Introduction to vector quantization and its applications for numerics*

G. Pagès

ESAIM: PROCEEDINGS AND SURVEYS, January 2015, Vol. 48, p. 29-79
N. Champagnat, T. Lelièvre, A. Nouy, Editors

INTRODUCTION TO VECTOR QUANTIZATION AND ITS APPLICATIONS FOR NUMERICS [*]

Gilles Pagès[1]

Abstract. We present an introductory survey to optimal vector quantization and its first applications to Numerical Probability and, to a lesser extent to Information Theory and Data Mining. Both theoretical results on the quantization rate of a random vector taking values in $\mathbb{R}^d$ (equipped with the canonical Euclidean norm) and the learning procedures that allow to design optimal quantizers ($CLVQ$ and Lloyd's procedures) are presented. We also introduce and investigate the more recent notion of *greedy quantization* which may be seen as a sequential optimal quantization. A rate optimal result is established. A brief comparison with Quasi-Monte Carlo method is also carried out.

1. Introduction to vector quantization

1.1. Signal transmission, information

The history of optimal vector quantization theory goes back to the 1950's in the Bell laboratories where researches were carried out to optimize signal transmission by appropriate discretization procedures. Two kinds of "stationary" signal can be naturally considered: either a deterministic – more or less periodic – signal, denoted by $(x_t)_{t\geq 0}$, or a stochastic signal, denoted by $(X_t)_{t\geq 0}$, considered under its stationary regime and supposed to be ergodic. In both cases, these signals share an averaging property as will be seen further on. Vector quantization can be briefly introduced as follows.

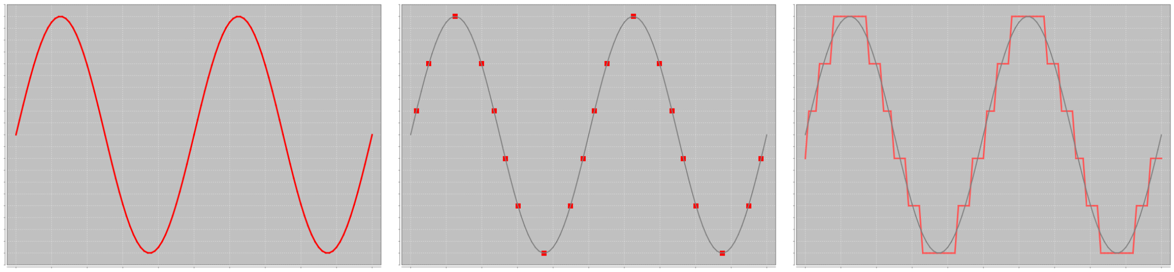

Figure 1. Quantization of a scalar (periodic) signal (B. Wilbertz)

[*] *The author thanks B. Jourdain and the referee for their careful reading of the manuscript and S. Graf for fruitful comments on source coding. All errors are mine.*

[1] Laboratoire de Probabilités et Modèles aléatoires, UMR 7599, UPMC, case 188, 4, pl. Jussieu, F-75252 Paris Cedex 5, France. E-mail: gilles.pages@upmc.fr

Article published online by EDP Sciences and available at http://www.esaim-proc.org or http://dx.doi.org/10.1051/proc/201448002

Let $\Gamma = \{x_1, \ldots, x_{_N}\}$ be a subset of $\mathbb{R}^d$ ($d \geq 1$) of size (at most) $N \geq 1$, called a *quantization grid* or simply a *quantizer* at *level* N if Γ has exactly cardinality N *i.e.* if the *elementary quantizers* x_i are pairwise distinct. When $d = 1$ the numbering of the elementary quantizers x_i is *a priori* made consistent with the natural order on the real line so that $i \mapsto x_i$ is non-decreasing.

In what follows, except specific mention, $|\,.\,|$ will denote the canonical Euclidean norm on $\mathbb{R}^d$ (although many of the stated results remain true or admit variants for more general norms).

A Γ-valued *quantization function* (also called *quantizer*) is simply a Borel function $q : \mathbb{R}^d \to \Gamma$. A naive idea is to transmit at time t the stochastic signal $q(X_t)$ instead of X_t itself inducing a resulting pointwise error

$$|X_t - q(X_t)|.$$

One proceeds likewise for a deterministic signal with a resulting error $|x_t - q(x_t)|$.

▷ *Deterministic signal*: Let $p \in (0, +\infty)$. Assume that the empirical measure $\dfrac{1}{t}\displaystyle\int_0^t \delta_{x(s)} ds$ weakly converges as $t \to +\infty$ toward a distribution μ on $(\mathbb{R}^d, \mathcal{B}or(\mathbb{R}^d))$ such that $\displaystyle\int_{\mathbb{R}^d} |\xi|^p \mu(d\xi) < +\infty$. If the quantization function q is μ-*a.s.* continuous and, *e.g.*, $\displaystyle\limsup_{t \to +\infty} \frac{1}{t}\int_0^t |x(s)|^r ds < +\infty$ for some $r > p$, then

$$\lim_{t \to +\infty} \left(\frac{1}{t}\int_0^t |x(s) - q(x(s))|^p ds \right)^{\frac{1}{p}} = \left(\int_{\mathbb{R}^d} |\xi - q(\xi)|^p \mu(d\xi) \right)^{\frac{1}{p}} = \|\xi - q(\xi)\|_{L^p(\mu)} < +\infty.$$

▷ *Stationary ergodic stochastic signal*: We consider again $p \in (0, +\infty)$. Assume the process $(X_t)_{t \geq 0}$ is stationary. Then, X_t has the same marginal distribution, say μ, for every $t \in \mathbb{R}_+$. Moreover, if $\mathbb{E}\,|X_t|^p = \int_{\mathbb{R}^d} |\xi|^p \mu(d\xi) < +\infty$, then

$$\|X_t - q(X_t)\|_{L^p(\mathbb{P})} = \|X_0 - q(X_0)\|_{L^p(\mathbb{P})} = \|\xi - q(\xi)\|_{L^p(\mu)} < +\infty.$$

Moreover, if the process $(X_t)_{t \geq 0}$ is ergodic, ergodic pointwise Birkhoff's Theorem ensures that

$$\mathbb{P}\text{-}a.s. \lim_{t \to +\infty} \left(\frac{1}{t}\int_0^t |X_s - q(X_s)|^p ds \right)^{\frac{1}{p}} = \|\xi - q(\xi)\|_{L^p(\mu)} < +\infty.$$

At this stage, several questions arise to optimize the transmission. Based on what precedes, we will mainly adopt from now on the *static* point of view of an $\mathbb{R}^d$-valued random vector X, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with distribution μ. It corresponds to the value of X_t at any time t or to the asymptotic behavior of the signal $(x(t))_{t \geq 0}$. More general situations of quantization or coding can be investigated in Information Theory which take into account the dynamics of the (ergodic) process leading to the most general Shannon's source coding theorem. For these deeper aspects from Information Theory, we refer to the general distortion theory as analyzed by large deviation methods in [23] and the references therein.

