mathcal{F}_k)).$$

$(U_k)_{0\leqslant k\leqslant n}$ is found to satisfy the dynamic programming principle

$$\begin{aligned} U_n &:= h_n(X_n), \\ U_k &:= \max(h_k(X_k), \mathbb{E}(U_{k+1} + f_{k+1}(X_{k+1}, U_{k+1})|\mathcal{F}_k)), \qquad 0 \leqslant k \leqslant n-1. \end{aligned} \tag{18}$$

It is thus the (h, f)-*Snell envelope* of X. Henceforth $\mathbb{E}(\cdot|\mathcal{F}_k)$ will be written simply as $\mathbb{E}_k(\cdot)$.

At this stage, a straightforward induction using the Markov property shows that, for every $k \in \{0, \ldots, n\}$, $U_k = u_k(X_k)$, where, u_k is recursively defined by

$$\begin{aligned} u_n &:= h_n, \\ u_k &:= \max(h_k, P(u_{k+1} + f_{k+1}(\cdot, u_{k+1}))), \qquad 0 \leqslant k \leqslant n-1. \end{aligned} \tag{19}$$

Example. The time discretization of an RBSDE corresponds to functions

$$f_k(x, u) := \frac{T}{n} f\left(\frac{kT}{n}, x, u\right) \quad \text{and} \quad h_k(x) := h\left(\frac{kT}{n}, x\right), \qquad 0 \leqslant k \leqslant n. \tag{20}$$

1.2.1. The quantization tree algorithm: a pseudo-Snell envelope

The starting point of the method is to discretize at every step $k \in \{0, \dots, n\}$ the random vector X_k using a $\sigma(X_k)$-measurable random vector $\hat{X}_k$ that takes finitely many values. The random variable $\hat{X}_k$ is called a *quantization* of X_k. One may always associate with $\hat{X}_k$ a Borel function $q_k : \mathbb{R}^d \to \mathbb{R}^d$ such that $\hat{X}_k = q_k(X_k)$. The function q_k is often called a *quantizer* (this terminology comes from signal processing and information theory; see Section 3.1). It will be convenient to call the finite subset $\Gamma_k := q_k(\mathbb{R}^d) = \hat{X}_k(\Omega)$ a *quantization grid*, or simply *grid*. The size of a grid Γ_k will be denoted by N_k. The elements of a quantization grid are called *elementary quantizers*. We will denote by $N := N_0 + N_1 + \dots + N_n$ the total number of elementary quantizers used to quantize all the X_k, $0 \leqslant k \leqslant n$.

We now wish to approximate the Snell envelope $(U_k)_{0 \leqslant k \leqslant n}$ by a sequence $(\hat{U}_k)_{0 \leqslant k \leqslant n}$ formally defined by a dynamic programming algorithm similar to (18), except that the random vector X_k is replaced by its quantization $\hat{X}_k$, for every $k \in \{0, \dots, n\}$, and the conditional expectation $\mathbb{E}_k$ (i.e the past of the filtration $\mathcal{F}$ up to time k) is replaced by the conditional expectation given $\hat{X}_k$ *i.e.* $\mathbb{E}(\cdot|\hat{X}_k)$. Henceforth, for the sake of simplicity, $\mathbb{E}(\cdot|\hat{X}_k)$ will denoted $\hat{\mathbb{E}}_k(\cdot)$.

Assume temporarily that, for every $k \in \{0, 1, \dots, n\}$, we have access to an appropriate quantization $\hat{X}_k = q_k(X_k)$ of X_k. The optimal choice of the grid Γ_k and the quantizer q_k that yield the best possible approximation will be investigated in Section 3.1.

Thus, the *pseudo-Snell envelope* is defined by mimicking the original one (18) as follows:

$$\hat{U}_n := h_n(\hat{X}_n),$$

$$\hat{U}_k := \max(h_k(\hat{X}_k), \hat{\mathbb{E}}_k(\hat{U}_{k+1} + f_{k+1}(\hat{X}_{k+1}, \hat{U}_{k+1})), \qquad 0 \leqslant k \leqslant n-1. \tag{21}$$

The main reason for considering conditional expectation with respect to $\hat{X}_k$ is that the sequence $(\hat{X}_k)_{k \in \mathbb{N}}$ does not satisfy the Markov property. The *quantization tree algorithm* then simply consists in rewriting the pseudo-Snell envelope in distribution.

Proposition 4 (Quantization tree algorithm). *For every $k \in \{0, \dots, n\}$, let $\Gamma^k := \{x_1^k, \dots, x_{N_k}^k\}$ denote a quantization grid of the distribution $\mathcal{L}(X_k)$ and q_k its quantizer. For every $k \in \{0, \dots, n-1\}$, $i \in \{1, \dots, N_k\}$, $j \in \{1, \dots, N_{k+1}\}$*

$$\pi_{ij}^k := \mathbb{P}(\hat{X}_{k+1} = x_j^{k+1} | \hat{X}_k = x_i^k). \tag{22}$$

One defines the functions $\hat{u}_k$ by the backward induction

$$\hat{u}_n(x_i^n) := h_n(x_i^n), \qquad i \in \{0, \ldots, N_n\},$$

$$\hat{u}_k(x_i^k) := \max\left(h_k(x_i^k), \sum_{j=1}^{N_{k+1}} \pi_{ij}^k \Big(\hat{u}_{k+1}(x_j^{k+1}) + f_{k+1}(x_j^{k+1}, \hat{u}_{k+1}(x_j^{k+1})) \Big) \right), \tag{23}$$

$$k \in \{0, \ldots, n-1\}, \qquad i \in \{1, \ldots, N_k\}.$$

Then $\hat{u}_k(\hat{X}_k) = \hat{U}_k$, $0 \leqslant k \leqslant n$, *is the pseudo-Snell envelope defined by* (21).

Note that if $\mathcal{L}(X_0) := \delta_{x_0}$, then $\hat{u}_0(\hat{X}_0) = \hat{u}_0(x_0)$ is deterministic, otherwise

$$\mathbb{E}\,\hat{u}_0(\hat{X}_0) = \sum_{i=1}^{N_0} p_i^0\,\hat{u}_0(x_i^0) \qquad \text{with} \quad p_i^0 := \mathbb{P}(\hat{X}_0 = x_i^0),\ 1 \leqslant i \leqslant N_0.$$

Implementing procedure (23) on a computer raises two questions. The first is how to estimate numerically the above coefficients π_{ij}^k. The second is whether the complexity of the quantization tree algorithm is acceptable.

As far as practical implementation is concerned, the ability to compute the π_{ij}^k (and the p_i^0) at a reasonable cost is the key to the whole method. The most elementary solution is simply to process a wide-scale *Monte Carlo simulation of the Markov chain* $(X_k)_{0 \leqslant k \leqslant n}$ (see Section 3.3). The estimation of the coefficients is based on the representation formula (22) of π_{ij}^k as expectations of simple functions of (X_k, X_{k+1}). Furthermore, the a priori error bounds for $\|U_k - \hat{U}_k\|_p$ that will be derived in Theorem 2 below all rely on the L^p-*quantization errors* $\|X_k - \hat{X}_k\|_p$, $0 \leqslant k \leqslant n$, which can be simultaneously approximated. So, the parameters of interest can be evaluated provided that independent paths of the Markov chain $(X_k)_{0 \leqslant k \leqslant n}$ can be simulated at a reasonable cost. This amounts *to the efficient simulation of some* $P(x, \mathrm{d}y)$*-distributed random vectors for every* $x \in \mathbb{R}^d$.

We will see in Section 3.3.1 that this first approach can be improved by combining this Monte Carlo simulation with the grid optimization procedure.

Turning now to the complexity of the algorithm, a quick look at the structure of the quantization tree algorithm (23) shows that going from layer $k+1$ down to layer k requires $\kappa N_k N_{k+1}$ elementary computations (where $\kappa > 0$ denotes the average number of computations per link '$i \to j$'). Hence, the cost of completing the tree descent is

$$\text{Complexity} = \kappa(N_0 N_1 + \ldots + N_k N_{k+1} + \ldots + N_{n-1} N_n),$$

so that

$$\kappa \frac{n}{(n+1)^2} N^2 \leqslant \text{ Complexity } \leqslant \kappa \frac{N^2}{4}.$$

The lower bound holds for $N_k = N/(n+1), 0 \leqslant k \leqslant n$, the upper one for $N_0 = N_1 = N/2$, $N_k = 0$, $2 \leqslant k \leqslant n$, which is clearly unrealistic. The optimal dispatching (see, for example, the practical comments in Section 6.1) leads to a complexity close to the lower bound.

However this is a very pessimistic analysis of the complexity. In fact, in most examples –

such as the Euler scheme – the Markov transition $P(x, \mathrm{d}y)$ is such that, at each step k, most coefficients of the quantized transition matrix $[\pi_{ij}^k]$ are so small that their estimates produced by the Monte Carlo simulation turn out to be 0. This is taken into account to speed up the computer procedure so that the practical complexity of the algorithm is $O(N)$. This can be compared to the complexity of a Cox-Ross-Rubinstein one-dimensional binomial tree with $\sqrt{2N}$ time steps which contains approximately N points.

2.2.2. Convergence and rate

The aim now is to provide some a priori L^p-error bounds for $\|U_k - \hat{U}_k\|_p$, $0 \leqslant k \leqslant n$, based on the L^p-quantization errors $\|X_k - \hat{X}_k\|_p$, $0 \leqslant k \leqslant n$ (keeping in mind that these quantities can simply be estimated during the Monte Carlo simulation of the chain).

The main necessary assumption on the Markov chain here is that its transition $P(x, \mathrm{d}y)$ is *Lipschitz* (see Definition 1). This assumption is natural, as emphasized by the above Proposition 2: the transitions of a diffusion and of its Euler scheme are both Lipschitz if its coefficients are Lipschitz continuous. The first task is to evaluate the Lipschitz regularity of the functions u_k defined by (19) in that setting.

Proposition 5. *Assume that the functions h and f are Lipschitz continuous, uniformly with respect to k, that is, for every* $k \in \{0, \ldots, n\}$,

$$\forall x, x' \in \mathbb{R}^d, \qquad |h_k(x) - h_k(x')| \leqslant [h]_{\mathrm{Lip}} |x - x'|, \tag{25}$$

$$\forall x, x' \in \mathbb{R}^d, \forall u, u' \in \mathbb{R}, \qquad |f_k(x, u) - f_k(x', u')| \leqslant [f]_{\mathrm{Lip}} (|x - x'| + |u - u'|) \tag{26}$$

If the transition P is K-Lipschitz, then the functions u_k defined by (19) *are Lipschitz continuous. Furthermore, setting $L := K(1 + [f]_{\mathrm{Lip}})$, one obtains*

$$[u_k]_{\mathrm{Lip}} \leqslant \begin{cases} L^{n-k}\left([h]_{\mathrm{Lip}} + \dfrac{K}{L-1}[f]_{\mathrm{Lip}}\right), & \textit{if } L > 1, \\ [h]_{\mathrm{Lip}} + (n-k)[f]_{\mathrm{Lip}} & \textit{if } L = 1, \\ \max\left([h]_{\mathrm{Lip}}, \dfrac{K}{1-L}[f]_{\mathrm{Lip}}\right), & \textit{if } L < 1. \end{cases}$$

Remark 6. If $f \equiv 0$ (i.e. regular optimal stopping), the above inequalities read as follows

$$[u_k]_{\mathrm{Lip}} \leqslant (K \vee 1)^{n-k} [h]_{\mathrm{Lip}}.$$

For practical applications, for example to the Euler scheme or to simulatable diffusions, $L \sim 1 + c/n$ so the coefficient L^{n-k} does not explode as the number n of time steps goes to infinity.

Proof. As $u_n = h_n$, $[u_n]_{\mathrm{Lip}} \leqslant [h]_{\mathrm{Lip}}$. Then, using the inequality

$$|\max(a, b) - \max(a', b')| \leqslant \max(|a - a'|, |b - b'|),$$

it easily follows from the dynamic programming equality (19) that

$$\begin{aligned}[u_k]_{\mathrm{Lip}} &\leqslant \max([h]_{\mathrm{Lip}}, [P(u_{k+1} + f_{k+1}(\cdot, u_{k+1}))]_{\mathrm{Lip}}) \\ &\leqslant \max([h]_{\mathrm{Lip}}, K([u_{k+1}]_{\mathrm{Lip}} + [f]_{\mathrm{Lip}}(1 + [u_{k+1}]_{\mathrm{Lip}}))) \\ &\leqslant \max([h]_{\mathrm{Lip}}, L[u_{k+1}]_{\mathrm{Lip}} + K[f]_{\mathrm{Lip}}).\end{aligned}$$

By induction,

$$\begin{aligned}[u_k]_{\mathrm{Lip}} &\leqslant L^{-k} \max_{k \leqslant i \leqslant n} (L^i [h]_{\mathrm{Lip}} + (L^k + \ldots + L^{i-1})[f]_{\mathrm{Lip}}) \\ &\leqslant \max_{0 \leqslant j \leqslant n-k} (L^j [h]_{\mathrm{Lip}} + \frac{L^j - 1}{L - 1} K [f]_{\mathrm{Lip}}) \qquad (\text{if } L \neq 1),\end{aligned}$$

and we are done. □

We now move on to the main result of this section: some a priori estimates for $\|U_k - \hat{U}_k\|_p$ as a function of the quantization error $\|X_k - \hat{X}_k\|_p$.

Theorem 2. *Assume that the transition* $(P(x, \mathrm{d}y))_{x \in \mathbb{R}^d}$ *is K-Lipschitz and that the functions h and f satisfy the Lipschitz assumptions* (25) *and* (26). *Then,*

$$\forall p \geqslant 1, \forall k \in \{0, \ldots, n\}, \qquad \|U_k - \hat{U}_k\|_p \leqslant \frac{1}{(1 + [f]_{\mathrm{Lip}})^k} \sum_{i=k}^{n} d_i \|X_i - \hat{X}_i\|_p,$$

with

$$\begin{aligned} d_i &:= ([h]_{\mathrm{Lip}} + [f]_{\mathrm{Lip}} + (2 - \delta_{2,p}) K(([u_{i+1}]_{\mathrm{Lip}} + 1)([f]_{\mathrm{Lip}} + 1) - 1))(1 + [f]_{\mathrm{Lip}})^i, \\ &\quad 0 \leqslant i \leqslant n - 1, \\ d_n &:= ([h]_{\mathrm{Lip}} + [f]_{\mathrm{Lip}})(1 + [f]_{\mathrm{Lip}})^n. \end{aligned} \tag{27}$$

Proof. Set $\Phi_k := P(u_{k+1} + f_{k+1}(., u_{k+1}))$, $k = 0, \ldots, n-1$, and $\Phi_n \equiv 0$ so that $\mathbb{E}(U_{k+1} + f_{k+1}(X_{k+1}, U_{k+1}) | X_k) = \Phi_k(X_k)$. Define $\hat{\Phi}_k$ similarly by the equality $\hat{\Phi}_k(\hat{X}_k) := \hat{\mathbb{E}}_k(\hat{U}_{k+1} + f_{k+1}(\hat{U}_{k+1}, \hat{X}_{k+1}))$ and $\hat{\Phi}_n \equiv 0$. Then

$$\begin{aligned}|U_k - \hat{U}_k| &\leqslant |h_k(X_k) - h_k(\hat{X}_k)| + |\Phi_k(X_k) - \hat{\Phi}_k(\hat{X}_k)| \\ &\leqslant [h]_{\mathrm{Lip}} |X_k - \hat{X}_k| + |\Phi_k(X_k) - \hat{\mathbb{E}}_k(\Phi_k(X_k))| + |\hat{\mathbb{E}}_k(\Phi_k(X_k)) - \hat{\Phi}_k(\hat{X}_k)|.\end{aligned}$$

Now

$$\begin{aligned}|\Phi_k(X_k) - \hat{\mathbb{E}}_k \Phi_k(X_k)| &\leqslant |\Phi_k(X_k) - \Phi_k(\hat{X}_k)| + \hat{\mathbb{E}}_k |\Phi_k(X_k) - \hat{\mathbb{E}}_k(\Phi_k(\hat{X}_k))| \\ &\leqslant [\Phi_k]_{\mathrm{Lip}}(|X_k - \hat{X}_k| + \hat{\mathbb{E}}_k |X_k - \hat{X}_k|).\end{aligned}$$

(Note that $\hat{X}_k$ is $\mathcal{F}_k$-measurable.) Hence,

$$\|\Phi_k(X_k) - \hat{\mathbb{E}}\Phi_k(X_k)\|_p \leqslant 2[\Phi_k]_{\mathrm{Lip}}\|X_k - \hat{X}_k\|_p.$$