Question 1 *How to optimally choose the Γ-valued quantization function q (Geometric optimization)?*

It is clear that, whatever the quantization function $q : \mathbb{R}^d \to \Gamma$ is, one has

$$|\xi - q(\xi)| \geq \mathrm{dist}(\xi, \Gamma)$$

where $\mathrm{dist}(\xi, A) = \inf_{a \in A} |\xi - a|$ denotes the distance of ξ to the set $A \subset \mathbb{R}^d$ (with respect to the current norm). One easily checks that equality holds in the above inequality if and only if q is a Borel nearest neighbour projection *i.e.* $q = \pi_{_\Gamma}$ defined for every $\xi \in \mathbb{R}^d$ by

$$\pi_{_\Gamma}(\xi) = \sum_{i=1}^{N} x_i \mathbf{1}_{C_i(\Gamma)}(\xi)$$

where the N-tuple of subsets $\big(C_i(\Gamma)\big)_{1\le i\le N}$ is a Borel partition of $(\mathbb{R}^d, \mathcal{B}or(\mathbb{R}^d))$ satisfying

$$\forall\, i=1,\ldots,N, \quad C_i(\Gamma)\subset \Big\{\xi\in\mathbb{R}^d\ :\ |\xi-x_i| = \min_{1\le j\le N}|\xi-x_j|\Big\}.$$

Such a partition of $\mathbb{R}^d$ is called a *Voronoi partition* (or sometimes *tessellation*) induced by Γ. When the norm $|\,.\,|$ is Euclidean, the closures $\overline{C}_i(\Gamma)$ of the *cells* are non-empty polyhedral closed convex sets (intersection of finitely many half-spaces defined by median hyperplanes of the couples of points (x_i,x_j), $i\neq j$). One easily shows that

$$\Big\{\xi\in\mathbb{R}^d\ :\ |\xi-x_i| < \min_{1\le j\le N,\, j\neq i}|\xi-x_j|\Big\}\subset \overset{\circ}{C}_i(\Gamma)\subset \overline{C}_i(\Gamma)\subset \Big\{\xi\in\mathbb{R}^d\ :\ |\xi-x_i| = \min_{1\le j\le N}|\xi-x_j|\Big\}.$$

The inclusions at both ends of the inclusion chain can be replaced by equalities *in an Euclidean framework.*

Then, for a given (static) random vector having values in $\mathbb{R}^d$, one defines a *Voronoi* Γ*-quantization* of X by Γ as

$$\widehat{X}^{\Gamma} = \pi_{\Gamma}(X).$$

Remark. For more developments on the non-Euclidean framework, like *e.g.* the ℓ^r-norms defined by $|\xi|_r = \big(|\xi^1|^r+\cdots+|\xi^d|^r\big)^{\frac{1}{r}}$, $r\in[1,+\infty)$, or $|\xi|_\infty = \max_{1\le j\le d}|\xi^j|$, $\xi=(\xi^1,\ldots,\xi^d)\in\mathbb{R}^d$, we refer to [33], Chapter 1.

This leads us to define for every $p\in(0,+\infty)$ the L^p-mean quantization error induced by a grid Γ as

$$\begin{aligned} e_p(\Gamma,X) &= \big\|X-\pi_\Gamma(X)\big\|_{L^p(\mathbb{P})} = \big\|\mathrm{dist}(X,\Gamma)\big\|_{L^p(\mathbb{P})} = \big\|\min_{1\le i\le N}|X-x_i|\big\|_{L^p(\mathbb{P})} &(1.1)\\ &= \big\|\min_{1\le i\le N}|\xi-x_i|\big\|_{L^p(\mu)} = \Big(\int_{\mathbb{R}^d}\min_{1\le i\le N}|\xi-x_i|^p\mu(d\xi)\Big)^{\frac{1}{p}}. &(1.2)\end{aligned}$$

Note that, from a computational point of view, the computation of $\pi_\Gamma(\xi)$ is very demanding when the size N is large since it amounts to a nearest neighbour search. We will come back to that point further on in Section 3 devoted to numerical aspects of (optimal) quantization grid computation.

Question 2 *How to choose* Γ *in order to improve the transmission?*

The underlying idea is to try selecting (or designing) a grid Γ with size at most N which optimally "fits" to the distribution μ of X, with in mind an approximation in the L^p-sense when $X\in L^p_{\mathbb{R}^d}(\mathbb{P})$. To this end, we introduce the *L^p-distortion function.*

Definition 1.1. *Let $p\in(0,+\infty)$ and $X\in L^p_{\mathbb{R}^d}(\mathbb{P})$. The $\mathbb{R}_+$-valued function $\mathcal{G}_{p,N}$ defined on $(\mathbb{R}^d)^N$ by*

$$\mathcal{G}_{p,N}\ :\ (x_1,\ldots,x_N)\longmapsto \mathbb{E}\Big(\min_{1\le i\le N}|X-x_i|^p\Big) = e_p(\Gamma,X)^p = \big\|\mathrm{dist}(X,\Gamma)\big\|^p_{L^p(\mathbb{P})}$$

is called the L^p-distortion function.

It is clear that, if we define the optimal L^p-mean quantization problem by

$$e_{p,N}(X) = \inf_{\Gamma,\,\mathrm{card}(\Gamma)\le N} e_p(\Gamma,X) \tag{1.3}$$

where $\mathrm{card}(\Gamma)$ denotes the cardinality of the grid $\Gamma\subset\mathbb{R}^d$, then

$$e_{p,N}(X) = \inf_{(x_1,\ldots,x_N)\in(\mathbb{R}^d)^N}\big(\mathcal{G}_{p,N}(x_1,\ldots,x_N)\big)^{\frac{1}{p}}.$$

Note that, in fact, $e_{p,N}(X)$ only depends on the distribution μ of X. So we will occasionally write $e_{p,N}(\mu)$ instead of $e_{p,N}(X)$. This follows from the easy remark that a grid Γ with less than N elements can always be represented by an N-tuple in which each element of the grid appears as a component at least once.

Proposition 1.1. *Let $p \in (0,+\infty)$. Assume that $X \in L^p_{\mathbb{R}^d}(\mathbb{P})$ i.e. $\int_{\mathbb{R}^d} |\xi|^p \mu(d\xi) < +\infty$ so that the distortion function $\mathcal{G}_{p,N}$ is finite everywhere on $(\mathbb{R}^d)^N$.*

(*a*) *The distortion function $\mathcal{G}_{p,N}$ attains a minimum at an N-tuple $x^{(N,p)} = (x_1^{(N,p)}, \ldots, x_N^{(N,p)})$.*

(*b*) *If $\mathrm{card}\big(\mathrm{supp}(\mu)\big) \geq N$, then the corresponding grid $\Gamma^{(N,p)} = \big\{x_1^{(N,p)}, \ldots, x_N^{(N,p)}\big\}$ has full size N and for every Voronoi partition $\big(C_i(\Gamma^{(N)})\big)_{1\leq i\leq N}$ of $\mathbb{R}^d$ induced by $\Gamma^{(N)}$, $\mathbb{P}(X \in C_i(\Gamma^{(N)}) > 0$.*

(*c*) *The sequence $N \mapsto e_{p,N}(X)$ (strictly) decreases as long as $N \leq \mathrm{cardsupp}(\mu)|$ and*

$$\lim_N e_{p,N}(X) = 0.$$

The proof of this proposition is postponed to Section 2.1. The grid $\Gamma^{(N,p)}$, the corresponding N-tuples $x^{(N,p)}$ (there are $N!$ N-tuples obtained by permutations of the components if the grid has full size N) as well as the (Borel) nearest neighbour projections $\pi_{\Gamma^{(N,p)}}$ are all called L^p-*optimal quantizers*.

Of course a crucial question in view of possible applications is to compute such L^p-*optimal quantizers* at level N, especially in higher dimension.

When $d = 1$ and $\mu = U([0,1])$, then, for any $p \in (0,+\infty)$, the *mid-point grid* $\Gamma^{(N,p)} = \big\{\frac{2i-1}{2N},\ i = 1,\ldots,N\big\}$ is the unique optimal L^p-quantizer at level N. The attached weights are all equal to $w_i^{(p,N)} = \frac{1}{N}$, $i = 1,\ldots,N$, and the resulting optimal L^p-quantization error is given for every $N \geq 1$ by

$$e_{p,N}\big(U([0,1])\big) = \frac{1}{2(1+p)^{1/p}N}. \tag{1.4}$$

More generally the question of the *rate of decay* of $e_{p,N}(X)$ is the central question of optimal vector quantization theory. It will be investigated further on in Section 2.3.

1.2. Application to signal transmission (source coding)

As mentioned in the introduction, this application of (optimal) quantization goes back to the very origin of quantization theory in the 1950's. Imagine one has access to an L^p-optimal quantization grid, say for $p = 2$ (quadratic case in an Euclidean setting). For convenience, we assume that $\Gamma = \{x_1,\ldots,x_N\}$ is a grid (possibly optimal) such that $\mathbb{P}\big(X \in \bigcup_{1\leq i\leq N} \partial C_i(\Gamma)\big) = \mu\big(\bigcup_{1\leq i\leq N} \partial C_i(\Gamma)\big) = 0$ *e.g.* because μ assigns no mass to hyperplanes.

What is the information "contained" in $\widehat{X}^\Gamma = \pi_\Gamma(X)$? Or equivalently, in probabilistic terms, what are the *characteristics* of the distribution of $\widehat{X}^\Gamma$?

(1) Its state space $\Gamma = \{x_1,\ldots,x_N\}$,

(2) Its "companion" weights $w_i = w_i^{(\Gamma)} = \mathbb{P}(\widehat{X}^\Gamma = x_i) = \mathbb{P}(X \in C_i(\Gamma)) = \mu(C_i(\Gamma))$, $i = 1,\ldots,N$.

If X is a random vector with a known simulatable distribution μ, one can pre-compute these weights w_i with an arbitrary accuracy by a large scale Monte Carlo simulation since, owing to the Strong Law of large Numbers,

$$w_i = \mathbb{P}\text{-}a.s. \lim_{M\to+\infty} \frac{\mathrm{card}\Big\{1 \leq m \leq M\ :\ |X_m - x_i| < \min_{j\neq i}|X_m - x_j\Big\}}{M},\ i = 1,\ldots,M,$$

where $(X_m)_{m\geq 1}$ is a sequence of i.i.d. random vectors with distribution μ. In case of a not too large dataset (a situation commonly met in signal transmission) one computes the weights w_i by assigning its Voronoi class to each data (when the dataset is too large a statistical approach is adopted).

1.2.1. *Coding the (quantized) signal*

Let $\Gamma = \{x_1 \ldots, x_{_N}\} \subset \mathbb{R}^d$ be a grid of size N, possibly sub-optimal at this stage, and let $\mathcal{P}(\Gamma)$ be the set of distributions whose support is exactly Γ. In order to transmit a Γ-valued signal from a sender A to a receiver B, A will transmit a *codeword* $C_i = C(x_i)$ *representative* of x_i instead of (an accurate enough approximation of) x_i itself. For simplicity we will assume that the coding function C maps Γ into the set $\{0,1\}^{(\mathbb{N})}$ of finite $\{0,1\}$-valued sequences. This means that we adopt a *dyadic* coding procedures. The set $\{0,1\}$ is called a 2-*alphabet* ([1]). Our first request on the function C is *identifiability i.e.* that B can always recover x_i from C_i or equivalently that C is injective. To design the codewords $(C_i)_{1\le i\le N}$, one aims at minimizing the *mean transmission cost* κ, also known as the *mean length* of the message. This is in fact a very old problem which goes back to the origins of Information Theory introduced by Claude Shannon in [71].

Let us focus for a while on this coding problem. The mean transmission cost $\kappa(N)$ for a grid of size N is clearly defined by

$$\kappa(N) = \sum_{i=1}^{N} w_i \times \text{length}(C_i).$$

A first (not so) naive idea is to re-index the points x_i by a permutation σ so that $i \mapsto w_{\sigma(i)}$ is non-increasing. Without loss of generality, we may assume from now on that σ is identity (though, for one-dimensional distributions, it is not consistent in general with the natural order of the points x_i on the real line). Then, it is intuitive (but in fact not mandatory) to devise the coding function C so that $i \mapsto \text{length}(C_i)$ is non-decreasing since, doing so, the more often a code is transmitted, the shorter it will be. In case of equality, like for the uniform distribution over Γ, assignment conventions have to be made.

The naive approach is to simply code x_i through the regular dyadic expression $\bar{i}^2$ of i which needs $1 + \lfloor \log_2 i \rfloor$ digits (where $\lfloor \xi \rfloor$ denotes the lower integer part of $\xi \in \mathbb{R}$). This yields

$$\kappa(N) = \sum_{i=1}^{N} w_i\big(1 + \lfloor \log_2 i \rfloor\big) = 1 + \sum_{i=1}^{N} w_i \lfloor \log_2 i \rfloor \le 1 + \lfloor \log_2 N \rfloor.$$

The transmission relies on the fact that both A and B share the *codebook i.e.* a one-to-one correspondence

$$x_i \longleftrightarrow \bar{i}^{2}. \tag{1.5}$$

A toy example. Imagine that, to transmit a uniformly distributed signal over the unit interval $[0,1]$, we first optimally quantize it using the mid-point grid $\Gamma^{(N)} = \Big\{\dfrac{2i-1}{2N},\ i = 1, \ldots, N\Big\}$. This is equivalent to transmit a uniformly distributed signal over $\{1, \ldots, N\}$ thanks to the codebook so that, as far as transmission is concerned, the grid $\Gamma^{(N)}$ itself plays no role. The resulting mean transmission cost $\kappa(N)$ is equal to

$$\kappa(N) = 1 + \frac{1}{N}\sum_{i=1}^{N} \lfloor \log_2 i \rfloor \sim \log_2\left(N/e\right) \quad \text{as} \quad N \to +\infty.$$

To be more precise, once noted that the *dyadic entropy* $H_2\big(\hat{\mu}_{_N}^{Unif}\big)$ of the uniform distribution $\hat{\mu}_{_N}^{Unif}$ over $\{1, \ldots, N\}$ (or equivalently on $\Gamma^{(N)}$) is equal to $\log_2 N$, we can show that

$$c_- = \liminf_N \Big(\kappa(N) - H_2\big(\hat{\mu}_{_N}^{Unif}\big)\Big) \le \limsup_N \Big(\kappa(N) - H_2\big(\hat{\mu}_{_N}^{Unif}\big)\big)\Big) = c_+$$

where $c_- \approx -2,8792$ and $c_+ \approx -0.9139$.

[1] More generally, if C is $\mathcal{X}^{(\mathbb{N})}$-valued, then $\mathcal{X}$ is called a card$(\mathcal{X})$-alphabet.

1.2.2. *Instantaneous coding.*

However, this approach is definitely too naive. In practice, A does not send *one* isolated codeword but a sequence of codewords. Such a coding is not satisfactory, mainly because it is not *self-punctuated*. To be decodable, an extra symbol (space, comma, etc) is needed to isolate the codewords. Doing so amounts to adding one symbol to the alphabet (with a special status since it cannot be repeated, like the large space in Morse coding). But this lowers the global performance of the coding system since it induces *de facto* switching from a 2-alphabet to a 3-alphabet coding function C, the third symbol having moreover a lower status of "under-symbol". To overcome this problem, the idea, again due to Shannon in his seminal 1948 paper [71], is to devise *self-punctuated codes*. This relies on two conditions. First we ask the coding process to be *uniquely decodable* in the sense that the *concatenation* of codewords $C(x_1)\cdots C(x_{_N})$ uniquely characterizes the concatenation $x_1\cdots x_{_N}$. The additional condition which defines an *instantaneous coding system* is that a codeword can never be the prefix of another or, equivalently, no codeword can be obtained as the concatenation of another codeword and further symbols of the alphabet (here 0 and 1 digits). One easily checks that an instantaneous coding procedure is always self-punctuated.