When $p = 2$, one may drop the factor 2 since the very definition of the conditional expectation as a projection in a Hilbert space implies that

$$\|\Phi_k(X_k) - \hat{\mathbb{E}}\Phi_k(X_k)\|_2 \leqslant \|\Phi_k(X_k) - \Phi_k(\hat{X}_k)\|_2 \leqslant [\Phi_k]_{\mathrm{Lip}}\|X_k - \hat{X}_k\|_2.$$

On the other hand, coming back to the definition of $\Phi_k(X_k)$ and $\hat{\Phi}_k(\hat{X}_k)$, one obtains, using the fact that $\hat{\mathbb{E}}_k \circ \mathbb{E}_k = \hat{\mathbb{E}}_k$ and that conditional expectation is an L^p-contraction,

$$|\hat{\mathbb{E}}_k(\Phi_k(X_k)) - \hat{\Phi}_k(\hat{X}_k)| \leqslant \hat{\mathbb{E}}_k|U_{k+1} + f_{k+1}(X_{k+1}, U_{k+1}) - \hat{U}_{k+1} - f_{k+1}(\hat{X}_{k+1}, \hat{U}_{k+1})|$$

$$\begin{aligned}\|\hat{\mathbb{E}}_k(\Phi_k(X_k)) - \hat{\Phi}_k(\hat{X}_k)\|_p &\leqslant \|U_{k+1} + f_{k+1}(X_{k+1}, U_{k+1}) - \hat{U}_{k+1} - f_{k+1}(\hat{X}_{k+1}, \hat{U}_{k+1})\|_p \\ &\leqslant (1 + [f]_{\mathrm{Lip}})\|U_{k+1} - \hat{U}_{k+1}\|_p + [f]_{\mathrm{Lip}}\|X_{k+1} - \hat{X}_{k+1}\|_p.\end{aligned}$$

Finally, for every $k \in \{0, \dots, n-1\}$,

$$\begin{aligned}\|U_k - \hat{U}_k\|_p &\leqslant [h]_{\mathrm{Lip}}\|X_k - \hat{X}_k\|_p + \|\Phi_k(X_k) - \hat{\mathbb{E}}_k(\Phi_k(X_k))\|_p + \|\Phi_k(X_k) - \hat{\Phi}_k(\hat{X}_k)\|_p \\ &\leqslant (1 + [f]_{\mathrm{Lip}})\|U_{k+1} - \hat{U}_{k+1}\|_p + ([h]_{\mathrm{Lip}} + (2 - \delta_{p,2})[\Phi_k]_{\mathrm{Lip}})\|X_k - \hat{X}_k\|_p \\ &\quad + [f]_{\mathrm{Lip}}\|X_{k+1} - \hat{X}_{k+1}\|_p.\end{aligned}$$

Using the inequality $\|U_n - \hat{U}_n\|_p \leqslant [h]_{\mathrm{Lip}}\|X_n - \hat{X}_n\|_p$, standard computations yield

$$\begin{aligned}\|U_k - \hat{U}_k\|_p &\leqslant \sum_{i=k}^{n}\left([h]_{\mathrm{Lip}} + (2 - \delta_{p,2})[\Phi_i]_{\mathrm{Lip}} + \frac{[f]_{\mathrm{Lip}}}{1 + [f]_{\mathrm{Lip}}}\right)(1 + [f]_{\mathrm{Lip}})^{i-k}\|X_i - \hat{X}_i\|_p \\ &\quad - [f]_{\mathrm{Lip}}\|X_k - \hat{X}_k\|_p \\ &\leqslant \frac{1}{(1 + [f]_{\mathrm{Lip}})^k}\sum_{i=k}^{n}(1 + [f]_{\mathrm{Lip}})^i\left([h]_{\mathrm{Lip}} + \frac{[f]_{\mathrm{Lip}}}{1 + [f]_{\mathrm{Lip}}} + (2 - \delta_{p,2})[\Phi_i]_{\mathrm{Lip}}\right) \\ &\quad \times \|X_i - \hat{X}_i\|_p.\end{aligned}$$

Finally, the definition of Φ_i and the Lipschitz property of $P(x, \mathrm{d}y)$ imply that

$$[\Phi_i]_{\mathrm{Lip}} = [P(u_{i+1} + f_{i+1}(\cdot, u_{i+1}))]_{\mathrm{Lip}} \leqslant K(1 + [f]_{\mathrm{Lip}})[u_{i+1}]_{\mathrm{Lip}} + K[f]_{\mathrm{Lip}}, 1 \leqslant i \leqslant n-1.$$

□

2.2.3. *Approximation of the (lowest) optimal stopping time*

The second quantity of interest in optimal stopping theory is the (set of) optimal stopping time(s). A stopping time τ_{opt} is optimal for the (h, f)-Snell envelope if

$$U_0 = \mathbb{E}_0\left(h_{\tau_{\mathrm{opt}}}(X_{\tau_{\mathrm{opt}}}) + \sum_{i=1}^{\tau_{\mathrm{opt}}} f_i(X_i, U_i)\right).$$

We know (see, Neveu, 1971) that the lowest optimal stopping time is given by

$$\tau_* := \min\{k | U_k = h_k(X_k)\}.$$

In the case of non-uniqueness of the optimal stopping times, τ_* plays a special role because it turns out to be the easiest to approximate. Thus when dealing with quantization of Markov chains, it is natural to introduce its counterpart for the quantized process, that is,

$$\hat{\tau}_* := \min\big\{k | \hat{u}_k(\hat{X}_k) = h_k(\hat{X}_k)\big\}.$$

In fact, the estimation of the error $\|\tau_* - \hat{\tau}_*\|_p$ seems out of reach, essentially because it is quite difficult to bound these stopping times from below. Nevertheless, we will be able to approximate τ_* (in probability). Let $\delta > 0$. Set

$$\tau_\delta := \min\{k | u_k(X_k) \leqslant h_k(X_k) + \delta\},$$

$$\hat{\tau}_\delta := \min\big\{k | \hat{u}_k(\hat{X}_k) \leqslant h_k(\hat{X}_k) + \delta\big\}.$$

Proposition 6. (*a*) *For every* $\delta > 0$, $\tau_\delta \geqslant \tau_*$, $\hat{\tau}_\delta \geqslant \hat{\tau}_*$. *Furthermore*

$$\tau_\delta \downarrow \tau_* \quad \textit{and} \quad \hat{\tau}_\delta \downarrow \hat{\tau}_* \quad \textit{as} \quad \delta \downarrow 0.$$

(These stopping times being integer-valued, τ_δ *and* $\hat{\tau}_\delta$ *are eventually equal to their limit.)*

(*b*) *For every* $\delta > 0$,

$$\mathbb{P}(\hat{\tau}_\delta \notin [\tau_{3\delta/2}, \tau_{\delta/2}]) \leqslant \tfrac{1}{\delta} \sum_{k=0}^{n} (kd_k + [h]_{\mathrm{Lip}}) \|\hat{X}_k - X_k\|_1.$$

Proof. Part (*a*) is an obvious corollary of the definitions of τ_δ and $\hat{\tau}_\delta$.

(*b*) Set $Z_k := u_k(X_k) - \hat{u}_k(\hat{X}_k) + h_k(\hat{X}_k) - h_k(X_k)$. Then, we may write

$$\hat{\tau}_\delta = \min\{k | u_k(X_k) \leqslant h_k(X_k) + \delta + Z_k\}$$

so that, on the event $\{\max_{0 \leqslant k \leqslant n} |Z_k| \leqslant \delta/2\}$, $\tau_{3\delta/2} \leqslant \hat{\tau}_\delta \leqslant \tau_{\delta/2}$. Subsequently,

$$\mathbb{P}(\hat{\tau}_\delta \notin [\tau_{3\delta/2}, \tau_{\delta/2}]) \leqslant \mathbb{P}(\max_{0 \leqslant k \leqslant n} |Z_k| > |\frac{\delta}{2}) \leqslant \frac{2}{\delta} \,\mathbb{E} \max_{\leqslant k \leqslant n} |Z_k|.$$

Now, using the notation of Theorem 2, one has

$$|Z_k| \leqslant |\hat{U}_k - U_k| + [h]_{\mathrm{Lip}} |\hat{X}_k - X_k|$$

and thus

$$\begin{aligned} \mathbb{E} \max_{0 \leqslant k \leqslant n} |Z_k| &\leqslant \sum_{k=0}^{n} \|\hat{U}_k - U_k\|_1 + [h]_{\mathrm{Lip}} \|\hat{X}_k - X_k\|_1 \\ &\leqslant \sum_{k=0}^{n} (kd_k + [h]_{\mathrm{Lip}}) \|\hat{X}_k - X_k\|_1. \end{aligned}$$

□

3. Optimization of the quantization

After some brief background material on optimal quantization of a $\mathbb{R}^d$-valued random vector, this section is devoted to the optimal quantization method of a Markov chain. For a modern and rigorous overview of quantization of random vectors, see Graf and Luschgy (2000) and the references therein.

3.1. Optimal quantization of a random vector X

Let $X \in L^p_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$. Following the terminology introduced in Section 2.2.1, the L^p-quantization ($p \geqslant 1$) consists in studying the best possible L^p-approximation of X by a random vector $\hat{X} := q(X)$, where $q : \mathbb{R}^d \to \mathbb{R}^d$ is a Borel function (*quantizer*) taking at most N values called *elementary quantizers*. Set the quantization grid $\Gamma := q(\mathbb{R}^d) := \{x_1, \ldots, x_N\}$, $x_1, \ldots, x_N \in \{\mathbb{R}\}^d$. Minimizing the L^p-quantization error $\|X - q(X)\|_p$ consists in two phases:

1. Having set a grid $\Gamma \subset (\mathbb{R}^d)^N$, $|\Gamma| \leqslant \mathrm{N}$, find a/the Γ-valued quantizer q_Γ that minimizes $\|X - p_\Gamma(X)\|_p$ among all Γ-valued quantizers q (if any).

2. Find a grid Γ of size $|\Gamma| \leqslant \mathrm{N}$ which achieves the infimum of $\|X - q_\Gamma(X)\|_p$ among all the grids having at most N points (if any).

The solution to phase 1 is provided by any *Voronoi quantizer* of the grid Γ, also called a *projection following the closest-neighbour rule* and defined by

$$q_\Gamma := \sum_{x_i \in \Gamma} x_i \mathbf{1}_{C_i(\Gamma)},$$

where $(C_i(\Gamma))_{1 \leqslant i \leqslant N}$ is a Borel partition of $\mathbb{R}^d$ called a *Voronoi tessellation*, satisfying

$$C_i(\Gamma) \subset \{\xi \in \mathbb{R}^d | |x_i - \xi| = \min_{1 \leqslant j \leqslant N} |\xi - x_j|\}.$$

A given grid of size $N \geqslant 2$ clearly has infinitely many Voronoi tessellations, essentially due to median hyperplanes. However, all the $C_i(\Gamma)$ have the same *convex* closure and boundary, included in at most $N - 1$ hyperplanes. If the distribution of X weights no hyperplane, that is, $\mathbb{P}(X \in H) = 0$ for any hyperplane H, then the Voronoi tessellation is $\mathbb{P}$-essentially unique.

The Voronoi Γ-quantization, denoted by $\hat{X}^\Gamma := q_\Gamma(X)$, induces an L^p*-quantization error* $\|X - \hat{X}^\Gamma\|_p$ (in information theory $\|X - \hat{X}^\Gamma\|_p^p$ is called L^p*-distortion*) given by

$$\|X - \hat{X}^\Gamma\|_p^p = \sum_{x_i \in \Gamma} \mathbb{E}\big(\mathbf{1}_{C_i(\Gamma)} |X - x_i|^p\big) = \mathbb{E}\Big(\min_{1 \leqslant i \leqslant N} |X - x_i|^p\Big) = \int_{\mathbb{R}^d} \min_{1 \leqslant i \leqslant N} |\xi - x_i|^p \, \mathbb{P}_X(\mathrm{d}\xi). \tag{28}$$

Notice that the quantization error only depends on the distribution of X, whereas the Voronoi quantizer q_Γ only depends on Γ (and the Euclidean norm). Equality (28) will be the key to the numerical optimization of the grid. Finally, one can easily show that

$$\|X - \hat{X}^{\Gamma}\|_p = \min\{\|X - Y\|_p,\, Y : \Omega \to \Gamma,\, |Y(\Omega)| \leqslant N\}. \tag{29}$$

To carry out phase 2 (grid optimization), one derives from (28) that the quantization error $\|X - \hat{X}^{\Gamma}\|_p$ behaves as *symmetric Lipschitz continuous function* of the components of the grid $\Gamma := \{x_1, \ldots, x_{\scriptscriptstyle N}\}$ (with the temporary convention that some elementary quantizers x_i may be 'stuck' so that $|\Gamma| \leqslant N$). One shows (see Abaya and Wise 1992; Pagès 1997; Graf and Luschgy 2000) that

$$\Gamma \mapsto \|X - \hat{X}^{\Gamma}\|_p,\, |\Gamma| \leqslant N, \text{ always reaches a minimum}$$

at some grid Γ^* which takes its values in the convex hull of the support of $\mathbb{P}_{\scriptscriptstyle X}$. One proceeds by induction. If $N = 1$, the existence of a minimum is obvious by convexity; then, one may assume without loss of generality that $|X(\Omega)| \geqslant N$. Then $\{\Gamma|\, \|X - \hat{X}^{\Gamma}\|_p \leqslant m_{N-1} - \varepsilon,$ $|\Gamma| \leqslant N\}$, with $m_{N-1} := \min\{\|X - \hat{X}^{\Gamma}\|_p,\ |\Gamma| \leqslant N-1\} = \|X - \hat{X}^{\Gamma^{*},N-1}\|_p$, is a non-empty compact set for small enough $\varepsilon > 0$ since it contains $\Gamma^{*,N-1} \cup \{\xi\}$ for some appropriate $\xi \in X(\Omega)\backslash\Gamma^{*,N-1}$. This implies the existence of an optimal grid $\Gamma^{*,N}$. Then, following (29), $\hat{X}^{\Gamma^{*},N}$ is the best L^p-approximation of X over the random vectors taking at most N values, that is,

$$\|X - \hat{X}^{\Gamma^{*,N}}\|_p = \min\Big\{\|X - Y\|_p,\, Y : \Omega \to \mathbb{R}^d,\, |Y(\Omega)| \leqslant N\Big\}. \tag{30}$$

As an example, an optimal L^2-quantization of the normal distribution is given in Figure 1.

We will need the following properties (see Pagès 1997; Graf and Luschgy 2000; and references therein).

Property 1. *If $\mathbb{P}_{\scriptscriptstyle X}$ has an infinite support, any 'N-optimal' grid Γ^* has N pairwise distinct components, $\mathbb{P}_{\scriptscriptstyle X}(\cup_{i=1}^N \partial C_i(\Gamma^*)) = 0$ and $N \mapsto \min_{|\Gamma|\leqslant N}\|X - \hat{X}^{\Gamma}\|_{\scriptscriptstyle p}$ is decreasing.*

Property 2. *If the support of $\mathbb{P}_{\scriptscriptstyle X}$ is everywhere dense in its convex hull $\mathcal{H}_X$, any N-optimal grid lies in $\mathcal{H}_X$ and any locally N-optimal grid lying in $\mathcal{H}_X$ has exactly N distinct components.*

Property 3. *The minimal L^p-quantization error goes to 0 as $N \to \infty$:*

$$\lim_N \min_{|\Gamma|\leqslant N} \|X - \hat{X}^{\Gamma}\|_p = 0.$$

As a matter of fact, set $\Gamma_{\scriptscriptstyle N} := \{z_1, \ldots, z_{\scriptscriptstyle N}\}$ where $(z_k)_{k\in\mathbb{N}}$ is everywhere dense in $\mathbb{R}^d$. Then, $\min_{|\Gamma|\leqslant N}\|X - \hat{X}^{\Gamma}\|_p \leqslant \|X - \hat{X}^{\Gamma_N}\|_p$ goes to 0 by the Lebesgue dominated convergence theorem.

The rate of this convergence to zero turns out to be a much more challenging problem. The solution, often referred to as Zador's theorem, was completed by several authors (Zador 1982; Bucklew and Wise 1982; Graf and Luschgy 2000).