Unfortunately, it is also straightforward to check that the naive dyadic coding (1.5) formerly mentioned which consists in writing in base 2 every index i *is not* an instantaneous coding system since, *e.g.*, $\bar{2}^2 = 10$ and $\bar{5}^2 = 101$.

Let us illustrate on a simple example how an instantaneous coding procedure look. We consider the following coding procedure of the set of indices $\{1,2,3,4\}$:

$$C(1) = 0,\; C(2) = 10,\; C(3) = 110,\; C(4) = 111.$$

Such a code is uniquely decodable (*e.g.* 0110111100110 can be uniquely decoded as the string 134213). Furthermore it is clearly instantaneous (thus 010111110010 can be parsed on line as $0, 10, 111, 110, 10$ *i.e.* the string 12432).

If we consider the uniform distribution $\hat{\mu}_{_4}^{Unif}$ over $\{1,2,3,4\}$, the resulting mean transmission cost is equal to $\kappa\big(\hat{\mu}_{_4}^{Unif}\big) := \dfrac{1}{4}(1+2+3+3) = \dfrac{9}{4}$ whereas the naive dyadic coding of the indices seemingly yields $\frac{8}{4} = 2$. However, the *implementable* version of this naive dyadic coding (1.5), *i.e.* including an extra symbol like ",", has a mean length equal to $3 > \frac{9}{4}$. This can be up to 30% more symbol consuming than the above instantaneous code!

Now, let us consider a *general distribution* $\hat{\mu}_{_N}$ exactly supported by $\{1,\ldots,N\}$ (or equivalently by a grid $\Gamma_{_N}$ of size N) and *a priori not uniform*. Assume we have access to the distribution $\hat{\mu}$ itself *i.e.* to the weights $w_i = \hat{\mu}_{_N}\big(\{i\}\big)$. We define the dyadic entropy $H_2(\hat{\mu})$ of $\hat{\mu}$ by

$$H_2(\hat{\mu}) = -\sum_{i=1}^{N} w_i \log_2 w_i.$$

Then, the following classical theorem from Information Theory holds (see [20], Chapter 5, Theorem 5.3.1 and Section 5.4).

Theorem 1.1. *For any instantaneous dyadic coding procedure $C : \{1,\ldots,N\} \to \{0,1\}^{(\mathbb{N})}$ of the distribution $\hat{\mu}$, its mean transmission cost $\kappa_{\hat{\mu}}(N)$ satisfies*

$$\kappa(\hat{\mu}_{_N}) \geq H_2(\hat{\mu}_{_N}). \tag{1.6}$$

Furthermore, there exists (at least) one instantaneous coding procedure such that

$$\kappa_{\hat{\mu}}(N) < H_2(\hat{\mu}_{_N}) + 1. \tag{1.7}$$

For a proof of this result based on *Kraft's inequality*, which is too far from the scope of this paper, we refer to [20]. Furthermore, when a sequence $(Y_n)_{n\geq 0}$ of $\{1,\dots,N\}$-valued signals to be transmitted is stationary with marginal invariant distribution $\hat{\mu}_{_N}$ and ergodic, it is possible by aggregating n of them to show (with obvious notations, see again [20]) that

$$\kappa(Y_1,\dots,Y_n) \to H_2(\hat{\mu})\ a.s. \quad \text{ as } n \to +\infty. \tag{1.8}$$

Examples. (a) *The Huffman code:* It was the first optimal instantaneous code – devised in Huffman's PhD thesis (see also [38]). Its length sequence $(\ell_i^*)_{1\leq i\leq N}$ can be obtained as the solution to the integer optimization problem (ℓ_i denotes the length of a code C_i):

$$\ell^* = \underset{\sum 2^{-\ell_i}\leq 1}{\operatorname{argmin}} \sum w_i \ell_i$$

so that $H_2(\hat{\mu}_{_N}) \leq \kappa_{Huf}(\hat{\mu}_{_N}) = \sum w_i \ell_i^* \leq H_2(\hat{\mu}_{_N}) + 1$. For an explicit construction of the Huffman code – and not only of its length sequence!) – we refer again to [20], Chapter 5. Let us simply mention that the codes are obtained by the concatenation of labels given to the edges, say 1 for "right" edges, 0 for "left edges" starting from the root, of successive trees built from the increasing monotony of the weights w_i. The successive trees are obtained by summing up the lower probabilities, starting from $\widetilde{w}_{_{N-1}} := w_N + w_{N-1}$, with appropriate conventions in case of equality like with uniform distributions.

(b) *The Shannon coding* (see exercise 5.28 in [20]): Still assume that the weights of the distribution $\hat{\mu}_{_N}$ satisfy $0 < w_{_N} \leq \dots \leq w_1 < 1$. Let $F^{\hat{\mu}_N}$ denote the *strict*-cumulative distribution function of $\hat{\mu}_{_N}$ defined by

$$F_i^{\hat{\mu}_N} = \sum_{j<i} w_j.$$

Set

$$\ell_i = \lceil -\log_2 w_i \rceil \quad \text{ and } \quad C_i = \lfloor 2^{\ell_i} F_i^{\hat{\mu}} \rfloor, \quad i = 1,\dots,N,$$

where $\lceil \xi \rceil$ denotes the upper integer part of the real number ξ. Elementary computations show that Shannon's code is instantaneous and that its mean transmission cost $\kappa_{Shan}S(\hat{\mu}_{_N})$ also satisfies

$$H_2(\hat{\mu}_{_N}) \leq \kappa_{Shan}S(\hat{\mu}_{_N}) < H_2(\hat{\mu}_{_N}) + 1.$$

1.2.3. *Global error induced by the transmission of a quantized signal*

Let us bring back quantization into the game by considering a continuous signal which needs to be quantized in order to reduce its transmission cost. Let us briefly compare from a quantitative viewpoint two modes of transmission for a signal.

▷ *Direct transmission.* Let $(X_t)_{t\geq 0}$ be a stochastic stationary signal with marginal distribution μ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $\Gamma = \{x_1,\dots,x_{_N}\}$. To transmit the Γ-quantization $\widehat{X}^{\Gamma}$ of the random signal $X = X_{t_0}$ at time t_0, the resulting quadratic mean quantization error is equal to

$$\left\| X - \widehat{X}^{\Gamma} \right\|_{L^2(\mathbb{P})} + 2^{-r} = e_2(\Gamma, \mu) + 2^{-r}$$

where 2^{-r} is the dyadic transmission accuracy of any of the elementary quantizers x_i. In fact this corresponds to a fixed transmission cost $\kappa = r + 1$ *i.e.* the *number of dyadic digits* used to transmit these values. Common values for r lie between 10 and 20 (having in mind that $2^{-10} = \frac{1}{1024} \approx 10^{-3}$).

$\triangleright$ *Signal transmission using the codebook.* If the receiver B uses the codebook $(C_i \longleftrightarrow x_i)_{1\le i\le N}$ for the decoding phase ([2]), the resulting mean quadratic transmission error will be equal to

$$\big\|X - \widehat{X}^{\Gamma}\big\|_{L^2(\mathbb{P})} = e_2(\Gamma, \mu)$$

whereas the mean unitary transmission cost is $\kappa_{\hat{\mu}}(N)$ where $\hat{\mu}$ is the distribution of the quantized signal $\widehat{X}^{\Gamma}$. In this second case, there is a connection between the transmission error and the transmission cost that will be made more precise in Section 2.3 when the grid Γ is L^2-optimal at level N for μ.

However, in the very simple case of the uniform distribution $U([0,1])$ over the unit interval, we can establish a direct relation between quadratic mean transmission error and mean transmission cost κ *when both the quantization and the instantaneous coding are optimal.* The optimal quadratic quantization of $U([0,1])$ is the uniform distribution $\hat{\mu}_N^{Unif}$ over the N-mid-point whose dyadic entropy is exactly $H_2(\hat{\mu}_N^{Unif}) = \log_2 N$. Plugging this equality in (1.7) yields $\kappa_{\hat{\mu}_N} \le \log_2(N)$. In turn, plugging this inequality in the quantization error bound (1.4) yields that *the lowest achievable mean transmission error*, for a prescribed mean transmission cost κ, approximately satisfies

$$\frac{2^{-(\kappa+1)}}{\sqrt{3}} \le L^2\text{-Mean transmission error}(\kappa) \le \frac{2^{-\kappa}}{\sqrt{3}}.$$

A less sharp (reverse) formulation is

$$-\log_2\big(\text{Transmission error}(\kappa)\big) \sim \kappa \quad \text{as} \quad \kappa \to +\infty.$$

This result appears as the most elementary version of Shannon's source coding theorem, here in one dimension. Its extension to more general distributions μ on $\mathbb{R}^d$ will be possible, once stated the sharp convergence rate of the L^2-optimal mean quantization error for general distributions on $\mathbb{R}^d$ in Section 2.3 (Zador's Theorem).

We focused in the above lines on a static random signal presentation but the adaptation to a stationary process or a quasi-periodic signal, as defined above in terms of weak convergence of its time empirical measure, is straightforward. In particular for stationary ergodic signal one may take advantage of the improvement provided by (1.8), using n-aggregates of the signal, to reduce the range of the two-sided inequality (1.6)-(1.7) in Theorem 1.1.

1.3. What else is quantization for?

1.3.1. *Data mining, clustering, automatic classification*

Let $(\xi_k)_{1\le k\le n}$ be an $\mathbb{R}^d$-valued dataset and let μ be the uniform distribution over this dataset – the empirical measure of the dataset – defined by

$$\mu = \frac{1}{n}\sum_{k=1}^{n} \delta_{\xi_k} \tag{1.9}$$

where δ_a denotes the Dirac mass at $a \in \mathbb{R}^d$. In such a framework, n is usually large, say 10^6 or more, and optimal quantization can be viewed as a model for clustering *i.e.* the design of a set of N *prototypes* of the dataset, with $N \ll n$, obtained as a solution to the mean quadratic (or more generally L^p-) optimal quantization at level $N \ge 1$ of the distribution μ ($p \in (0,+\infty)$ being fixed). This reads as the L^p-minimization problem

$$\min_{(x_1,\ldots,x_N)\in(\mathbb{R}^d)^N} \frac{1}{n}\sum_{k=1}^{n} \min_{1\le i\le N} |\xi_k - x_i|^p.$$

[2] The sender A only needs a codebook to discriminate the elementary quantizers x_i *i.e.* a codebook where all x_i are known with a fixed length ℓ_1 (dyadic) bits in its dyadic representation. The receiver B may need arbitrary accurate values for the elementary quantizers x_i in his/her codebook in order to reconstruct *a posteriori* the signal.

The existence of such an optimal N-quantization grid $\Gamma^{(N,p)}$ of prototypes follows from the above Proposition 1.1. Such a distribution does assign mass to hyperplanes and in particular to the boundaries of polyhedral Voronoi cells. However, owing to Theorems 4.1 and 4.2 in [33] (p.38), we know that the boundaries of the Voronoi cells induced by an optimal grid $\Gamma^{(N,p)}$ are always μ-negligible.

Once an optimized grid of N prototypes has been computed (see Section 3 devoted to the algorithmic aspects), it can be used to produce an automatic classification of the dataset by making up "clusters" of points of the dataset following the nearest neighbour rule among the prototypes. Formulated equivalently, one defines the N clusters as the "trace" of the dataset on the N Voronoi cells $C_i(\Gamma^{(N,p)})$, $i = 1, \dots, N$.

From a mathematical point of view, investigations on this topic are carried out be replacing the deterministic dataset $(\xi_k)_{1\le k\le n}$ by a sequence of i.i.d. random vectors $(X_k)_{k\ge 0}$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with distribution μ. The quantities of interest become, in short, the sequence of optimization problems induced by the random empirical measures $\mu_n(\omega, d\xi) = \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k(\omega)}(d\xi)$, $\omega\in\Omega$. This has given rise to a huge literature in Statistics and has known a kind of renewal with the emergence of clustering methods in the "Big Data" world, see [10]. We consider, for every $\omega\in\Omega$, the optimization problem

$$\min_{(\mathbb{R}^d)^N}\left[\frac{1}{n}\sum_{k=1}^{n}\min_{1\le i\le N}|X_k(\omega)-x_i|^p = \int_{\mathbb{R}^d}\min_{1\le i\le N}|\xi - x_i|^p\mu_n(\omega, d\xi)\right]. \tag{1.10}$$

The main connection with optimal quantization is the following. assume that $\mu(B(0;1)) = 1$. For every $\omega\in\Omega$, there exists (at least) an optimal N-tuple $x^{(N)}(\omega, n)$ for the above problem which satisfies

$$\mathbb{E}\Big(e_2\big(x^{(N)}(\omega,n),\mu\big)\Big) - e_{2,N}(\mu) \le C\min\left(\sqrt{\frac{Nd}{n}}, \sqrt{\frac{d\,N^{1-\frac{2}{d}}\log n}{n}}\right)$$

where $C > 0$ is a positive universal real constant. For other results, we also refer to [34] devoted to the quantization rate of empirical measures.

1.3.2. *From Numerical integration (I) ...*

Another way to take advantage of optimal quantization emerged in the 1990's (see [55]). As we know, for a sequence $(\Gamma^{(N,p)})_{N\ge1}$ of L^p-optimal grids of size N with $N\to+\infty$, we have

$$\|X - \widehat{X}^{\Gamma^{(N,p)}}\|_{L^p(\mathbb{P})} = e_{p,N}(X) \to 0$$

i.e. $\widehat{X}^{\Gamma^{(N,p)}} \to X$ in L^p as $N\to+\infty$ (hence in distribution). It can be shown (see [22]) that, in fact, this convergence also holds in an *a.s.* sense although we will make little use of this feature in what follows. In particular, if a function $F:\mathbb{R}^d\to\mathbb{R}$ is bounded and continuous, then $\mathbb{E}\,F(\widehat{X}^{\Gamma^{(N,p)}})\to\mathbb{E}\,F(X)$ as $N\to+\infty$. On the other hand, using the characteristics $(x_i^{(N)}, w_i^{(N)})_{1\le i\le N}$ of the distribution of $\widehat{X}^{\Gamma^{(N,p)}}$, we derive a very simple weighted *cubature formula*

$$\mathbb{E}\,F\big(\widehat{X}^{\Gamma^{(N,p)}}\big) = \sum_{i=1}^{N} w_i^{(N)}F\big(x_i^{(N)}\big). \tag{1.11}$$

When F has more regularity and is possibly not bounded, precise error bounds for this quantization based cubature formula can also be established, as we will see now.