Theorem 3 (Asymptotics). *If $\mathbb{E}|X|^{p+\eta} < +\infty$ for some $\eta > 0$, then*

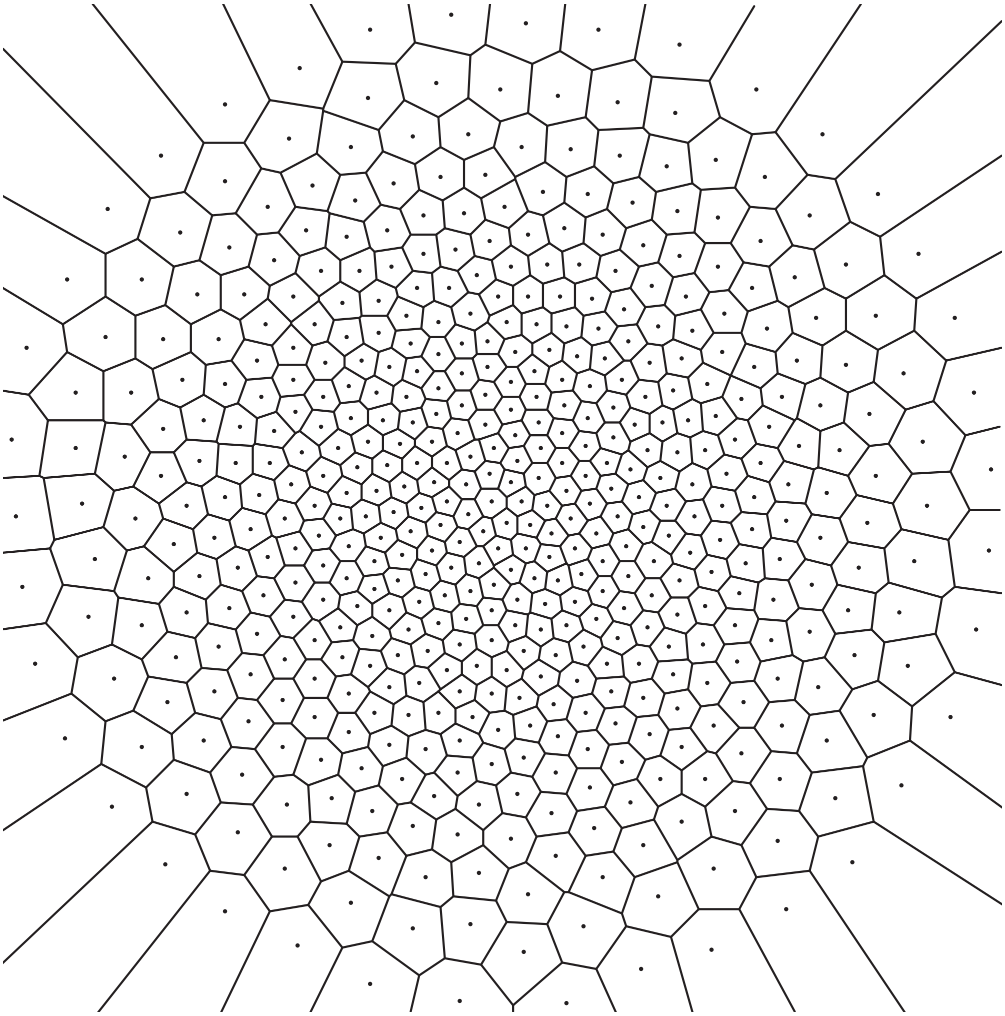

Figure 1. Optimal L^2-quantization of the normal distribution $\mathcal{N}(0, I_2)$ with a 500-tuple and its Voronoi tessellation

$$\lim_N \left(N^{p/d} \min_{|\Gamma| \leqslant N} \|X - \hat{X}^\Gamma\|_p^p \right) = J_{p,d} \left(\int |g|^{d/(d+p)}(u)\, \mathrm{d}u \right)^{1+p/d}, \tag{31}$$

where $\mathbb{P}_X(\mathrm{d}u) = g(u).\lambda_d(\mathrm{d}u) + \nu$, $\nu \perp \lambda_d$ (λ_d *Lebesgue measure on* $\mathbb{R}^d$). *The constant* $J_{p,d}$ *corresponds to the case of the uniform distribution on* $[0, 1]^d$.

Little is known about the true value of the constant $J_{p,d}$ except in one-dimension, where $J_{p,1} = \frac{1}{2^p(p+1)}$ and in two-dimensions, where, for example, $J_{2,2} = \frac{5}{18\sqrt{3}}$ (see Gersho and Gray

1982; Graf and Luschgy 2000). Nevertheless some bounds are available, based on the introduction of random quantization grids (see Zador 1982; Cohort 2003). Thus, as $d \to \infty$, $J_{p,d} \sim (d/2\pi e)^{p/2}$ (see Graf and Luschgy 2000).

Theorem 3 says that $\min_{|\Gamma|\leqslant N} \|X - \hat{X}^\Gamma\|_p = O(N^{-1/d})$: this means that optimal quantization of a distribution $\mathbb{P}_X$ produces (for every grid size N) some grids with the same rate as that obtained with uniform lattice grids (when $N = m^d$) for $U([0, 1]^d)$ distributions (in fact, even then, uniform lattice grids are never optimal if $d \geqslant 2$).

Example (*Numerical integration*). On the one hand, for every $p \geqslant 1$,

$$\|X - \hat{X}^\Gamma\|_p^p = \max\left\{\int_{\mathbb{R}^d} |\varphi - \varphi \circ q_\Gamma|^p \mathrm{d}\,\mathbb{P}_X,\ \varphi \text{ Lipschitz continuous },\ [\varphi]_{\mathrm{Lip}} \leqslant 1\right\}$$

(the equality stands for the function $\varphi : \xi \mapsto \min_{x_i \in \Gamma} |\xi - x_i|$). This induces a propagation of the L^p-quantization error by Lipschitz functions (already used in the proof of Theorem 2). On the other hand, from a numerical viewpoint,

$$|\mathbb{E}\,\varphi(\hat{X}^\Gamma) - \mathbb{E}\,\varphi(X)| \leqslant [\varphi]_{\mathrm{Lip}} \|X - \hat{X}^\Gamma\|_1 \leqslant [\varphi]_{\mathrm{Lip}} \|X - \hat{X}^\Gamma\|_p. \tag{32}$$

with

$$\mathbb{E}\varphi(\hat{X}^\Gamma) = \int_{\mathbb{R}^d} \varphi \,\mathrm{d}q_\Gamma(\mathbb{P}_X) = \sum_{i=1}^{N} \mathbb{P}(X \in C_i(\Gamma))\, \varphi(x_i). \tag{33}$$

(The parameter of interest is mainly $p = 2$, for algorithmic reasons.) The numerical computation of $\mathbb{E}\,\varphi(\hat{X}^\Gamma)$ for any (known) function φ relies on the grid $\Gamma = \{x_1, \ldots, x_N\}$ and its 'Voronoi $\mathbb{P}_X$-weights' $(\mathbb{P}_X(C_i(\Gamma)))_{1\leqslant i\leqslant N}$, whereas the error evaluation relies on $\|X - \hat{X}^\Gamma\|_p$. See Pagès (1997) and Fort and Pagès (2002) for further numerical integration formulae (Hölder, $\mathcal{C}^2$, locally Lipschitz continuous functions).

3.2. How to obtain optimal quantization

As far as numerical applications of optimal quantization of a random vector X are concerned, it has been emphasized above that we need an algorithm which produces an optimal (or at least a suboptimal) grid $\Gamma^* := \{x_1^*, \ldots, x_N^*\}$, the $\mathbb{P}_X$-mass of its Voronoi tessellation $(\mathbb{P}_X(C_i(\Gamma^*)))_{1\leqslant i\leqslant N}$, and the resulting L^p-quantization error $\|X - \hat{X}^{\Gamma^*}\|_p$.

3.2.1. One-dimensional quadratic setting ($d = 1$, $p = 2$)

In one-dimension, an algorithm, known as *Lloyd's Method I*, appears as a by-product of the uniqueness problem for optimal grids (in fact *stationary* grids, see (36) below): for every grid Γ of size N, one sets

$$T(\Gamma) = \left(\int_{C_i(\Gamma)} \xi\,\mathbb{P}_X(\mathrm{d}\xi) | \mathbb{P}_X(C_i(\Gamma))\right)_{1\leqslant i\leqslant N}.$$

The grid $T(\Gamma)$ has size N and satisfies $\|X - \hat{X}^{T(\Gamma)}\|_2 \leqslant \|X - \hat{X}^{\Gamma}\|_2$. If $\mathbb{P}_X$ has a *log-concave density function*, then T is contracting (see Kieffer 1982; Lloyd 1982; Trushkin 1982). Its unique fixed point Γ^* is clearly an optimal grid and the resulting (deterministic) iterative algorithm, $\Gamma^{t+1} = T(\Gamma^t)$, converges exponentially fast towards Γ^*.

3.2.2. Multidimensional setting ($d \geqslant 2$)

Lloyd's Method I formally extends to higher dimensions. Since there are always several suboptimal quantizers, T can no longer be contracting, although it still converges to some stationary quantizer (see (36) below), usually not optimal, even locally. Its major drawback is that it involves at each step several integrals $\int_{C_i(\Gamma)} \varphi \, \mathrm{d}\,\mathbb{P}_X$ (another is that it only works for $p = 2$). This suggests randomizing Lloyd's Method I by computing these integrals using Monte Carlo simulations of $\mathbb{P}_X$-distributed randomvectors.

There is another randomized procedure that can be called upon to find the critical points of a function *when its gradient admits an integral representation* with respect to a probability distribution: *stochastic gradient descent* (the stochastic counterpart of deterministic gradient descent). Let us recall briefly what this procedure is.

Let V be a differentiable *potential* function $V : \mathbb{R}^M \to \mathbb{R}$ such that $\lim_{|y| \to +\infty} V(y) = +\infty$, $|\nabla V|^2 = O(V)$, ∇V is Lipschitz continuous, $\{\nabla V = 0\}$ is locally finite, and ∇V has an integral representation $\nabla V(y) = \mathbb{E}\, \nabla_y\, v(y, X)$, where X is an $\mathbb{R}^L$-valued random vector. Let $(\gamma_t)_{t \geqslant 1}$ be a sequence of positive gain parameters satisfying $\sum_t \gamma_t = +\infty$ and $\sum_t \gamma_t^2 < +\infty$. Classical stochastic approximation theory says that the recursive algorithm

$$Y^{t+1} = Y^t - \gamma_{t+1} \nabla_y v(Y^t, \xi^{t+1}), \qquad \xi^t \text{ i.i.d., } \xi^1 \overset{\mathcal{L}}{\sim} X, \tag{34}$$

converges almost surely towards some critical point $y^* \in \{\nabla V = 0\}$ of V (for various results in this direction, see Duflo 1997; Kusher and Yin 1997). Under some additional assumptions, one shows that y^* is necessary a local minimum (see Pemantle 1990; Lazarev 1992; Brandière and Duflo 1996).

Let us return to our optimal quantization problem. We have already noticed that equation (28) defines a symmetric continuous function on $(\mathbb{R}^d)^N$, namely

$$D_N^{X,p}(x_1, \ldots, x_N) := \|X - \hat{X}^{\{x_1,\ldots,x_N\}}\|_p^p = \mathbb{E}(d_N^{X,p}(x, X)),$$

with

$$d_N^{X,p}(x_1, \ldots, x_N, \xi) := \min_{1 \leqslant i \leqslant N} |\xi - x_i|^p, \qquad (x_1, \ldots, x_N) \in (\mathbb{R}^d)^N,\ \xi \in \mathbb{R}^d.$$

$D_N^{X,p}$ and $d_N^{X,p}$ are called distortion and local distortion functions, respectively. One can show for $p = 2$, see Graf and Luschgy 2000; otherwise, see Pagès 1997 that for every $p > 1$, $D_N^{X,p}$ is continuously differentiable at every N-tuple $x := (x_1, \ldots, x_N) \in (\mathbb{R}^d)^N$ such that $x_i \neq x_j$, $i \neq j$ and $\mathbb{P}_X(\bigcup_{i=1}^N \partial C_i(x)) = 0$. The gradient $\nabla D_N^{X,p}(x)$ is given by a formal differentiation,

$$\nabla D_N^{X,p}(x) := \mathbb{E}(\nabla_x d_N^{X,p}(x, X)) = \left(\int_{\mathbb{R}^d} \frac{\partial d_N^{X,p}(x, \xi)}{\partial x_i} \mathbb{P}_X(\mathrm{d}\xi)\right)_{1 \leqslant i \leqslant N}, \tag{35}$$

with

$$\frac{\partial d_N^{X,p}}{\partial x_i}(x, \xi) := p \mathbf{1}_{C_i(x)} |x_i - \xi|^{p-1} \frac{x_i - \xi}{|x_i - \xi|}, \qquad 1 \leqslant i \leqslant N$$

(set $\vec{0}/\|\vec{0}\| := \vec{0}$). Equality (35) still holds when $p = 1$ if $\mathbb{P}_X$ is continuous. A grid $\{x_1, \ldots, x_N\}$ (with N pairwise distinct components and $\mathbb{P}_X$-negligible Voronoi tessellation boundaries) is

$$\text{a stationary grid if } \nabla D_N^{X,p}(x_1, \ldots, x_N) = 0. \tag{36}$$

Then, following Property 1, any optimal grid is stationary.

Setting $V := D_N^{X,p}$ and plugging the above formula for $\nabla_x d_N^{X,p}$ into the abstract stochastic gradient procedure (34) yields (reverting to the grid notation $\Gamma^t = \{X_1^t, \ldots, X_N^t\}$)

$$\Gamma^{t+1} = \Gamma^t - \frac{\gamma_{t+1}}{p} \nabla_x d_N^{X,p}(\Gamma^t, \xi^{t+1}), \qquad \Gamma^0 \subset \mathbb{R}^d, |\Gamma^0| = N, \xi^t \text{ i.i.d.}, \xi^1 \sim \mathbb{P}_X, \tag{37}$$

where *the step sequence* $(\gamma_t)_{t \geqslant 1}$ is $(0, 1)$-*valued* and satisfies, as usual, $\sum_t \gamma_t = +\infty$ and $\sum_t \gamma_t^2 < +\infty$. The fact that $\gamma_t \in (0, 1)$ for every $t \in \mathbb{N}$ ensures that $|\Gamma^t| = N$.

Unfortunately, the assumptions that make the stochastic gradient descent almost surely converge are never satisfied by the L^p-distortion function $D_N^{X,p}$ which is not a true potential function for at least two reasons. First, the gradient $\nabla D_N^{X,p}$ does not exist at N-tuples of $(\mathbb{R}^d)^N$ having stuck components (although it remains locally bounded). So, it cannot be Lipschitz or even Hölder continuous. Second, the L^p-distortion $D_N^{X,p}(x)$ does not go to infinity as $|x| := |x_1| + \ldots + |x_N| \to +\infty$ (only if $\min_{1 \leqslant i \leqslant N} |x_i| \to +\infty$).

However, $D_N^{X,p}$ turns out to be a fairly good potential function for practical implementation, especially in the quadratic case $p = 2$ (see Figures 1 and 3 and the CLVQ algorithm below). One does not observe in simulations some components of Γ^t becoming asymptotically stuck in (37) when $t \to +\infty$ in spite of the structure of the potential function. Although many stationary grids exist, it does converge toward a grid Γ^*, apparently close to optimality.

This quadratic case corresponds to a very commonly implemented procedure in automatic classification called Competitive Learning Vector Quantization (CLVQ), also known as the Kohonen algorithm with 0 neighbour. As we are motivated by numerical applications, we need to compute the target grid Γ^* with great accuracy. This usually means that significantly more iterations of the stochastic optimization procedure (37) are required than in other applications. When $p = 2$, the recursive procedure (37) can be described in a more geometric way. Set $\Gamma^t := \{X_1^t, \ldots, X_N^t\}$.

1. *Competitive phase*. Select $i(t+1) \in \text{argmin}_i |X_i^t - \xi^{t+1}|$ (closest neighbour).
2. *Learning phase*. Set

$$X_{i(t+1)}^{t+1} := X_{i(t+1)}^t - \gamma_{t+1}(X_{i(t+1)}^t - \xi^{t+1}),$$

$$X_i^{t+1} := X_i^t, \qquad i \neq i(t = 1).$$

Concerning numerical implementations of the algorithm, notice that, at each step the grid Γ^{t+1} lives in the convex hull of Γ^t and ξ^{t+1} since the learning phase is simply a homothety centred at ξ^{t+1} with ratio $1 - \gamma_{t+1} > 0$ (see Figure 2). This has a stabilizing effect on the procedure, which explains why one verifies in simulations that the CLVQ algorithm does behave better than its non-quadratic counterparts. Finally, one often 'refines' the CLVQ by processing a randomized Lloyd's I method (see Pagès and Printems 2003).

Figure 3 shows some planar quantizations obtained using the CLVQ algorithm for some simulations in the quadratic case, see Pagès and Printems 2003).