First order error bound for the quantization based cubature formula. Assume F is locally α-Hölder continuous in the sense that there exists $\alpha\in(0,1]$, $\beta\ge0$, and a real constant $[F]_{\alpha,\beta}$ such that

$$\forall\, x,\, y\in\mathbb{R}^d,\quad |F(x)-F(y)| \le \quad [F]_{\alpha,\beta}|x-y|^{\alpha}\big(1+|x|^{\beta}+|y|^{\beta}\big).$$

Let $\Gamma \subset \mathbb{R}^d$ be a quantization grid. Then, for every conjugate Hölder exponents $(p,q) \in [1,+\infty]$,

$$\begin{aligned}\big|\mathbb{E}\,F(X) - \mathbb{E}\,F(\widehat{X}^\Gamma)\big| &\le [F]_{\alpha,\beta}\mathbb{E}\left(|X-\widehat{X}^\Gamma|^\alpha\big(1+|X|^\beta+|\widehat{X}^\Gamma|^\beta)\big)\right)\\ &\le [F]_{\alpha,\beta}\|X-\widehat{X}^\Gamma\|^\alpha_{L^{\alpha p}(\mathbb{P})}\left(1+\|X\|^\beta_{L^{\beta q}(\mathbb{P})}+\|\widehat{X}^\Gamma\|^\beta_{L^{\beta q}(\mathbb{P})}\right).\end{aligned}$$

In particular, if $p = \frac{1}{\alpha}$, one gets

$$\big|\mathbb{E}\,F(X) - \mathbb{E}\,F(\widehat{X}^\Gamma)\big| \le [F]_{\alpha,\beta}\|X-\widehat{X}^\Gamma\|^\alpha_1\left(1+\|X\|^\beta_{L^{\frac{\beta}{1-\alpha}}(\mathbb{P})}+\|\widehat{X}^\Gamma\|^\beta_{L^{\frac{\beta}{1-\alpha}}(\mathbb{P})}\right) \tag{1.12}$$

with the convention $\|\,.\,\|^0_{L^{\frac{\beta}{1-1}}(\mathbb{P})} = 1$. If F is α-Hölder continuous with Lipschitz coefficient $[F]_\alpha = \frac{1}{3}[F]_{1,0}$, then

$$\big|\mathbb{E}\,F(X) - \mathbb{E}\,F(\widehat{X}^\Gamma)\big| \le F]_\alpha\|X-\widehat{X}^\Gamma\|^\alpha_{L^\alpha(\mathbb{P})}. \tag{1.13}$$

From the cubature formula (1.13) and using that bounded Hölder functions characterize the weak convergence of probability measures, we derive the following corollary about L^p-optimal quantizers (by considering $\alpha = p\wedge 1$).

Corollary 1.1. *Let $X \in L^p_{\mathbb{R}^d}(\mathbb{P})$, $p \in (0,+\infty)$, with distribution μ. Let $(\Gamma^{(N)})_{N\ge1}$ be a sequence of quantizers, with $\Gamma^{(N)}$ of size N, satisfying $e_p(\Gamma^{(N)},\mu)\to 0$ as $N\to+\infty$. Let $\hat{\mu}_N$ denote the distribution of the quantization $\widehat{X}^{\Gamma^{(N)}}$. Then*

$$\hat{\mu}_N = \sum_{i=1}^N \mu\big(C_i(\Gamma^{(N)})\big)\delta_{x_i^{(N)}} \xrightarrow{(w)} \mu \quad \textit{as} \quad N\to+\infty \tag{1.14}$$

where $\Gamma^{(N)} = \{x_1^{(N)},\ldots,x_N^{(N)}\}$ and $\xrightarrow{(w)}$ denotes the weak convergence of distributions.

1.3.3. *. . . to Numerical Probability (conditional expectation)*

One of the main problem investigated in the past twenty years in Numerical Probability has been the *numerical computation of conditional expectations*, mostly motivated by problems arising in finance for the pricing of derivative products of American style or more generally known as "callable". It is also a challenging problem for the implementation of numerical schemes for Backward Stochastic Differential Equations (see [2,3]), Stochastic PDEs (see [32]), for non-linear filtering [57, 68] or Stochastic Control Problems (see [13, 14, 58]). Further references are available in the survey paper [62] devoted to applications of optimal vector quantization to Numerical Probability. The specificity of these problems in the probabilistic world is that, whatever the selected method is, it suffer in some way or another from the *curse of dimensionality*. Optimal quantization trees (introduced in [2]) is one of the numerical methods designed to cope with this problem (with regression and Monte Carlo-Malliavin method, see [46], [28]). The precise connection between vector quantization and conditional expectation computation can be summed up in the proposition below.

We consider a couple of random vectors $(X,Y) : (\Omega,\mathcal{A},\mathbb{P}) \to \mathbb{R}^d\times\mathbb{R}^q$ and the regular version Q of the conditional distribution operator of X given Y, defined on every bounded or non-negative Borel function $f : \mathbb{R}^d \to \mathbb{R}$, by

$$Qf(y) = \mathbb{E}\big(f(X)\,|\,Y=y\big).$$

Then, Qf is a Borel function on $\mathbb{R}^d$. We define the Lipschitz ratio of a function $f : \mathbb{R}^d \to \mathbb{R}$ by $[f]_{\rm Lip} = \sup_{x\neq y}\frac{|f(y)-f(x)|}{|x-y|} \le +\infty$. We make the following *Lipschitz continuity propagation* assumption on Q: there exists $[Q]_{\rm Lip} \in \mathbb{R}_+$ such that

$$\forall\, f : \mathbb{R}^d \to \mathbb{R},\ \text{Borel function},\quad [Qf]_{\rm Lip} \le [Q]_{\rm Lip}[f]_{\rm Lip}. \tag{1.15}$$

Proposition 1.2. *Assume that the conditional distribution operator Q of X given Y satisfies the above Lipschitz continuity propagation property (1.15). Let $\Gamma_X \subset \mathbb{R}^d$ and $\Gamma_Y \subset \mathbb{R}^q$ be two quantization grids of X and Y respectively.*

(a) Quadratic case. *Assume $X, Y \in L^2(\mathbb{P})$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz continuous function and let $g : \mathbb{R}^d \to \mathbb{R}$ be a Borel function with linear growth. Then*

$$\big\|\mathbb{E}\big(f(X)\,|\,Y\big) - \mathbb{E}\big(g(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big)\big\|^2_{L^2(\mathbb{P})} \le [Qf]^2_{\rm Lip}\big\|Y-\widehat{Y}^{\Gamma_Y}\big\|^2_{L^2_{\mathbb{R}^q}(\mathbb{P})} + \big\|f(X)-g(\widehat{X}^{\Gamma_X})\big\|^2_{L^2(\mathbb{P})}$$

so that if $g = f$,

$$\big\|\mathbb{E}\big(f(X)\,|\,Y\big) - \mathbb{E}\big(f(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big)\big\|^2_{L^2_{\mathbb{R}^q}(\mathbb{P})} \le [Qf]^2_{\rm Lip}\big\|Y-\widehat{Y}^{\Gamma_Y}\big\|^2_{L^2(\mathbb{P})} + [f]^2_{\rm Lip}\big\|X-\widehat{X}^{\Gamma_X}\big\|^2_{L^2(\mathbb{P})}.$$

(b) L^p-case. *Assume $X, Y \in L^p(\mathbb{P})$, $p \in [1, +\infty)$ and let f and g be like in (a). Then*

$$\big\|\mathbb{E}\big(f(X)\,|\,Y\big) - \mathbb{E}\big(g(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big)\big\|_{L^p_{\mathbb{R}^q}(\mathbb{P})} \le (2-\delta_{p,2})[Qf]_{\rm Lip}\big\|Y-\widehat{Y}^{\Gamma_Y}\big\|_{L^p(\mathbb{P})} + \big\|f(X)-g(\widehat{X}^{\Gamma_X})\big\|_{L^p(\mathbb{P})}$$

where $\delta_{p,p'}$ denotes the Kronecker symbol. In particular, if $g = f$, one has

$$\big\|\mathbb{E}\big(f(X)\,|\,Y\big) - \mathbb{E}\big(f(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big)\big\|_{L^p(\mathbb{P})} \le (2-\delta_{p,2})[Qf]_{\rm Lip}\big\|Y-\widehat{Y}^{\Gamma_Y}\big\|_{L^p_{\mathbb{R}^q}(\mathbb{P})} + [f]_{\rm Lip}\big\|X-\widehat{X}^{\Gamma_X}\big\|_{L^p(\mathbb{P})}.$$

Proof. (a) We decompose $\mathbb{E}\big(f(X)\,|\,Y\big) - \mathbb{E}\big(f(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big)$ into two ($L^2(\mathbb{P})$-orthogonal) terms

$$\begin{aligned}\mathbb{E}\big(f(X)\,|\,Y\big) - \mathbb{E}\big(f(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big) \;=\;& \Big(\underbrace{\mathbb{E}\big(f(X)\,|\,Y\big) - \mathbb{E}\big(\mathbb{E}(f(X)\,|\,Y)\,|\,\widehat{Y}^{\Gamma_Y}\big)}_{(1)}\Big)\\ &+\Big(\underbrace{\mathbb{E}\big(\mathbb{E}(f(X)\,|\,Y)\,|\,\widehat{Y}^{\Gamma_Y}\big) - \mathbb{E}\big(g(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big)}_{(2)}\Big).\end{aligned}$$

To check the announced $L^2(\mathbb{P})$-orthogonality, we note that (2) is $\sigma(\widehat{Y}^{\Gamma_Y})$-measurable; hence, the characterization of conditional expectation given $\widehat{Y}^{\Gamma_Y}$ implies $\mathbb{E}\,(1)\times(2) = 0$. On the other hand, the very definition of conditional expectation given $\widehat{Y}^{\Gamma_Y}$ as the best approximation in $L^2_{\mathbb{R}^q}(\mathbb{P})$ by a square integrable $\sigma(\widehat{Y}^{\Gamma_Y})$-measurable random vector implies in turn

$$\begin{aligned}\mathbb{E}\,(1)^2 \;&=\; \mathbb{E}\big(Qf(Y) - \mathbb{E}(Qf(Y)\,|\,\widehat{Y}^{\Gamma_Y})\big)^2 \le \mathbb{E}\big(Qf(Y) - Qf(\widehat{Y}^{\Gamma_Y})\big)^2\\ &\le\; [Qf]^2_{\rm Lip}\big\|Y-\widehat{Y}^{\Gamma_Y}\big\|^2_{L^2(\mathbb{P})}.\end{aligned}$$

On the other hand, using that $\widehat{Y}^{\Gamma_Y}$ is $\sigma(Y)$-measurable, we first derive from the chain rule for conditional expectation that

$$(2) = \mathbb{E}\big(f(X)\,|\,\widehat{Y}^{\Gamma_Y}\big) - \mathbb{E}\big(g(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big) = \mathbb{E}\big(f(X) - g(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big).$$

Using now that conditional expectation is an L^2-contraction, we deduce that

$$\mathbb{E}\,(2)^2 \le \|f(X) - g(\widehat{X}^{\Gamma_X})\|^2_{L^2(\mathbb{P})} \le \|f(X) - g(\widehat{X}^{\Gamma_X})\|^2_{L^2_{\mathbb{R}^q}(\mathbb{P})}.$$

When $g = f$, the conclusion is straightforward.

(*b*) We start from the classical Minkowski Inequality

$$\begin{aligned}\big\|\mathbb{E}\big(f(X)\,|\,Y\big)-\mathbb{E}\big(g(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big)\big\|_{L^p(\mathbb{P})} &\le \big\|Qf(Y)-\mathbb{E}(Qf(Y)\,|\,\widehat{Y}^{\Gamma_Y})\big\|_{L^p(\mathbb{P})}\\ &\quad+\big\|\mathbb{E}\big(f(X)\,|\,\widehat{Y}^{\Gamma_Y}\big)-\mathbb{E}\big(g(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big)\big\|_{L^p(\mathbb{P})}\end{aligned}$$

where we used like in (*a*) that $\mathbb{E}\big(Qf(Y)\,|\,\widehat{Y}^{\Gamma_Y}\big)=\mathbb{E}\big(f(Y)\,|\,\widehat{Y}^{\Gamma_Y}\big)$. Now, still owing to Minkowski's Inequality,

$$\big\|Qf(Y)-\mathbb{E}(Qf(Y)\,|\,\widehat{Y}^{\Gamma_Y})\big\|_{L^p(\mathbb{P})}\le\big\|Qf(Y)-Qf(\widehat{Y}^{\Gamma_Y})\big\|_{L^p(\mathbb{P})}+\big\|\mathbb{E}\big(Qf(\widehat{Y}^{\Gamma_Y})-Qf(Y)\,|\,\widehat{Y}^{\Gamma_Y}\big)\big\|_{L^p(\mathbb{P})}$$

so that

$$\begin{aligned}\big\|Qf(Y)-\mathbb{E}(Qf(Y)\,|\,\widehat{Y}^{\Gamma_Y})\big\|_{L^p(\mathbb{P})} &\le 2\big\|Qf(Y)-Qf(\widehat{Y}^{\Gamma_Y})\big\|_{L^p(\mathbb{P})}\\ &\le 2[Qf]_{\rm Lip}\big\|Y-\widehat{Y}^{\Gamma_Y}\big\|_{L^p_{\mathbb{R}^q}(\mathbb{P})}.\end{aligned}$$

Note that when $p=2$ the above coefficient 2 can be cancelled using again, like in (*a*), that conditional expectation given $\widehat{Y}^{\Gamma_Y}$ is the best *approximator* in $L^2(\mathbb{P})$ by $\sigma(\widehat{Y}^{\Gamma_Y})$-measurable square integrable random vectors. On the other hand,

$$\big\|\mathbb{E}\big(f(X)\,|\,\widehat{Y}^{\Gamma_Y}\big)-\mathbb{E}\big(g(\widehat{X}^{\Gamma_X})\,|\,\widehat{Y}^{\Gamma_Y}\big)\big\|_{L^p(\mathbb{P})}\le\big\|f(X)-g(\widehat{X}^{\Gamma_X})\,\big\|_{L^p(\mathbb{P})}.$$

The case $g=f$ follows immediately. This completes the proof. □

To conclude this section, we make the connection between these cubature formulas and the L^1-Wasserstein distance $\mathcal{W}_1$ defined by

$$\mathcal{W}_1(\mu,\nu)=\inf\Big\{\mathbb{E}_{\mathbb{P}}|X-Y|,\;X,\,Y:(\Omega,\mathcal{A},\mathbb{P})\to\mathbb{R}^d,\;X\overset{d}{=}\mu,\;Y\overset{d}{=}\;\nu\Big\}$$

where $\overset{d}{=}$ denotes the identity in distribution.