The main drawback of the CLVQ algorithm is that it is slow: roughly speaking, like most recursive stochastic algorithms, its rate of convergence is ruled by a central limit theorem (CLT) at rate $1/\sqrt{\gamma_t}$ (cf. Duflo 1997). Moreover, at each iteration, the computation of the winning index i_0 in the learning phase is time-consuming if N is large. Speeding up the algorithm requires both these aspects to be addressed. First, one may use some deterministic sequences with low discrepancy instead of pseudo-random numbers to implement the algorithm (see Lapeyre *et al*. 1990) or call upon some averaging methods which reduce the variance in the CLT theorem (see Pelletier 2000, and the references therein). To cut down

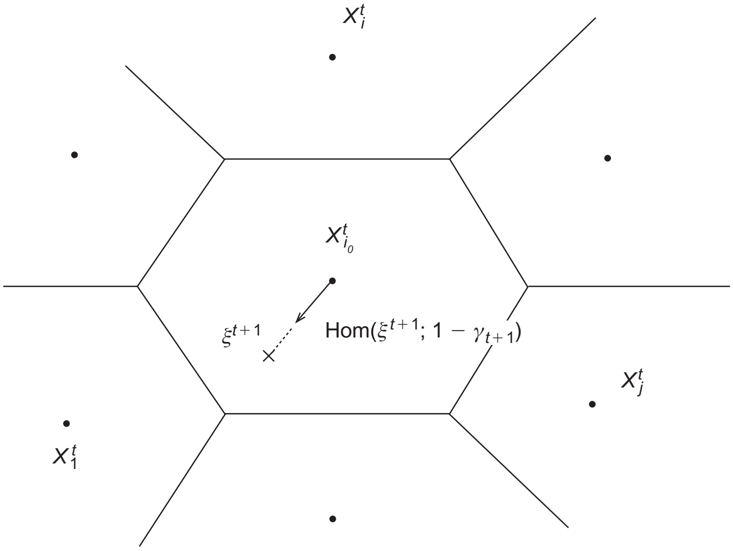

Figure 2. Competitive Learning Vector Quantization

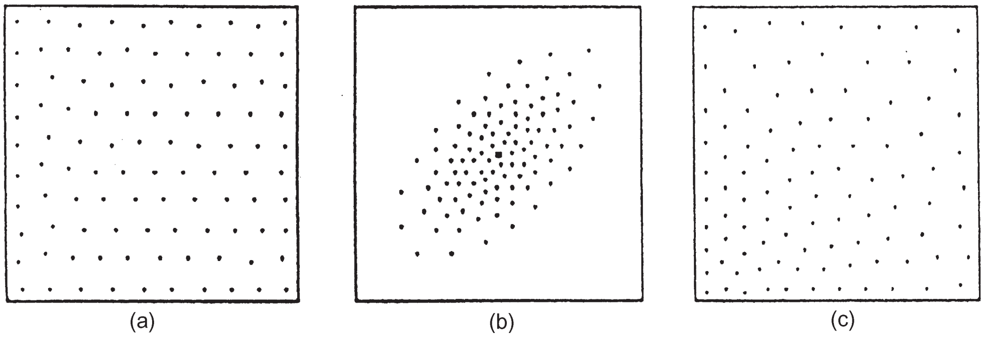

Figure 3. Planar quantizations obtained by CLVQ: (a) $\mathbb{P}_X = U([0, 1]^2)$; (b) $\mathbb{P}_X := \mathcal{E}(1/2)^{\otimes 2}$; (c) $\mathbb{P}_X := \mathcal{N}\left(0\begin{bmatrix} 1 & \sqrt{2}/2 \\ \sqrt{2}/2 & 1 \end{bmatrix}\right)$.

the winning index search time, one may implement some fast (approximate) search procedures. For recent developments, see Gersho and Gray (1992, pp. 332 and 479).

The last practical question of interest is the choice of the starting grid Γ^0. This question is clearly connected with the existence of several local minima for the distortion. One way is to start from a random N-grid obtained by simulation. An alternative is to process a so-called *splitting method*: one progressively adds some (optimal) quantizers with smaller size to the current state grid: the heuristic idea is that if an N-grid Γ_N^* (almost) achieves the the minimal distortion, then the $(N+\nu)$-grid $(\nu \ll N)$ $\Gamma_N^* \cup \Gamma_\nu^*$ is likely to be inside the attracting basin of the absolute minimum of the $(N+\nu)$-distortion. These computational aspects are developed in Pagès and Printems (2003) in the important framework of Gaussian distributions.

Finally, one can polish up the converging phase of the grid produced by the CLVQ stochastic gradient by processing a randomized Lloyd's Method I procedure.

Turning to theoretical aspects of the CLVQ algorithm, some rigorous almost sure convergence results have been established in Pagès (1997), only for compactly supported distributions $\mathbb{P}_X$ on $\mathbb{R}^d$. When $d \geqslant 2$ this convergence only holds in the Kushner–Clark (or conditional) sense, whereas standard almost sure convergence towards a true N-grid holds in one dimension (see the Appendix for details).

As a conclusion, the CLVQ and its L^p-counterparts compress the information on $\mathbb{P}_X$ provided by the sequence $(\xi^t)_{t\in\mathbb{N}^*}$: it appears to be a *compressed Monte Carlo method*.

We now turn to the estimation of the companion parameters. The proposition below shows that the weight vector $(\mathbb{P}_X(C_i(\Gamma^*)))_{1\leqslant i\leqslant N}$ and the induced L^p-quantization error $\|X - \hat{X}^{\Gamma^*}\|_p$ can be obtained on-line for free as a by-product of the CLVQ stochastic gradient descent as soon as Γ^t converges to Γ^* (this holds true for any $p \geqslant 1$).

Proposition 7. *Let* $p \geqslant 1$. *Assume that* $X \in L^{p+\eta}$, $\eta > 0$, *and* $\mathbb{P}_X$ *weights no hyperplane. Set, for every* $t \geqslant 1$,

$$p_i^t := \frac{|\{1 \leqslant s \leqslant t | \xi^s \in C_i(\Gamma^{s-1})\}|}{t},$$

the empirical frequency of the ξ^s *falling in the ith tessellation* $C_i(\Gamma^{s-1})$ *of the (moving) Voronoi tessellation of* Γ^{s-1} *up to time* t $(i = 1, \ldots, N)$*. Also set*

$$D^{\beta,t} := \frac{1}{t}\sum_{s=1}^{t} \min_{1 \leqslant i \leqslant N} |\xi^s - X_i^{s-1}|^\beta, \qquad \beta \in (0,\ p+\eta),$$

the average of the lowest distance of ξ^s *(to the power* β*) to the moving quantization grid* $\Gamma^{s-1} = \{X_1^{s-1}, \ldots, X_N^{s-1}\}$, $1 \leqslant s \leqslant t$.

Let Γ^* *be a (stationary) grid (see* (36)*) and let* $A_{\Gamma^*} := \{\Gamma^t \to \Gamma^*\}$ *be the set of convergence of* Γ^t *towards* Γ^**. Then, on the event* A_{Γ^*},

$$\forall i \in \{1, \ldots, n\}, \qquad p_i^t \xrightarrow{a.s.} p_i := \mathbb{P}_X(C_i(\Gamma^*)) \quad \textit{as } t \to +\infty,$$

$$\forall \beta \in (0,\ p), \qquad D^{\beta,t} \xrightarrow{a.s.} D_N^{X,\beta}(\Gamma^*) \quad \textit{as } t \to +\infty.$$

Remark 7. The above quantities p^t and $D^{\beta,t}$ obviously satisfy the recursive procedures

$$p_i^t = p_i^{t-1} - \frac{1}{t}(p_i^{t-1} - \mathbf{1}_{\{\xi^t \in C_i(X^{t-1})\}}), \qquad D^{\beta,t} = D^{\beta,t-1} - \frac{1}{t}\left(D^{\beta,t-1} - \min_{1 \leqslant i \leqslant N}|X_i^{t-1} - \xi^t|^\beta\right). \tag{38}$$

In fact, the conclusion of Proposition 7 still holds if $(1/t)_{t \geqslant 1}$ is replaced in (38) by any positive sequence $(\tilde{\gamma}_t)_{t \geqslant 1}$ satisfying $\sum_t \tilde{\gamma}_t = +\infty$ and $\sum_t \tilde{\gamma}_t^{(p+\varepsilon)/\beta} < +\infty$. Natural choices for $\tilde{\gamma}_t$ are $1/t$ or the original step γ_t used in the learning phase of the CLVQ algorithm, depending on the range of the simulation (see Pagès and Printems 2003).

Proof. For notational convenience, a generic N-tuple $x = (x_1, \ldots, x_N) \in (\mathbb{R}^d)^N$ will be denoted by its grid notation $\Gamma = \{x_1, \ldots, x_N\}$. Let $\Phi : (\mathbb{R}^d)^N \times \mathbb{R}^d \to \mathbb{R}$ be a Borel function satisfying $|\Phi(\Gamma, \xi)| \leqslant M|\xi|^\beta$ for some real constants $M > 0$ and $0 \leqslant \beta < p + \eta$, and such that the function defined by $\varphi(\Gamma) := \int_{\mathbb{R}^d} \Phi(\Gamma, \xi)\mathbb{P}_X(d\xi)$ is bounded and continuous at Γ^*. One verifies that $(\Phi(\Gamma^t, \xi^{t+1}) - \varphi(\Gamma^t))_{t \geqslant 0}$ is an $L^{(p+\eta)/\beta}$-bounded sequence of martingale increments. Now the Chow theorem (see Duflo 1997, p. 22) and $\sum_{t \geqslant 1} t^{-(p+\eta)/\beta} < +\infty$ imply that the martingale $\sum_{1 \leqslant s \leqslant t}(\Phi(\Gamma^{s-1}, \xi^s) - \varphi(\Gamma^{s-1}))/s$ converges almost surely towards a finite random variable Z_∞. In turn, the Kronecker lemma finally implies that

$$\frac{1}{t}\sum_{s=1}^{t} \Phi(\Gamma^{s-1}, \xi^s) - \varphi(\Gamma^{s-1}) \xrightarrow{a.s.} 0.$$

Finally, the continuity of φ at Γ^* yields

$$\frac{1}{t}\sum_{s=1}^{t} \Phi(\Gamma^{s-1}, \xi^s) \xrightarrow{a.s.} \varphi(\Gamma^*) \text{ on } A_{\Gamma^*}.$$

One first applies this result to the indicator functions $\Phi_i(\Gamma, \xi) := \mathbf{1}_{C_i(\Gamma)}(\xi)$, $1 \leqslant i \leqslant n$; the associated φ functions are continuous at Γ^* because $\mathbb{P}_X$ weights no hyperplane.

Then set

$$\Phi(\Gamma, \xi) := \rho(\Gamma) \min_{1 \leqslant i \leqslant N} \|\xi - x_i\|^\beta$$

where ρ is a continuous $[0, 1]$-valued function with compact support on $(\mathbb{R}^d)^N$ satisfying $\rho(\Gamma^*) = 1$. $\varphi(\Gamma) := \rho(\Gamma) D_N^{X,\beta}(\Gamma)$ is continuous and, on A_{Γ^*}, $\Phi(\Gamma^s, \xi^{s+1}) = \min_{1 \leqslant i \leqslant N} \|\xi^{s+1} - X_i^s\|^\beta$ for large enough s. □

3.3. Optimal quantization of a simulatable Markov chain

Let us now move on to the optimal quantization of an $\mathbb{R}^d$-valued Markov chain $(X_k)_{0 \leqslant k \leqslant n}$ with transition $P(x, \mathrm{d}y)$ and initial distribution $\mu_0 = \mathcal{L}(X_0)$. Assume that, for every $x \in \mathbb{R}^d$, the distributions $P(x, \mathrm{d}y)$ can be easily simulated on a computer, as well as μ_0. A typical example is the Gaussian Euler scheme of a diffusion (see Section 2.1). Assume that, for every $k = 0, \ldots, n$, $X_k \in L^{p+\eta}$ $(\eta > 0)$, and that

$$\text{the distribution of } X_k \text{ weights no hyperplane in } \mathbb{R}^d. \tag{39}$$

3.3.1. *The extended* CLVQ *algorithms for Markov chain optimal quantizations*

The principle is to modify the Monte Carlo simulation briefly outlined in Section 2.2.1 by processing a CLVQ algorithm at each step k. One starts from a large-scale Monte Carlo simulation of the Markov chain $(X_k)_{0 \leqslant k \leqslant n}$ that we will denote by $\xi^0 := (\xi_0^0, \ldots, \xi_n^0)$, $\xi^1 := (\xi_0^1, \ldots, \xi_n^1), \ldots, \xi^t := (\xi_0^t, \ldots, \xi_n^t), \ldots$. Our aim is to produce for every $k \in \{0, \ldots, n\}$ some optimal grids $\Gamma_k := \{x_1^k, \ldots, x_{N_k}^k\}$ with size N_k, their transition kernels $[\pi_{ij}^k]$ and their L^p-quantization errors. Note that, if one sets

$$p_i^k := \mathbb{P}(X_k \in C_i(\Gamma_k)), \qquad p_{ij}^k := \mathbb{P}(\{X_{k+1} \in C_j(\Gamma_{k+1})\} \cap \{X_k \in C_i(\Gamma_k)\}),$$

then

$$\pi_{ij}^k := \mathbb{P}(\hat{X}_{k+1} = x_j^{k+1} | \hat{X}_k = x_i^k) = \frac{p_{ij}^k}{p_i^k},$$

$$i = 1, \ldots, N_k,\ j = 1, \ldots, N_{k+1},\ k = 0, \ldots, n-1.$$

In the quadratic case $(p = 2)$, the *extended* CLVQ algorithm proceeds as follows:

1. *Initialization phase.*
   - Initialize the $n+1$ starting grids $\Gamma_k^0 := \{x_1^{0,k}, \ldots, x_{N_k}^{0,k}\}$, $k = 0, \ldots, n$, of the $n+1$ CLVQ algorithms that will quantize the distributions μ_k.
   - Initialize the 'marginal counter' vectors $\alpha_i^{k,0} := 0$, $i = 1, \ldots, N_k$, for every $k = 0, \ldots, n$.

- Initialize the 'transition counters' $\beta_{ij}^{k,0} := 0$, $i = 1, \ldots, N_k$, $j = 1, \ldots, N_{k+1}$, $k = 0, \ldots, n-1$.

2. *Updating* $t \rightsquigarrow t+1$. At step t, the $n+1$ grids Γ_k^t, $k = 0, \ldots, n$, have been obtained. We now use the sample ξ^{t+1} to carry on the optimization process, building up the grids Γ_k^{t+1} as follows. For every $k = 0, \ldots, n$:
   - Simulate ξ_k^{t+1} (using ξ_{k-1}^{t+1} if $k \geqslant 1$).
   - Select the 'winner' in the kth CLVQ algorithm, that is, the only index $i_k^{t+1} \in \{1, \ldots, N_k\}$ satisfying
   $$\xi_k^{t+1} \in C_{i_k^{t+1}}(\Gamma_k^t).$$
   - Update the kth CLVQ algorithm:
   $$\forall i \in \{1, \ldots, N_k\}, \qquad \Gamma_{k,i}^{t+1} = \Gamma_{k,i}^t - \gamma_{t+1} \mathbf{1}_{\{i = i_k^{t+1}\}} (\Gamma_{k,i}^t - \xi_k^{t+1}).$$
   - Update the kth marginal counter vector $\alpha^{k,t} := (\alpha_i^{k,t})_{1 \leqslant i \leqslant N_k}$:
   $$\forall i \in \{1, \ldots, N_k\}, \qquad \alpha_i^{k,t+1} := \alpha_i^{k,t} + \mathbf{1}_{\{i = i_k^{t+1}\}}.$$
   - Update the (quadratic) distortion estimator $D^{k,t}$:
   $$D^{k,t+1} := D^{k,t} + \frac{1}{t+1}(|\Gamma_{k,i_k^{t+1}}^t - \xi^{t+1}|^2 - D^{k,t}).$$
   - Update the transition counters $\beta^{k,t} := (\beta_{ij}^{k,t})_{1 \leqslant i \leqslant N_{k-1}, 1 \leqslant j \leqslant N_k}$ ($k \geqslant 1$):
   $$\forall i \in \{1, \ldots, N_{k-1}\}, \forall j \in \{1, \ldots, N_k\}, \qquad \beta_{ij}^{k-1,t+1} := \beta_{ij}^{k-1,t} + \mathbf{1}_{\{i = i_{k-1}^{t+1}, j = i_k^{t+1}\}}.$$
   - Update the transition kernels $(\pi_{ij}^{k,t})_{1 \leqslant i \leqslant N_{k-1}, 1 \leqslant j \leqslant N_k}$ ($k \geqslant 1$):
   $$\pi_{ij}^{k,t+1} := \frac{\beta_{ij}^{k,t+1}}{\alpha_i^{k,t+1}} \qquad \text{(possibly only once at the end of the simulation process!).}$$

Following Proposition 7 one has, on the event $\{\Gamma_k^t \to \Gamma_k^*\}$, $D^{k,t} \overset{t \to +\infty}{\longrightarrow} D_{N_k}^{X_k,2}(\Gamma_k^*)$ and

$$\frac{\alpha^{k,t}}{t} \overset{t \to +\infty}{\longrightarrow} (p_i^{*,k})_{1 \leqslant i \leqslant N_k} = (\mathbb{P}(X_k \in C_i(\Gamma_k^*)))_{1 \leqslant i \leqslant N_k} \qquad \text{(thanks to (39)).}$$

The same martingale approach shows that, on the event $\{\Gamma_k^t \to \Gamma_k^*\} \cap \{\Gamma_{k+1}^t \to \Gamma_{k+1}^*\}$,

$$\frac{\beta^{k,t}}{t} \overset{t \to +\infty}{\longrightarrow} p_{ij}^{*,k} = (\mathbb{P}(X_k \in C_i(\Gamma_k^*), X_{k+1} \in C_j(\Gamma_{k+1}^*)))_{1 \leqslant i \leqslant N_k, 1 \leqslant j \leqslant N_{k+1}}$$

so that, on the event $\{\Gamma_k^t \overset{t \to +\infty}{\longrightarrow} \Gamma_k^*, k = 0, \ldots, n\}$, for every $k \in \{1, \ldots, n\}$,

$$\pi_{ij}^{k,t} \overset{t \to +\infty}{\longrightarrow} \pi_{ij}^{*,k}, \qquad 1 \leqslant i \leqslant N_k, 1 \leqslant j \leqslant N_{k+1}.$$

The main features of this algorithm are essentially those of the original CLVQ algorithm. Moreover, note that the forward optimization of the grids and the weight computation are not recursive in k, so there is no deterioration of the optimization process as k increases.

When $p \neq 2$ we employ an *L^p-optimization procedure (extended L^p-CLVQ)* by using the general L^p-formula (37) in the grid updating phase.

3.3.2. A priori *optimal dispatching of optimal quantizer sizes*

The above grid optimization procedures for Markov chains (extended CLVQ and its L^p-variants) produce optimal grids Γ_k for a given dispatching of their sizes N_k. This raises the optimization problem of how to dispatch a priori the sizes $N_0, \ldots, N_n$ of the quantization grids, assumed to be L^p-optimal, if one wishes to use at most $N \geqslant N_0 + \ldots + N_n$ elementary quantizers.

Let $p \geqslant 1$. Formula (27) in Theorem 2 provides some positive real coefficients $d_0, d_1, \ldots, d_n$ such that, for *any sequence of quantizations* $(\hat{X}_k)_{0 \leqslant k \leqslant n}$,

$$\|u_0(X_0) - \hat{u}_0(\hat{X}_0)\|_p \leqslant \sum_{i=0}^{n} d_i \|X_i - \hat{X}_i\|_p. \tag{40}$$

Then, the best we can do is to specify *the sizes N_k that minimize the right-hand side of* (40) *when all the quantization vectors $\hat{X}_k$'s are L^p-optimal, that is,*

$$\min_{(N_0+\ldots+N_n \leqslant N)} \sum_{i=0}^{n} d_i \times \|X_i - \hat{X}_i\|_p \quad \text{with } \|X_i - \hat{X}_i\|_p = \min_{|\Gamma| \leqslant N_i} \|X_i - \hat{X}_i^{\Gamma}\|_p, \; 0 \leqslant i \leqslant n.$$

This will also produce an asymptotic bound for the resulting L^p-error $\|\hat{u}_0(\hat{X}_0) - u_0(X_0)\|_p$. The key is Theorem 3, which says that $N_k^{1/d} \min_{|\Gamma| \leqslant N_k} \|X_k - \hat{X}_k^{\Gamma}\|_p$ converges to some positive constant as $N_k \to +\infty$.

Proposition 8. *Assume that all the distributions $\mathcal{L}(X_k)$ have an absolutely continuous part φ_k, $0 \leqslant k \leqslant n$. Let $N \in \mathbb{N}^*$. Set, for every $i \in \{0, \ldots, n\}$,*

$$N_i := [\underline{\rho}_i N] \quad \text{and} \quad \underline{\rho}_i := \frac{a_i}{\sum_{0 \leqslant j \leqslant n} a_j} \qquad \text{with } a_i := \left(\|\varphi_i\|_{d/(d+p)}^{1/p} d_i\right)^{d/(d+1)}, \quad 0 \leqslant i \leqslant n. \tag{41}$$

Assume that all the quantizations $\hat{X}_k$ of the X_k are L^p-optimal with size N_k. Then,

$$\overline{\lim_N}\, N^{1/d} \max_{0 \leqslant i \leqslant n} \|u_i(X_i) - \hat{u}_i^{(N_i)}(\hat{X}_i)\|_p \leqslant J_{p,d}^{1/p} \left(\sum_{i=0}^{n} a_i\right)^{1+1/d} \tag{42}$$

where $J_{p,d}$ is defined in Theorem 3.

The following simple lemma solves the 'continuous bit allocation' problem.

Lemma 1. *Let $a_0, \ldots, a_n$ be some positive real numbers. Then the function $\rho := (\rho_0, \ldots, \rho_n) \mapsto \sum_{i=0}^{n} a_i \rho_i^{-1/d}$ defined on the set $\{\rho_0 + \ldots + \rho_n = 1, \rho_i \geqslant 0, 0 \leqslant i \leqslant n\}$ reaches its minimum at*

$$\underline{\rho} := \left(\frac{a_i^{d/(d+1)}}{a_0^{d/(d+1)} + \ldots + a_n^{d/(d+1)}}\right)_{0\leqslant i\leqslant n}$$

so that

$$\sum_{i=0}^{n} a_i \underline{\rho}_i^{-1/d} = \left(\sum_{i=0}^{n} a_i^{d/(d+1)}\right)^{1+1/d}.$$

Note that this minimum value is non-decreasing as a function of the a_i *and that*

$$\left(\sum_{j=0}^{n} a_j^{d/(d+1)}\right)^{1+1/d} \leqslant (n+1)^{1/d} \sum_{i=0}^{n} a_i.$$

Proof of Proposition 8. First, rewrite (40) as follows:

$$N^{1/d}\|u_0(X_0) - \hat{u}_0^{(N_0)}(\hat{X}_0)\|_p \leqslant \sum_{i=0}^{n} d_i \left(\frac{N_i}{N}\right)^{-1/d} (N_i^{1/d}\|X_i - \hat{X}_i\|_p),$$

keeping in mind that every $\hat{X}_k$ is L^p-optimal with size N_k. Then,

$$\begin{aligned}\overline{\lim_N}\, N^{1/d} \max_{0\leqslant i\leqslant n} \|u_i(X_i) - \hat{u}_i^{(N_i)}(\hat{X}_i)\|_p &\leqslant \sum_{i=0}^{n} d_i\, \overline{\lim_N} \left\{\left(\frac{N}{N_i}\right)^{1/d}\Big(N_i^{1/d}\|X_i - \hat{X}_i\|_p\Big)\right\}\\ &= \sum_{i=0}^{n} d_i \underline{\rho}_i^{-1/d}\, \overline{\lim_N}\Big(N_i^{1/d}\|X_i - \hat{X}_i\|_p\Big).\end{aligned}$$

Now, as $N_i \to +\infty$ for every $i \in \{0, \ldots, n\}$, Theorem 3 implies that $\overline{\lim}_N N_i^{1/d}\|X_i - \hat{X}_i\|_p = (J_{p,d}\|\varphi_i\|_{d/(d+p)})^{1/p}$. Then

$$\begin{aligned}\overline{\lim_N}\, N^{1/d} \max_{0\leqslant i\leqslant n} \|u_i(X_i) - \hat{u}_i^{(N_i)}(\hat{X}_i)\|_p &\leqslant J_{p,d}^{1/p} \sum_{i=0}^{n} d_i \|\varphi_i\|_{d/d+p}^{1/p} \underline{\rho}_i^{-1/d}\\ &\leqslant J_{p,d}^{1/p}\left(\sum_{i=0}^{n} \|\varphi_i\|_{d/(d+p)}^{d/(\,p(d+1))} d_i^{d/(d+1)}\right)^{1+1/d}.\end{aligned}$$

and the theorem is proved. □

Remark 8. The above asymptotic bound (42) is not fully satisfactory: if one wishes to optimize the number n of time steps as a function of the number N of elementary quantizers, one needs some *a priori* error bounds for a given couple (n, N). To this end we will need to control the distributions of the X_k, namely to 'dominate' them by a fixed distribution up to some affine time scaling. This can be done, for example, with some uniformly elliptic diffusions $(X_{t_k})_{0\leqslant k\leqslant n}$ (see Section 6.1) or with their Doléans exponentials (see Bally *et al*. 2002b).

4. Weight estimation in the quantization tree: statistical error

In this short section we summarize some results developed in Bally and Pagès (2003). First, although the question of the rate of convergence of the quantization grid optimization and the companion parameter estimation is natural, one must keep in mind that this is *a one-shot phase of the process*.

We will not discuss rigorously the error induced on the coefficients π^k_{ij} by the on-line estimation procedure (38): it would lead us too far, given the partial results available on the rate of convergence of the CLVQ algorithm. However, let us mention that, under some appropriate assumptions (see Duflo 1997, p. 52), one can show that a recursive stochastic algorithm like (34) with step γ_t satisfies a CLT with rate $1/\sqrt{\gamma_t}$ as $t \to +\infty$. Thus, if $\gamma_t \sim \gamma/t$ (with γ large enough), a 'standard' CLT holds as for the regular Monte Carlo method (for more details concerning the rate of convergence of the CLVQ algorithm in one dimension, see Pagès and Printems 2003).

A less ambitious but still challenging problem is the following. Consider some fixed grids $\Gamma_0, \ldots, \Gamma_n$, their companion parameters and the related quantization tree algorithm (23). What is the error induced by the use of Monte Carlo estimated weights $\tilde{\pi}^k_{ij}$ instead of the true weights π^k_{ij}? This third type of error – the *statistical error* – is extensively investigated in Bally and Pagès (2003), but let us summarize the situation (a CLT is given in Bally 2002).

This error strongly depends on the structure of the nonlinearity. In the linear case (no reflection: $h(t, \cdot) := -\infty$ if $t \neq T$, and $f\colon \equiv 0$), the composition of the empirical frequency matrices $(\tilde{\pi}^k_{ij})_{0 \leqslant k \leqslant n}$ yields $\tilde{\alpha}^n_i = (1/M)\sum_{\ell=1}^M \mathbf{1}_{\{\hat{X}^\ell_n = x^n_i\}}$ so that $\tilde{U}_n = (1/M)\sum_{\ell=1}^M h(T, \hat{X}^\ell_n)$. This is but a standard Monte Carlo method for computing $\mathbb{E}h(T, X_T)$, ruled by the regular CLT: the statistical error is $O(1/\sqrt{M})$. In the nonlinear case, the empirical frequencies cannot be composed. In Bally and Pagès (2003) we focus on the regular Snell envelope of $(h(X_k))_{0 \leqslant k \leqslant n}$ where h is a *bounded* Lipschitz continuous function ($f \equiv 0$) and X_k stands either for the diffusion $X_{kT/n}$ (with Lipschitz continuous coefficients) or for its Euler scheme $\bar{X}_k$, $k = 0, \ldots, n$. Then the statistical error depends on the regularity of the obstacle h as follows:

$$\mathbb{E}|\tilde{u}_0(x_0) - \hat{u}_0(x_0)| \leqslant C_{b,\sigma,h,T}\frac{\sqrt{nN}\sum_{k=1}^n \|X_k - \hat{X}_k\|_2 + \rho_{n,N,M}}{\sqrt{M}}, \qquad C_{b,\sigma,h,T} > 0, \tag{43}$$

where $\rho_{n,N,M} = \sqrt{n} + N^2/\sqrt{M}$ if h is semi-convex and X_k is the diffusion $X_{kT/n}$, and $\rho_{n,N,M} = n^{3/4} + N^2/\sqrt{nM}$ otherwise. Optimality of the quantizations $\hat{X}_k$ is not required.

5. Finite-element method versus quantization: a first comparison

A quick comparison of the finite-element method, on the one hand, and the quantization method, on the other, shows that there is a strong analogy between them: in both cases one computes the approximation of the solution at a finite number of points using a weighted sum, starting from a final condition also evaluated at a finite number of points. The aim of this short section is to obtain a slightly deeper insight into this analogy. In order to compare the two methods we will use the simplest possible example, the heat equation

$$(\partial_t + \tfrac{1}{2}\Delta)u = 0, \qquad u_T = f. \tag{44}$$

Let $\langle \varphi | \psi \rangle := \int_{\mathbb{R}^d} \varphi(x)\psi(x)\mathrm{d}x$ denote the inner product of $L^2(\mathbb{R}^d, \lambda_d)$. The weak form of (44) is given by

$$\langle f|\varphi\rangle - \langle u_t|\varphi\rangle - \frac{1}{2}\int_t^T \sum_{i=1}^d \left\langle \frac{\partial u_s}{\partial x^i} \middle| \frac{\partial \varphi}{\partial x^i} \right\rangle \mathrm{d}s = 0,$$

for every $t \in [0, T]$ and every test function $\phi \in \mathcal{C}_c^2(\mathbb{R}^d, \lambda_d)$. The first step of both approaches is the time discretization. Let $n \in \mathbb{N}^*$ and $t_k = kh$, $h := T/n$ (we temporarily abandon the notation Δ for the step because of the Laplacian). Then, we consider the discrete problem

$$\langle f|\phi\rangle - \langle u_{t_{k_0}}|\phi\rangle - \frac{h}{2}\sum_{k=k_0}^n \sum_{i=1}^d \left\langle \frac{\partial u_{t_k}}{\partial x^i} \middle| \frac{\partial \phi}{x^i} \right\rangle = 0.$$

We use the dynamic programming principle in order to solve this problem by induction: we put $\bar{u}_n := f$ and define $\bar{u}_k$ to be the solution of the elliptic problem

$$\langle \bar{u}_{k+1}|\phi\rangle - \langle \bar{u}_k, \phi\rangle - \frac{h}{2}\sum_{i=1}^d \left\langle \frac{\partial \bar{u}_k}{\partial x^i} \middle| \frac{\partial \phi}{\partial x^i} \right\rangle = 0. \tag{45}$$

Equation (45) can be seen either as a Dirichlet problem on the whole of $\mathbb{R}^d$ with condition zero at infinity or as a problem restricted to a large enough ball. This is a technical point which requires some special treatment (an additional error appears if we restrict to a ball), but this is of no interest here.

At this stage, one calls upon the finite-element method (or the quantization) in order to solve (45). In both cases, one builds up a grid $\Gamma := \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ and looks for some weights π_{ij}^k in order to approximate $\bar{u}_k$ by

$$\hat{u}_k(x_i) = \sum_{1 \leqslant j \leqslant N} \pi_{ij}^k \, \hat{u}_{k+1}(x_j). \tag{46}$$

5.1. The finite-element method

In the finite-element method, one tries to find the grid which best fits in the geometry of the problem. The same aim appears in the quantization method when we look for an optimal quantization. The simplest grid used in the finite-element method is based on triangles. Each x_i is the vertex of several triangles and we may consider it as the centroid of the polygon completed by these triangles. We denote by $\mathcal{N}_i := \{x_{i_1}, \ldots, x_{i_k}\}$ the vertices of this polygon, that is, the points which are vertices of a triangle having x_i as a vertex. These are the neighbours of x_i.

Now, having defined a grid Γ, we focus on the construction of the weights π_{ij}^k and on their significance. We begin with the finite-element method. One constructs the trials T_i, $1 \leqslant i \leqslant N$, in the following way. One sets $T_i(x_i) := 1$ and $T_i(x_j) := 0$ for every $x_j \in \mathcal{N}_i$. Then one sets T_i to be linear on each triangle – so we obtain a pyramid centred at x_i whose basis is the polygon. The trial is null outside the polygon. The idea is to replace (45) by a finite-dimensional problem to be solved in the space $\mathcal{S}_N$ spanned by T_i, $1 \leqslant i \leqslant N$. Note that the trials are not orthogonal. In any case, each function $U \in \mathcal{S}_N$ may be written as $U(x) := \sum_{i=1}^N U_i\, T_i(x)$ (we use lower-case letters for general functions in L^2 and capital letters for functions in $\mathcal{S}_N$). Note also that $U_i = U(x_i)$ and that

$$\frac{\partial U}{\partial x_i}(x) = \sum_{j=1}^N U_j \frac{\partial T_j}{\partial x_i}(x).$$

There are two ways of solving the finite-dimensional problem, leading to the same result. The first one is the Galerkin method based on the weak form of the PDE and the second one is the Riesz–Reilich method based on the Dirichlet principle. We use the second here. The solution of (45) is given by the function $u \in H_0^1(\mathbb{R}^d)$ which minimizes the energy

$$e(u) =: \int_{\mathbb{R}^d} \left(\frac{h}{2} \sum_{i=1}^d \left| \frac{\partial u}{\partial x_i}(x) \right|^2 - u(x)\bar{u}_{k+1}(x) \right) \mathrm{d}x.$$

The discretization consists in solving the above minimum problem in $\mathcal{S}_N$. Since each function in this space may be identified with the vector $U := (U_1, \ldots, U_N)$, one may write the discrete problem as follows. Let

$$E(U) := \frac{h}{2} \sum_{i,j=1}^N U_i\, K_{ij}\, U_j - \sum_{i=1}^N U_i\, U_i^{k+1}$$

where $K_{ij} := \sum_{1 \leqslant r \leqslant N} \langle \partial T_i / \partial x_p | \partial T_j / \partial x_r \rangle$ and $U_i^{k+1} := \hat{u}_{k+1}(x_i)$, that is, the coefficients of $\hat{u}_{k+1}$. Note that $\bar{u}_{k+1}$ has been changed into $\hat{u}_{k+1}$: the hat stresses that we are working with the approximation computed at the step $k+1$ of the dynamic principle algorithm. Now we have a finite-dimensional problem

$$\hat{u}_k(x_i) = \sum_{1 \leqslant j \leqslant N} \pi_{ij}^k \hat{u}_{k+1}(x_j), \qquad \text{with } \pi_{ij}^k := \frac{2}{h}(K)_{ij}^{-1}$$

(see (46)), whose solution is given by $U^k = 2h^{-1}K^{-1}U^{k+1}$.

5.2. Quantization method

We turn now to the quantization method where $\pi_{ij}^k = \mathbb{P}(X_{k+1} \in C_j(\Gamma_{k+1}) | X_k \in C_i(\Gamma_k))$. Writing $P_h(x, \mathrm{d}y) := \mathbb{P}(X_{k+1} \in \mathrm{d}y | X_k = x)$, then

$$\pi_{ij}^k = \int_{C_i(\Gamma_k)} P_h(\mathbf{1}_{C_j(\Gamma_{k+1})})(\xi) \mathrm{d}\mathbb{P} \circ X_k^{-1}(\mathrm{d}\xi) \approx \frac{1}{h} P_h(\mathbf{1}_{C_j(\Gamma_{k+1})})(x_i).$$

So the part of the trial T_i in the finite-element method is played here by the indicator function of the Voronoi tessellation of x_i. Both $h^{-1}T_i$ and $h^{-1}\mathbf{1}_{C_i}$ are approximations of the Dirac mass at x_i. Finally, we note that

$$\sum_{1\leqslant j\leqslant N} \hat{u}_{k+1}(x_j)\pi_{ij}^k \approx P_h\hat{u}_{k+1}(x_i).$$

The Feynman–Kac formula shows that $\hat{u}_k$ is the solution of (45) with final condition $\hat{u}_{k+1}$.

5.3. Conclusion

In both methods the first step is a time discretization. Then, as a second step, both of them solve the same PDE problem (45) using a space approximation procedure. The finite-element method relies on the variational principle, whereas the quantization method is based on the probabilistic interpretation of the PDE. In the finite-element method this leads to a linear system. Solving it amounts to inverting the sparse matrix K (only neighbours yield non-null entries). In the quantization method, we directly obtain the solution of the system: the weights are computed using the Monte Carlo method. Once again, the matrix K is sparse, numerically speaking, since the Brownian motion does not go too far in one time step: many weights π_{ij}^k will be close to 0 except for the neighbours.

Finally, note that formula (46), however it arises, reads as a finite-difference scheme built on a grid which is no longer uniform with weights which are no longer h^{-1}. It looks like a finite-difference scheme *adapted to the geometry of the problem*.

Beyond these similarities, it appears that in the finite-element method *one projects the function* to be computed as an expectation, whereas in the quantization method *one projects the underlying process X* involved in the Feynman–Kac formula.

From a numerical point of view, there is a connection between the conditioning of the matrix K to be inverted in the finite-element method and the asymptotic variance term in the CLT that – heuristically – gives the rate of convergence of the CLVQ algorithm: when K has a small eigenvalue, the variance of the CLVQ is large *i.e.* it converges more slowly.

6. Applications to RBSDEs and American option pricing

6.1. RBSDEs and optimal stopping of Brownian diffusions

In Section 2.1 we pointed out that (h, f)-Snell envelopes $(U_k)_{0\leqslant k\leqslant n}$ and $(\overline{U}_k)_{0\leqslant k\leqslant n}$ of the sampled diffusion $(X_{kT/n})_{0\leqslant k\leqslant n}$ or its Euler scheme $(\overline{X}_k)_{0\leqslant k\leqslant n}$ each provide natural discretization schemes for the RBSDE (according to the ability to simulate the diffusion). For a time discretization step T/n, these Snell envelopes are both related to the functions $f_k(x, u) := (T/n)f(t_k, x, u)$ and $h_k(x) := h(t_k, x)$, $k = 0, \ldots, n$. They satisfy

$$U_n := h(T, X_T),\ U_k := \max(h_k(X_{kT/n}), \mathbb{E}(U_{k+1} + f_k(X_{(k+1)T/n}, U_{k+1})|\mathcal{F}_{kT/n})),\quad 0 \leqslant k \leqslant n-1, \quad (47)$$

$$\overline{U}_n := h(T, \overline{X}_n),\ \overline{U}_k := \max(h_k(\overline{X}_k),\ \mathbb{E}(\overline{U}_{k+1} + f_k(\overline{X}_{k+1},\ \overline{U}_{k+1})|\mathcal{F}_{kT/n})),$$
$$0 \leqslant k \leqslant n-1. \quad (48)$$

Throughout this section, we denote by $\bar{u}_k^{(n)}$ the function satisfying $\overline{U}_k := \bar{u}_k^{(n)}(\overline{X}_k)$. Lemma 2 below shows that the Lipschitz coefficients of functions $u_k^{(n)}$ and $\bar{u}_k^{(n)}$ do not explode as n goes to infinity. Its (easy) proof is left to the reader.

Lemma 2. *Assume that* (5)–(7) *hold.*

(a) *Diffusion. The* (h, f)*-Snell envelope* $(U_k)_{0 \leqslant k \leqslant n}$ *defined by* (11) *satisfies* $U_k = u_k^{(n)}(X_{t_k})$ *with*

$$[u_k^{(n)}]_{\mathrm{Lip}} \leqslant K_1 \exp(K_0(T - t_k)) + \frac{\varepsilon_{n,k}}{n},$$

where

$$K_1 := \gamma_0 + \frac{1}{1+\gamma_0/2}, \qquad K_0 := \gamma_0(2 + \gamma_0/2), \qquad \lim_n \sup_{0 \leqslant k \leqslant n} |\varepsilon_{n,k}| = 0.$$

If, furthermore, the transition is asymptotically flat with parameter $a > \gamma_0$*, then*

$$[u_k^{(n)}]_{\mathrm{Lip}} \leqslant K_2 + \frac{\varepsilon_{n,k}}{n}. \qquad (49)$$

with $K_2 := \gamma_0 \max(1,\ 1/(a - \gamma_0))$ *and* $\lim_n \sup_{0 \leqslant k \leqslant n} |\varepsilon_{n,k}| = 0$. *When* f *does not depend on u (regular optimal stopping problems) then* $K_2 := \gamma_0 \max(1,\ 1/a)$ *and* (49) *holds if the diffusion is asymptotically flat.*

(b) *Euler scheme. The* (h, f)*-Snell envelope* $(\overline{U}_k)_{0 \leqslant k \leqslant n}$ *defined by* (14) *satisfies* $\overline{U}_k = \bar{u}_k^{(n)}(\overline{X}_k)$ *and the functions* $u_k^{(n)}$ *are Lipschitz continuous. The same bounds as those obtained for the Lipschitz coefficients of* $u_k^{(n)}$ *in part (a) hold.*

Remark 9. In the asymptotically flat case, the minimal assumption on f is $a > [f]_{\mathrm{Lip}}$.

Remark 10. A less precise statement of Lemma 2 could be: there exist real constants $\tilde{K}_0, \tilde{K}_1 > 0$ depending only on b, σ, h and f, such that,

$$\forall n \geqslant 1, \forall k \in \{0, \ldots, n-1\}, \qquad [u_k^{(n)}]_{\mathrm{Lip}} \leqslant \tilde{K}_1\, \mathrm{e}^{\tilde{K}_0(T - t_k)}.$$

Furthermore, if the diffusion is 'asymptotically flat' enough, one may set $\tilde{K}_0 := 0$.

Lemma 2 yields the following bounds for the coefficients $d_i^{(n)}$ defined by equation (27).

Proposition 9. *Let* $p \geqslant 1$*. For every fixed n and every* $k \in \{0, \ldots, n\}$*,*

$$\forall\, i \in \{k, \ldots, n\}, \qquad \frac{d_i^{(n)}}{(1 + T/n\gamma_0)^k} = (\gamma_0 + (2 - \delta_{2,p})K_1\, \mathrm{e}^{K_0(T - t_i)})\mathrm{e}^{\gamma_0(t_i - t_k)} + \frac{\varepsilon_{i,k,n}}{n}$$

where

$$\lim_n \max_{0 \leqslant k \leqslant i \leqslant n} |\varepsilon_{i,k,n}| = 0$$

so that

$$d^\infty := \sup_{n \geqslant 1} \max_{0 \leqslant i \leqslant n} d_i^{(n)} < +\infty.$$

We use the notation for RBSDEs introduced in Section 1. The aim here is to derive some a priori error bounds for $\|Y_{t_k} - \hat{u}_k(\hat{X}_k)\|_p$ as a function of the total number N of elementary quantizers and of the number n of time steps, all real constants depending on b, σ, h and T. To this end, we will need some precise estimates for the probability densities of the diffusion process $(X_t)_{t\in[0,T]}$ in the uniformly elliptic case. Assume that diffusion parameters b and σ satisfy the assumptions

$$\begin{array}{l} \sigma\sigma^* \geqslant \varepsilon_0 I_d,\, \varepsilon_0 > 0 \qquad \text{(uniform ellipticity)}, \\ (b,\, \sigma \in \mathcal{C}_b^\infty(\mathbb{R}^d)) \text{ or } (b,\, \sigma \text{ Lipschitz continuous and bounded}). \end{array} \tag{50}$$

Then, there exist two real constants $\alpha, \beta > 0$ such that the diffusion process $(X_t^x)_{t\in[0,T]}$ starting at x has a probability density function $p_t(x, y)$ at every time $t \in [0, T]$ satisfying

$$p_t(x, y) \leqslant \frac{\alpha}{(2\pi t)^{d/2}} \exp\left(-\frac{|y-x|^2}{2\beta t}\right). \tag{51}$$

The *bounded Lipschitz setting* is due to Friedman (1975, Theorems 4.5 and 5.4), the *smooth sublinear setting* follows from Kusuoka and Stroock (1985). When the diffusion is only hypoelliptic and satisfies a non-degeneracy Hormander type assumption, it is also established in Kusuoka and Stroock (1985) that (51) holds with exponent $d' \geqslant d/2$. This could lead to different dispatching rules (provided that d' is known). An alternative approach could be to rely on similar results established by Bally and Talay (1996) for an 'excited' version of the Euler scheme.

Theorem 4. *Assume* (5)–(7) *hold, that the diffusion* $(X_t)_{t\in[0,T]}$ *satisfies* (50) *and* $X_0 = x$. *For* $N \geqslant n+1$, *assign*

$$N_k := \left\lceil \frac{t_k^{d/(2(d+1))} N}{t_1^{d/(2(d+1))} + \ldots + t_n^{d/(2(d+1))}} \right\rceil \geqslant 1$$

elementary quantizers to the optimal quantization grid Γ_k *of* $\overline{X}_k$ *or* X_{t_k}, $1 \leqslant k \leqslant n$, *and set* $N_0 = 1$. *(Note that in fact* $N \leqslant N_0 + \ldots + N_n \leqslant N + n + 1$.*)*

(a) *Diffusion. Let* $\hat{X}_k$ *denote the optimal* L^p*-quantization of the diffusion* X_{t_k}. *Then,*

$$\forall\, p \geqslant 1, \qquad \max_{0 \leqslant k \leqslant n} \|Y_{t_k} - u_k(\hat{X}_k)\|_p \leqslant C_p\, \mathrm{e}^{C_p T}\left(\frac{n^{1+1/d}}{N^{1/d}} + \frac{1+|x|}{n^\theta}\right), \tag{52}$$

where $\theta = 1$ *if* h *is semi-convex and* f *is* $C_{\mathrm{b}}^{1,2,2}$, *and* $\theta = \frac{1}{2}$ *otherwise.*

(b) *Euler scheme. Let* $\hat{X}_k$ *denote the optimal* L^p*-quantization of the Euler scheme* $\overline{X}_k$. *Then, the above error bound holds with* $\theta = \frac{1}{2}$.

Some practical comments about the dispatching are in order. First, one verifies using $t_k = kT/n$ that

$$N_k \sim \frac{3d+2}{2(d+1)}\left(\frac{k}{n}\right)^{d/(2(d+1))}\frac{N}{n} \qquad \text{as } n \to +\infty,\ N \to +\infty, \qquad \text{with } n = o(N).$$

Note that the optimized ratio N_k/N of the elementary quantifiers assigned to time t_k marginally depends upon the dimension d since

$$\frac{N_k}{N} \approx \frac{3}{2n}\sqrt{\frac{k}{n}} \qquad \text{when } d \text{ becomes 'large' (say } d \geqslant 5). \tag{53}$$

Then the (theoretical) complexity of the algorithm is approximately $9N^2\kappa/(8n)$, which is close to the lowest possible (see (24)). Note that, for example in the Lipschitz continuous setting, if $N := \lceil n^{1+3d/2} \rceil$,

$$\max_{0\leqslant k\leqslant n}\|Y_{t_k} - \hat{u}_k(\hat{X}_k)\|_p \leqslant \frac{C_p\,\mathrm{e}^{C_pT}(1+|x|)}{\sqrt{n}}.$$

Proof of Theorem 4. (a) One derives from Proposition 3 and Theorem 2 applied to the diffusion $(X_{t_k})_{0\leqslant k\leqslant n}$ that, for every $p \geqslant 1$,

$$\max_{0\leqslant k\leqslant n}\|Y_{t_k} - U_k\|_p \leqslant \frac{C_p(x)}{n^\theta}$$

with $C_p(x) := C_p\,\mathrm{e}^{C_pT}(1+|x|)$, $C_p > 0$, and

$$\max_{0\leqslant k\leqslant n}\|U_k - \hat{U}_k\|_p \leqslant \sum_{k=0}^{n} d_k^{(n)}\|X_{t_k} - \hat{X}_k\|_p.$$

Proposition 9 shows that, for every $n \geqslant 1$ and every $k \in \{0, \ldots, n\}$,

$$d_k^{(n)} = \gamma_0(1 + C_{b,\sigma,p}\,\mathrm{e}^{(\gamma_0+C_{b,\sigma})(T-t_k)})\mathrm{e}^{\gamma_0 t_k} + \frac{\varepsilon_{k,n}}{n} \leqslant C_{\gamma_0,T,p} + \frac{|\varepsilon_{k,n}|}{n}$$

with

$$C_{\gamma_0,T,p} := \gamma_0\,\mathrm{e}^{\gamma_0 T}\left(1 + \frac{(2-\delta_{2,p})(1+C_{\gamma_0})}{C_{\gamma_0}}\,\mathrm{e}^{(\gamma_0+C_{\gamma_0})T}\right), \qquad C_{\gamma_0} := \gamma_0(1+\gamma_0/2),$$

and

$$\lim_n \max_{0\leqslant k\leqslant n}|\varepsilon_{k,n}| = 0.$$

First, set $\hat{X}_0 := X_0 = x$. Setting $\varepsilon_n := \max_{0\leqslant k\leqslant n}|\varepsilon_{k,n}|$, one has

$$N^{1/d} \max_{0 \leqslant k \leqslant n} \| Y_{t_k} - \hat{u}_k(\hat{X}_k) \|_p \leqslant \left(C_{\gamma_0, T, p} + \frac{\varepsilon_n}{n} \right) \sum_{k=0}^{n} \left(\frac{N}{N_k} \right)^{1/d}$$

$$\times \; N_k^{1/d} \min_{|\Gamma| \leqslant N_k} \| X_{t_k} - \hat{X_{t_k}}^{\Gamma} \|_p + \frac{C_p(x)}{n^{\theta}}.$$

Now, the Gaussian domination inequality (51) implies that, for every $k \in \{1, \ldots, n\}$,

$$\forall x,\, y \in \mathbb{R}^d, \qquad p_{t_k}(x,\, y) \leqslant \alpha \beta^{d/2} \pi_{x + \sqrt{\beta_{t_k}} Z}(y),$$

where $Z \overset{\mathcal{L}}{\sim} \mathcal{N}(0, I_d)$ and $\pi_{Y}(x)$ is for the probability density function of a random vector Y. Hence, for every $k \in \{1, \ldots, n\}$, $N_k \geqslant 1$, and every grid $\Gamma := \{v_1, \ldots, v_{N_k}\} \subset \mathbb{R}^d$ of size N_k,

$$\| X_{t_k} - \hat{X}^{\Gamma}_{t_k} \|_p \leqslant \alpha \beta^{d/2} \| \min_{1 \leqslant \ell \leqslant N_k} | v_\ell - x - \sqrt{\beta t_k}\, Z | \, \|_p, \tag{54}$$

$$= \alpha \beta^{d/2} \sqrt{\beta t_k} \, \| Z - \hat{Z}^{(\Gamma - x)/\sqrt{\beta t_k}} \|_p.$$

Hence

$$\min_{\Gamma, |\Gamma| \leqslant N_k} \| X_{t_k} - \hat{X}^{\Gamma}_{t_k} \|_p \leqslant \alpha \beta^{d/2} \sqrt{\beta t_k} \min_{|\Gamma| \leqslant N_k} \| Z - \hat{Z}^{\Gamma} \|_p. \tag{55}$$

Applying Theorem 3 to Z (i.e. to the normal distribution $\mathcal{N}(0, I_d)$ on $\mathbb{R}^d$) yields

$$\min_{|\Gamma| \leqslant N_k} \| Z - \hat{Z}^{\Gamma} \|_p \leqslant (1 + \eta_N)^{1/p} \tilde{J}_{p,d} \, \| \pi_Z \|^{1/p}_{d/(d+p)} = (1 + \eta_N)^{1/p} \tilde{J}_{p,d} \, (1 + p/d)^{(d+p)/2p} \sqrt{2\pi}. \tag{56}$$

where $\tilde{J}_{p,d} := J^{1/p}_{p,d}$ and $\lim_N \eta_N = 0$. Combining (55) and (56) yields

$$N_k^{1/d} \min_{|\Gamma| \leqslant N_k} \| X_{t_k} - \hat{X}^{\Gamma}_{t_k} \|_p \leqslant \alpha \beta^{(d+1)/2} \tilde{J}_{p,d} (1 + \eta_{N_k})^{1/p} (1 + p)^{(d+p)/2p} \sqrt{2\pi t_k},$$

$$k \in \{0, \ldots, n\}.$$

$$N^{1/d} \max_{0 \leqslant k \leqslant n} \| Y_{t_k} - \hat{u}_k(\hat{X}_k) \|_p \leqslant \left(C_{\gamma_0, T, p} + \frac{\varepsilon_n}{n} \right) \left(1 + \max_{1 \leqslant k \leqslant n} \eta_{N_k} \right) \sqrt{2\pi} \tilde{J}_{p,d} \left(1 + \frac{p}{d} \right)^{(d+p)/2p}$$

$$\times \; \beta^{(d+1)/2} \sum_{k=1}^{n} \left(\frac{N}{N_k} \right)^{1/d} \sqrt{t_k} + \frac{C_p(x) N^{1/d}}{n^{\theta}}$$

$$\leqslant C_{n, \gamma_0, T, p, d} \sum_{1 \leqslant k \leqslant n} \rho_k^{-1/d} \sqrt{t_k} + \frac{C_p(x) N^{1/d}}{n^{\theta}},$$

where $\rho_k \propto t_k^{d/(2(d+1))}$, $1 \leqslant k \leqslant n$, $\rho_1 + \ldots + \rho_n = 1$ and $N_k := \lceil \rho_k N \rceil \geqslant 1$, and $C_{n, \gamma_0, T, p, d}$ is bounded as $n \to \infty$ by a real constant $\overline{C}_{\gamma_0, T, p, d}$. Following Lemma 1,

$$N^{1/d}\max_{0\leqslant k\leqslant n}\|Y_{t_k}-\hat{u}_k(\hat{X}_k)\|_p\leqslant \overline{C}_{\gamma_0,T,p,d}\left(\sum_{k=1}^{n}t_k^{d/(2(d+1))}\right)^{1+1/d}+\frac{C_p(x)N^{1/d}}{n^{\theta}}.$$

Now, Jensen's inequality implies that $\sum_{1\leqslant k\leqslant n}t_k^{d/(2(d+1))}\leqslant T^{d/(2(d+1))}n$. Setting $\overline{C}_p:=\mathrm{T}^{d/(2(d+1))}\overline{C}_{\gamma_0,T,p,d}$ yields the bound (52) by setting C_p at the appropriate value.

(b) The main modification lies in the above inequality (54). With the same notation,

$$\begin{aligned}\|\overline{X}_k-\hat{\hat{X}}{}^{\Gamma}_k\|_p&=\|\min_{1\leqslant\ell\leqslant N_k}|v_\ell-\overline{X}_k|\,\|_p\\&\leqslant\|\min_{1\leqslant\ell\leqslant N_k}|v_\ell-X_{t_k}|^p\|_p+\|X_{t_k}-\overline{X}_k\|_p\\&\leqslant\|X_{t_k}-\hat{X}^{\Gamma}_{t_k}\|_p+C_p\,\mathrm{e}^{C_pT}(1+|x|)\frac{1}{\sqrt{n}},\end{aligned}$$

using classical L^p-error bounds for the Euler scheme. The rest of the proof is as before. □

One can easily derive an optimized choice for the number n of time steps.

Corollary 1. (a) *Lipschitz setting (quantization of the Euler scheme or of the diffusion). The optimal number n of time steps and the resulting error bound satisfy*

$$n\approx\left(\frac{2d}{d+1}C_p(x)\right)^{2/(3d+2)}N^{2/(3d+2)}\quad\textit{and}\quad|u_0(x)-\hat{u}_0(x)|=O(N^{-1/(3d+2)})=O\left(\frac{1}{\sqrt{n}}\right),$$

where $C_p(x)\leqslant C_p\,\mathrm{e}^{C_pT}(1+|x|)$.

(b) *Semi-convex setting (quantization of the diffusion). The optimal number n of time steps and the resulting error bound satisfy*

$$n\approx\left(\frac{d}{d+1}C_p(x)\right)^{d/(2d+1)}N^{1/2(d+1)}\quad\textit{and}\quad|u_0(x)-\hat{u}_0(x)|=O(N^{-1/(2d+1)})=O\left(\frac{1}{n}\right).$$

6.2. Numerical pricing of American exchange options by quantization

6.2.1. The test model

One considers two risky assets, a stock S^1 with a geometric dividend rate λ and a stock S^2 without dividend. The interest rate r is deterministic and constant. Assume that (S^1, S^2) follows a Black–Scholes dynamics, so that, under the *risk-neutral* probability $\mathbb{P}$, one has

$$\mathrm{d}^1_t=S^1_t((r-\lambda)\,\mathrm{d}t+\sigma_1\mathrm{d}B^1_t),\qquad S^1_0:=s^1_0>0,$$

$$\mathrm{d}^2_t=S^2_t(r\,\mathrm{d}t+\sigma_2\mathrm{d}B^2_t),\qquad S^2_0:=s^2_0>0,$$

where (B^1, B^2) is a two-dimensional Brownian motion with covariance $\langle B^1, B^2\rangle_t=\rho\,t$, $\rho\in[-1, 1]$. One can verify that the discounted traded assets

$$\tilde{S}^1_t := \mathrm{e}^{-rt}(\mathrm{e}^{\lambda t} S^1_t) \quad \text{and} \quad \tilde{S}^2_t := \mathrm{e}^{-rt} S^2_t, \qquad t \in [0, T],$$

make up a two-dimensional $\mathbb{P}$-martingale with respect to the filtration $\underline{\mathcal{F}}^B$ of the two-dimensional Brownian motion $B := (B^1, B^2)$. The diffusion $S := (S^1, S^2)$ is obviously not *uniformly* elliptic, but $(\ln S^1, \ln S^2)$ clearly is.

An American exchange option with exchange rate μ is the right to exchange once and only once, at any time $t \in [0, T]$, μ units of asset S^2 for one unit of asset S^1. Or, to put it the other way round, the right to buy one unit of asset S^1 for μ units of asset S^2.

The discounted premium of such an option is defined as the $(h, 0)$-Snell envelope of

$$h_t := \mathrm{e}^{-rt}(S^1_t - \mu S^2_t)_+ = \max(\mathrm{e}^{-\lambda t}\tilde{S}^1_t - \mu \tilde{S}^2_t, 0),$$

where x_+ denotes the positive part of the real number x. Since h_t does not depend upon the interest rate r, we may assume without loss of generality that $r := 0$. So, if $\mathcal{E}x_t$ denotes the premium of this exchange American option at time t,

$$\mathcal{E}x_t := \operatorname{ess\,sup}_{t \leqslant \tau \leqslant T} \mathbb{E}(h_\tau | \mathcal{F}_t) := \mathcal{E}(t, S^1_t, S^2_t, \rho, \sigma_1, \sigma_2, \lambda).$$

One noticeable feature of this derivative is that the premium of its European counterpart, that is, the right to exchange *at time* T, μS^2 and S^1, admits a closed form given by

$$Ex_t := E(T - t, S^1_t, \mu S^2_t, \tilde{\sigma}, \lambda), \qquad \tilde{\sigma} := \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

where

$$E(t, x, y, \sigma, \lambda) := x\,\mathrm{e}^{-\lambda t}\operatorname{erf}(d_1(t, x, y, \sigma, \lambda)) - y\operatorname{erf}(d_1(t, x, y, \sigma, \lambda) - \sigma\sqrt{t}),$$

$$d_1(t, x, y, \sigma, \lambda) := \frac{\ln(x/y) + (\sigma^2/2 - \lambda)t}{\sigma\sqrt{t}}, \qquad \operatorname{erf}(x) := \int_{-\infty}^{x} \mathrm{e}^{-u^2/2}\frac{\mathrm{d}u}{\sqrt{2\pi}}.$$

American exchange options have two characteristics: the (discounted) contingent claim h_t only depends on the state variable $S_t := (S^1_t, S^2_t)$ at time t, and the European option related to h_T has a closed form. For such options, one uses the premium of the European option as a control variate to reduce the computations to the 'residual' American part of the option. That is to say, keeping in mind that $r = 0$, set

$$\forall\, t \in [0, T], \qquad k_t := h_t - Ex(T - t, S^1_t, \mu S^2_t, \tilde{\sigma}, \lambda).$$

Since $(E(T - t, S^1_t, \mu S^2_t, \tilde{\sigma}, \lambda))_{t \in [0,T]}$ is a $\mathbb{P}$-martingale, it follows that

$$\mathcal{E}x_t = \operatorname{ess\,sup}_{t \leqslant \tau \leqslant T} k_\tau + Ex_t.$$

So, pricing the American exchange option amounts to pricing the American option having k_t as discounted contingent claim. The interesting fact for numerical purposes is that $k_T \equiv 0$ and that k_t is always smaller than h_t. Moreover, standard computations show that one may write $k_t := k(\ln \tilde{S}^1_t, \ln \tilde{S}^2_t)$, where k is a Lipschitz continuous function.

6.2.2. Practical implementation and results

In a Black–Scholes model, exact simulation of $(S_{t_k})_{0\leqslant k\leqslant n}$ at times $0=: t_0 < t_1 < t_2 < \ldots < t_k < \ldots < t_n := T$ is possible. In fact, for numerical purposes, it is more convenient to consider the couple $(\ln \tilde{S}^1_t, \ln \tilde{S}^2_t)$ as the underlying variables of the pricing problem. One simulates this process recursively at times $t_k := kT/n$, $0 \leqslant k \leqslant n$, by setting $\tilde{S}^1_0 := s^1_0 > 0$, $\tilde{S}^2_0 := s^2_0 > 0$ and, for every $k \in \{0, n \ldots, n-1\}$,

$$\ln(\tilde{S}^1_{t_{k+1}}) := \ln(\tilde{S}^1_{t_k}) + \left(-\frac{\sigma_1^2}{2}\Delta + \sigma_1\sqrt{\Delta}\,\varepsilon_{2k}\right),$$

$$\ln(\tilde{S}^2_{t_{k+1}}) := \ln(\tilde{S}^2_{t_k}) + \left(-\frac{\sigma_2^2}{2}\Delta + \sigma_2\sqrt{\Delta}(\rho\,\varepsilon_{2k} + \sqrt{1-\rho^2}\,\varepsilon_{2k+1})\right)$$

where ε_k is an i.i.d. sequence of normal random variables and $\Delta := T/n$.

We carried out a simulation in which the parameters of the options were set as follows $T := 1$ (1 year), $\sigma_1 = \sigma_2 := 20\%$, $\lambda := 5\%$, $S^1_0 = 40$, $S^2_0 := 36$ or 44, $\mu := 1$ and $\rho \in \{-0.8; 0; 0.8\}$ (the price does not depend upon the interest rate r).

The quantization was processed with $N := 5722$ $\mathbb{R}^2$-valued elementary quantizers dispatched on 25 layers using the optimal dispatching rule (41). The estimation of the weights (and of the quadratic quantization error) was carried out using $M := 10^6$ trials in the CLVQ algorithm. The reference solution labelled VZ in Table 1 computed by a two-dimensional finite-difference algorithm devised by Villeneuve and Zanette (2002).

The ‘quantization’ premium of European style options was computed using the number N_{25} of elementary quantizers on the last (25th) layer, namely $N_{25} := 299$. The numerical experiments were carried out with $\mu = 1$. Of course, once the quantization is performed, the pricing of any American style options for any (reasonable) value of μ simply needs to rerun the (pseudo-)dynamic programming formula whose CPU cost is negligible. One could parametrize the starting values s^1_0 and s^2_0 similarly. Figure 4 displays the global results obtained for 25 maturities from (approximately) 2 weeks up to 1 year.

One important and promising fact is that quite similar results have been obtained by directly quantizing the standard Brownian motion itself instead of the geometric Brownian motions of the above Black–Scholes model. The first noticeable fact is that the functions h_k are no longer Lipschitz continuous: this speaks for the robustness of the method. This robustness of the method could play a role in the future development of this approach to

Table 1. ‘Quantization premia’ versus ‘reference premia’ for some European and American exchange options

$S^2_0 := 36$ ρ	-0.8	0	0.8	$S^2_0 := 44$ ρ	-0.8	0	0.8
Euro. *B & S*	6.6547	5.2674	3.0674	Euro. *B & S*	3.6390	2.2289	0.3217
Euro. Quantiz.	6.6297	5.2558	3.0639	Euro. Quantiz.	3.6133	2.2117	0.3151
Am. VZ	6.9754	5.6468	4.0000	Am. VZ	3.7692	2.3364	0.3595
Am. Quantiz.	6.9812	5.6520	4.000	Am. Quantiz.	3.7726	2.3398	0.3610

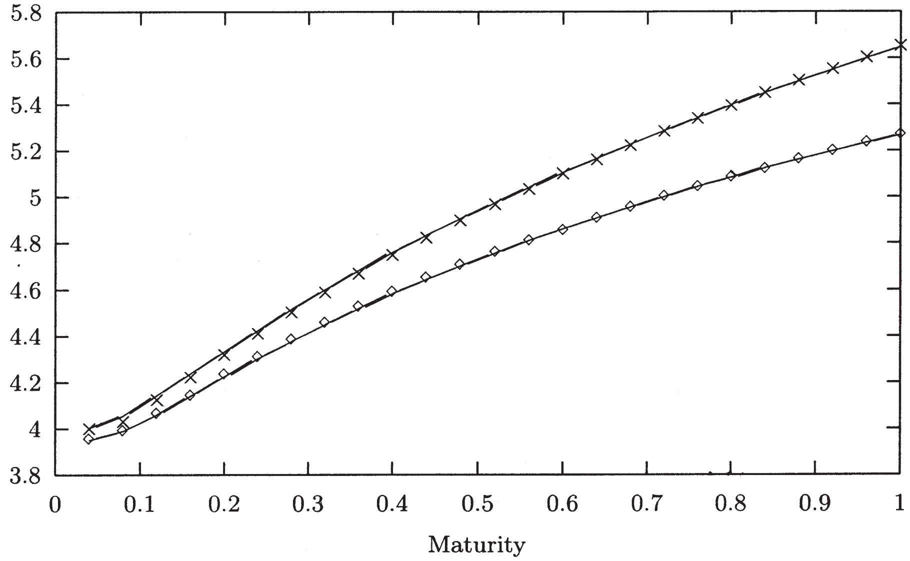

Figure 4. American and European style option prices as a function of the maturity ($S_0^1 := 40$, $S_0^2 := 36$, $\rho := 0$). Dashes depict the reference prices (V&Z for American style and B&S for European style options); crosses depict the quantization price for American style options; and diamonds depict the quantization price for European style options

finance since the quantization of a standard Brownian motion is parameter-free (except for the dimension!). Furthermore, the quantization of every layer can be achieved, up to an appropriate dilatation, by using some precomputed tables of optimal quantization of the standard d-dimensional Gaussian vector. (Generating such optimal quantization tables is done by a splitting method. At each step, 10^6 CLVQ trials are processed. The CPU time for producing an optimal $(N+1)$-grid from an optimal N-grid increases from 5 s (if $N = 25$) to 60 s ($N = 300$).) At this stage, if using these precomputed grids, all that remains is to estimate the transition weights π_{ij}^k using a standard Monte Carlo simulation. Exploratory experiments showed that 10 000 trials are enough to obtain results as accurate as above for American exchange options (the total CPU time required for this weight estimation is then 25 s). The quantization tree descent itself is instantaneous.

These results augur well of the future comparisons with former pricing methods for multidimensional American options (see Broadie and Glasserman 1997; Longstaff and Schwartz 2001). The simulations were performed by J. Printems (Univ. Paris 12). Some extensive numerical investigations have been carried out in Bally *et al*. (2002b).

6.2.3. Provisional remarks

There is a possible alternative to optimal quantization: one may build the grids of the quantization tree by settling the first $\overline{N} = N/n$ random paths of the Markov chain. The resulting theoretical quantization errors almost surely still follow a $O(\overline{N}^{-1/d})$-rate, with a worse sharp rate (see Cohort 2003). The companion parameter estimation is carried out by a standard Monte Carlo simulation.

Some developments (quantized hedging) are proposed in Bally *et al*. (2002b), and some first-order schemes based on correctors obtained using Malliavin calculus are proposed in Bally *et al*. (2003).

Appendix. Partial almost sure convergence results for the CLVQ algorithm

As the distortion $D_N^{X,2}$ does not enjoy the standard properties of a potential for a stochastic gradient, we are led to make the following restrictive assumption:

$$\text{supp}(\mathbb{P}_X) \text{ is a compact set.} \tag{57}$$

Hence, the convex hull of supp($\mathbb{P}_X$) is a convex compact set.

The one-dimensional case

In this very special setting, standard stochastic approximation theory applies and we obtain a satisfactory almost sure convergence result. The convex hull of supp($\mathbb{P}_X$) is an interval $[a, b]$. One first notices that the set of N-grids is one-to-one with the simplex

$$\Sigma_N^{a,b,+} := \{\Gamma := (x_1, \ldots, x_N) \in (a, b)^N \mid a < x_1 < \ldots < x_N < b\},$$

which is invariant for the algorithm provided that the starting value $\Gamma^0 \in \Sigma_N^{a,b,+}$. Then,

$$\forall\, \Gamma := (x_1, \ldots, x_N) \in \Sigma_N^{a,b,+}, \qquad \nabla D_N^{X,2}(\Gamma) := 2\left(\int_{\tilde{x}_i}^{\tilde{x}_{i+1}} (x_i - \xi)\,\mathbb{P}_X(\mathrm{d}\xi)\right)_{1\leqslant i\leqslant N}, \tag{58}$$

where $\tilde{x}_1 := a$, $\tilde{x}_i := (x_i + x_{i-1})/2$, $\tilde{x}_{N+1} := b$. Consequently, the distribution $\mathbb{P}_X$ being continuous (*i.e*. no single ξ is weighted), $\nabla D_N^{X,2}$ has a *continuous extension* on the compact set $\overline{\Sigma}_n^{a,b,+}$.

Theorem 5 (Pagès 1997). (*a*) *If* $\Gamma^0 \in \Sigma_N^{a,b,+}$, *then the algorithm* (37) *lives in* $\Sigma_N^{a,b,+}$. *Furthermore, if* $\mathbb{P}_X$ *is continuous, if* $\{\nabla D_N^{X,2} = 0\} \cap \overline{\Sigma}_N^{a,b,+}$ *is finite and if the step* γ_t *satisfies the usual decreasing step assumption* $\sum_{t\geqslant 1} \gamma_t = +\infty$ *and* $\sum_{t\geqslant 1} \gamma_t^2 < +\infty$, *then*

$$\Gamma^t \xrightarrow{a.s} \Gamma^* \in \{\nabla D_N^{X,2} = 0\}.$$

(*b*) *Moreover, if* $\mathbb{P}_X$ has a log-concave density *then* $\{\nabla D_N^{X,2} = 0\} \cap \overline{\Sigma}_N^{a,b,+} =$ argmin$_{\overline{\Sigma}_N^{a,b,+}} D_N^{X,2} = \{\Gamma^*\}$ *(see Kieffer 1982; Trushkin 1982; Lamberton and Pagès 1996; Graf and Luschgy 2000), with*

$$\Gamma^* = \left\{a + \frac{2k-1}{2N}(b-a),\ 1 \leqslant k \leqslant N\right\}$$

if $X \sim U([0, 1])$.

The multidimensional case ($d \geqslant 2$)

In the multidimensional setting, only partial results are proved, even for bounded distributions μ. We observe once again that the singularity of $\nabla D_N^{X,2}$ makes the standard theory inefficient. The result below follows from a specific proof. The assumption on the distribution of the stimuli ξ^t is now the following:

$$\mathbb{P}_X \text{ has a bounded density } f \text{ with a compact convex support.} \tag{59}$$

Roughly speaking, Theorem 6 below says that, almost surely, either the N components of Γ^t remain parted and converge to some stationary quantizer of $D_N^{X,2}$, or they get stuck into M aggregates which will converge, up to some subsequence, towards a stationary quantizer of $D_{_M}^{X,2}$. This is of very little help for simulations since it does not say precisely how often the elementary quantizers remain parted. In the worst case – all the components get stuck into a single aggregate at the average of $\mathbb{P}_{_X}$ – the resulting quantization would be simply useless. These partial theoretical results seem very pessimistic in view of the practical performance of the CLVQ: no aggregation phenomenon is usually observed when the algorithm is appropriately initialized (random or splitting method).

Theorem 6 (Pagès 1997). *Assume that* (59) *holds and that the step sequence satisfies* $\sum_{t\geqslant 1}\gamma_t = +\infty$ *and* $\sum_{t\geqslant 1}\gamma_t^2 < +\infty$.

(a) $\mathbb{P}$*-almost surely, either the elements of* Γ^t *remain asymptotically parted or at least two elements of* Γ^t *get asymptotically stuck as* $t \to +\infty$*, that is,*

$$\underline{\lim}_t \operatorname{dist}(\Gamma^t, {}^c\mathcal{S}_{_N}) > 0 \quad \textit{or} \quad \lim_t \operatorname{dist}(\Gamma^t, {}^c\mathcal{S}_{_N}) = 0,$$

where $\mathcal{S}_N$ *denotes the set of N-tuples with pairwise distinct components.*

(b) *On the event* $\{\underline{\lim}_t \operatorname{dist}(\Gamma^t, {}^c\mathcal{S}_N) > 0\}$ *(asymptotically parted components), there exists a 'level'* $\delta^* > 0$ *and a connected component* Γ^* *of* $\{\nabla D_N^{X,2} = 0\} \cap \{D_N^{X,2} = \delta^*\}$ *such that*

$$\Gamma^t \xrightarrow{a.s.} \Gamma^* \qquad \textit{as } t \to +\infty.$$

(c) *On the event* $\{\underline{\lim}_t \operatorname{dist}(\Gamma^t, {}^c\mathcal{S}_N) = 0\}$*, the components definitely get stuck, that is,*

$$\operatorname{dist}(\Gamma^t, {}^c\mathcal{S}_N) \overset{t\to+\infty}{\longrightarrow} \qquad \textit{as } t \to +\infty.$$

There is a partition $I_1 \cup \ldots \cup I_M$ *of* $\{1, \ldots, N\}$ *along which the components* Γ^t *make* M *aggregates as* $t \to +\infty$*. At least one of the limiting values of* Γ^t *is a zero of* $\nabla D_M^{X,2}$.

References

Abaya, E. and Wise, G. (1992) On the existence of optimal quantizers. *IEEE Trans. Inform. Theory*, **38**, 937–946.

Bally, V., Caballero, M.E., Fernandez, B. and El Karoui, N. (2002a) Reflected BSDE's, PDEs and variational inequalities. Technical Report no. 4455, Project MATHFI, INRIA, Le Chesnay, France.

Bally, V. (2002) The central limit theorem for a nonlinear algorithm based on quantization. Technical Report no. 4629, Project MATHFI, INRIA, Le Chesnay, France.

Bally, V. and Pagès, G. (2003) Error analysis of the quantization algorithm for obstacle problems, *Stochastic Process. Appl.*, **106**, 1–40.

Bally, V., Pagès, G. and Printems, J. (2002b) A quantization tree method for pricing and hedging multi-dimensional American options. Technical Report no. 753, Laboratoire de Probabilités et modèles Aléatoires, Université de Paris 6, France.

Bally, V., Pagès, G. and Printems, J. (2003) First order schemes in the numerical quantization method. *Math. Finance*, **13**, 1–16.

Bally, V. and Saussereau, B. (2002) Approximation of the Snell envelope and computation of American option prices in dimension one. *ESAIM Probab. Statist.*, **16**, 1–21.

Bally, V. and Talay, D. (1996) The law of the Euler scheme for stochastic differential equations (II): Convergence rate of the density. *Monte Carlo Methods Appl.*, **2**, 93–128.

Bensoussan, A. and Lions, J.L. (1982) *Applications of Variational Inequalities in Stochastic Control*. Amsterdam: North-Holland.

Bouchard-Denize, B. and Touzi, N. (2002) Discrete time approximation and Monte-Carlo simulation of backward stochastic differential equations. Technical Report no. 766, Laboratoire de Probabilités et Modèles Aléatoires, Université de Paris 6, France.

Bouton, C. and Pagès, G. (1997) About the multidimensional Competitive Learning Vector Quantization algorithm with constant gain. *Ann. Appl. Probab.*, **7**, 679–710.

Brandière, O. and Duflo, M. (1996) Les algorithmes stochastiques contournent-ils les pièges? *Ann. Inst. H. Poincaré Probab. Statist.*, **32**, 395–427.

Briand, P., Delyon, B. and Mémin, J. (2001) Donsker-type theorem for BSDEs. *Electron. Comm. Probab.*, **6**, 1–14.

Briand, P., Delyon, B. and Mémin, J. (2002) On the robustness of backward stochastic differential equations. *Stochastic Process. Appl.*, **97**, 229–253.

Broadie, M. and Glasserman, P. (1997) Pricing American-style securities using simulation. *J. Econom. Dynam. Control*, **21**, 1323–1352.

Bucklew, J. and Wise, G. (1982) Multidimensional asymptotic quantization theory with *rth* power distortion measures. *IEEE Trans. Inform. Theory*, **28**, 239–247.

Caverhill, A.P. and Webber, N. (1990) American options: theory and numerical analysis. In S. Hodges (ed.), *Options: Recent Advances in Theory and Practice.* Manchester: Manchester University Press.

Chevance, D. (1997) Numerical methods for backward stochastic differential equations. In L. Rogers and D. Talay (eds), *Numerical Methods in Finance*. Cambridge: Cambridge University Press.

Cohort, P. (2003) Limit theorems for the random normalized distortion. *Ann. Appl. Probab.* To appear.

Duflo, M. (1997) *Random Iterative Models*, Appl. Math. 34. Berlin: Springer-Verlag.

El-Karoui, N., Kapoudjan, C., Pardoux, É., Peng, S. and Quenez, M.C. (1997a) Reflected solutions of backward stochastic differential equations and related obstacle problems for PDEs. *Ann. Probab.*, **25**, 702–737.

El Karoui, N., Peng, S. and Quenez, M.C. (1997b) Backward stochastic differential equations in finance. *Math. Finance*, **7**, 1–71.

Fort, J.C. and Pagès, G. (1995) On the a.s. convergence of the Kohonen algorithm with a general neighborhood function. *Ann. Appl. Probab.*, **5**, 1177–1216.

Fort, J.C. and Pagès, G. (2002) Asymptotics of optimal quantizers for some scalar distributions. *J. Comput. Appl. Math.*, **146**, 253–275.

Fournié, É., Lasry, J.M., Lebouchoux, J., Lions, P.L. and Touzi, N. (1999) Applications of Malliavin calculus to Monte-Carlo methods in finance. *Finance Stochastics*, **3**, 391–412.

Fournié, É., Lasry, J.M., Lebouchoux, J. and Lions, P.L. (2001) Applications of Malliavin calculus to Monte-Carlo methods in finance II. *Finance Stochastics*, **5**, 201–236.

Friedman, A. (1975) *Stochastic Differential Equations and Applications*, Vol. 1. New York: Academic Press.

Gersho, A. and Gray, R. (1992) *Vector Quantization and Signal Compression*. Boston: Kluwer.

Gersho, A. and Gray, R. (eds.) (1982) Special issue on Quantization. *IEEE Trans. Inform. Theory*, **28**.

Graf, S. and Luschgy, H. (2000) *Foundations of Quantization for Probability Distributions*, Lecture Notes in Math. 1730. Berlin: Springer-Verlag.

Kieffer, J. (1982) Exponential rate of convergence for the Lloyd's Method I. *IEEE Inform. Theory*, **28**, 205–210.

Kohatsu-Higa, A. and Pettersson, R. (2002) Variance reduction methods for simulation of densities on Wiener space. *SIAM J. Numer. Anal.*, **40**, 431–450.

Kushner, H.J. (1977) *Probability Methods for Approximation in Stochastic Control and for Elliptic Equations*. New York: Academic Press.

Kushner, H.J. and Yin, G.G. (1997) *Stochastic Approximation Algorithms and Applications*. New York: Springer-Verlag.

Kusuoka, S. and Stroock, D. (1985) Applications of the Malliavin calculus, part II. *J. Fac. Sci. Univ. Tokyo*, **32**, 1–76.

Lazarev, V.A. (1992) Convergence of stochastic approximation procedures in the case of a regression equation with several roots. *Problems Inform. Transmission*, **28**, 66–78.

Lamberton, D. (1998) Error estimates for the binomial approximation of the American put option. *Ann. Appl. Probab.*, **8**, 206–233.

Lamberton, D. (2002) Brownian optimal stopping and random walks. *Appl. Math. Optim.*, **45**, 283–324.

Lamberton, D. and Pagès, G. (1990) Sur l'approximation des réduites. *Ann. Inst. H. Poincaré Probab. Statist.*, **26**, 331–355.

Lamberton, D. and Pagès, G. (1996) On the critical points of the 1-dimensional Competitive Learning Vector Quantization algorithm. *Proceedings of the ESANN'96*, Bruges, Belgium.

Lamberton, D. and Rogers, L.C. (2000) Optimal stopping and embedding. *J. Appl. Probab.*, **37**, 1143–1148.

Lapeyre, B., Pagès, G. and Sab, K. (1990) Sequences with low discrepancy. Generalisation and application to Robbins–Monro algorithm. *Statistics*, **21**, 251–272.

Lions, P.L. and Régnier, H. (2002) Calcul des prix et des sensibilités d'une option américaine par une méthode de Monte Carlo. Working paper.

Longstaff, F.A. and Schwartz, E.S. (2001) Valuing American options by simulation: a simple least-squares approach. *Rev. Financial Stud.*, **14**, 113–148.

Lloyd, S.P. (1982) Least squares quantization in PCM. *IEEE Trans Inform. Theory*, **28**, 129–137.

Ma, J., Protter, P., San Martín, J. and Torres, S. (2002) Numerical method for backward stochastic differential equations. *Ann. Appl. Probab.*, **12**, 302–316.

Neveu, J. (1971) *Martingales à Temps Discret.* Paris: Masson.

Pagès, G. (1997) A space vector quantization method for numerical integration. *J. Computat. Appl. Math.*, **89**, 1–38.

Pagès, G. and Printems, J. (2003) Optimal quantization for numerics: the Gaussian case. *Monte Carlo Methods Appl.*, **9**, (2).

Pardoux, É. and Peng, S. (1992) Backward stochastic differential equations and quasi-linear parabolic partial differential equations. In B.L. Rozovskii and R.B. Sowers (eds), *Stochastic Partial Differential Equations and Their Applications*, Lecture Notes in Control and Inform. Sci. 176, pp. 200–217. Berlin: Springer-Verlag.

Pelletier, M. (2000) Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM J. Control Optim.*, **39**, 49–72.

Pemantle, R. (1990) Nonconvergence to unstable points in urn models and stochastic approximations. *Ann. Probab.*, **18**, 698–712.

Trushkin, A. (1982) Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions. *IEEE Trans. Inform. Theory*, **28**, 187–198.

Villeneuve, S. and Zanette, A. (2002) Parabolic A.D.I. methods for pricing American option on two stocks. *Math. Oper. Res.*, **27**, 121–149.

Zador, P. (1982) Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Trans Inform. Theory*, **28**, 139–148.

Received May 2001 Revised January 2003