Proposition 1.3. *Let* $X\in L^p_{\mathbb{R}^d}(\mathbb{P})$, $p\in(0,1]$, *with distribution* μ *and let* $\Gamma=\{x_1,\ldots,x_N\}$.
(*a*) *For every* $p\in(0,1]$, $\|X-\widehat{X}^\Gamma\|^p_{L^p(\mathbb{P})}=\sup_{[F]_p\le 1}\big|\mathbb{E}F(X)-\mathbb{E}F\big(\widehat{X}^\Gamma\big)\big|$ *where* $[F]_p=\sup_{x\neq y}\frac{|F(x)-F(y)|}{|x-y|^p}$ *denotes the* p*-Hölder coefficient of the function* $F:\mathbb{R}^d\to\mathbb{R}$.
(*b*) *If* $\mathcal{P}_N$ *denotes the set of probability measures with a support having at most* N *points in* $\mathbb{R}^d$, *then*

$$\mathcal{W}_1(\mu,\mathcal{P}_N)=e_{1,N}(\mu).$$

Proof. (*a*) The inequality $\sup_{[F]_p\le 1}\big|\mathbb{E}F(X)-\mathbb{E}F\big(\widehat{X}^\Gamma\big)\big|\le\|X-\widehat{X}^\Gamma\|^p_{L^p(\mathbb{P})}$ is straightforward by setting $\alpha=p$ and $\beta=0$ in (1.12) and noting that $[F]_p=\frac{1}{3}[F]_{p,0}$. The equality follows by noting that the function F_p defined for every $\xi\in\mathbb{R}^d$ by $F_p(\xi)=\min_{1\le i\le N}|\xi-x_i|^p$ is p-Hölder with $[F_p]_p=1$.
(*b*) Let $X:(\Omega,\mathcal{A},\mathbb{P})\to\mathbb{R}^d$ with distribution $\mathbb{P}_X=\mu$. It is clear, as already seen, that if $Y:(\Omega,\mathcal{A},\mathbb{P})\to\mathbb{R}^d$ is such that $\Gamma_Y=Y(\Omega)$ has at most N values, then $|X-Y|\ge\text{dist}(X,\Gamma_Y)=|X-\widehat{X}^{\Gamma_Y}|$ so that $\|X-\widehat{X}^{\Gamma_Y}\|_1\le\mathbb{E}|X-Y|$. As a consequence $e_{1,N}(\mu)\le\mathcal{W}_1(\mu,\mathcal{P}_N)$. Conversely, it follows from the definition of $e_{1,N}(\mu)$ in (1.3) that $e_{1,N}(\mu)\ge\mathcal{W}_1(\mu,\mathcal{P}_N)$ since it is defined as an infimum over *less* random vectors, namely those of the form $Y=q(X)$ of X where $q:\mathbb{R}^d\to\mathbb{R}$ takes at most N values, than the L^1-Wasserstein distance. This completes the proof. □

1.4. **Application to Numerical Analysis**

1.4.1. *Representation and numerical approximation of the solution of parabolic PDE, Feynman-Kac's formula*

Let $b : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ and $a : [0,T] \times \mathbb{R}^d \to \mathbf{S}^+(d,\mathbb{R})$ be two continuous functions with at most linear and quadratic growth in x, uniformly with respect to $t \in [0,T]$, respectively ($\mathbf{S}^+(d,\mathbb{R})$ denotes the set of $d \times d$ symmetric non-negative matrices). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Borel function with polynomial growth. We want to solve numerically the following parabolic partial differential equation (PDE), either by a Monte Carlo simulation or by a quadrature formula

$$\frac{\partial u}{\partial t} + Lu = 0, \quad u(T,.) = f \tag{1.16}$$

where, denoting by $(.|.)$ the canonical inner product on $\mathbb{R}^d$,

$$Lu = (b|\nabla u) + \frac{1}{2}\mathrm{Tr}(a\nabla^2 u). \tag{1.17}$$

▷ Step 1 (*Feynman-Kac's representation formula*). This fundamental connection between diffusion process and (parabolic) PDEs is summed up in the following theorem.

Theorem 1.2 (Feynman-Kac's representation formula)**.** *Assume (for simplicity) that the functions b and a are such that the above PDE (1.17) has a unique $\mathcal{C}^{1,2}([0,T] \times \mathbb{R}^d)$ solution u whose gradient $\nabla_x u$ has polynomial growth in x, uniformly in $t \in [0,T]$. Let $\sigma : \mathbb{R}^d \to \mathcal{M}(d,q,\mathbb{R})$ ([3]) such that $a = \sigma\sigma^*$ (where * stands for matrix transposition). Assume that b and σ are continuous on $[0,T] \times \mathbb{R}^d$ and, at least, Lipschitz continuous in x, uniformly in $t \in [0,T]$.*

(a) *Then the function u admits the following representation as an expectation:*

$$\forall\, x \in \mathbb{R}^d,\ \forall\, t \in [0,T], \quad u(t,x) = \mathbb{E}\, f(X^{t,x}_{T})$$

where $(X^{x,t}_s)_{s\in[t,T]}$ denotes the unique solution to the Stochastic Differential Equation (SDE)

$$dX^{t,x}_s = b(s,X^{t,x}_s)ds + \sigma(s,X^{t,x}_s)dW_s,\ X^{t,x}_t = x,\ s \in [t,T], \tag{1.18}$$

starting from $x \in \mathbb{R}^d$ at time $t \in [0,T]$ and defined on $[t,T]$, where W is a q-dimensional standard Brownian motion defined on a probability space $(\Omega,\mathcal{A},\mathbb{P})$.

Owing to the Markov property, an alternative formulation is given by

$$\forall\, t \in [0,T], \quad \mathbb{E}\big(f(X_{T})\,|\,X_t\big) = u(t,X_t) \quad a.s.$$

for any solution $(X_t)_{t\in[0,T]}$ of the above SDE defined over the whole interval $[0,T]$ starting at a finite random vector X_0 independent of W. In particular $u(t,x) = \mathbb{E}\big(f(X_{T})\,|\,X_t = x\big)$ (in the sense that it is a regular version of the conditional expectation as x varies).

(b) Time homogeneous diffusion coefficients*: If $b(t,x) = b(x)$ and $\sigma(t,x) = \sigma(x)$ (no dependence of b and σ in t), then the representation can be written*

$$\forall\, x \in \mathbb{R}^d,\ \forall\, t \in [0,T], \quad u(t,x) = \mathbb{E}\, f(X^{0,x}_{T-t}). \tag{1.19}$$

Proof. (a) Itô's formula applied to the function u and the process $(s,X^{t,x}_s)_{s\in[t,T]}$ between t and T yields

$$u(T,X^{t,x}_{T}) = u(t,x) + \int_t^T \underbrace{\Big(\frac{\partial u}{\partial t} + Lu\Big)(s,X^{t,x}_s)}_{=0}\,ds + \int_t^T \big(\nabla_x u(s,X^{t,x}_s)|\sigma(s,X^{t,x}_s)dW_s\big).$$

[3] Space of matrices with d rows and q columns with real entries.

The integral in "ds" is zero since u satisfies the parabolic PDE (1.16) and one easily establishes that the local martingale null at 0 defined by the Brownian stochastic integral is a true martingale, null at 0, owing to the growth control assumption made on $\nabla_x u$. Then, one gets

$$\mathbb{E}\, u(T, X_T^{t,x}) = u(t,x).$$

(b) One writes Itô's formula between 0 and $T-t$ to $u(T-t, X_t^{0,x})$ and proceeds as above. □

Remark. In the time homogeneous case, one can proceed by verification. Under smoothness assumption on b and σ, say $\mathcal{C}^2$ with bounded existing derivatives and Hölder second order partial derivatives, one shows, using the tangent process of the diffusion, that the function $u(t,x)$ defined by (1.19) is $\mathcal{C}^{1,2}$ in (t,x). Then, the above claim (b) shows the existence of a solution to the parabolic PDE (1.16).

▷ Step 2a *(Monte Carlo simulation)*. Assume for the sake of simplicity that we want to compute a numerical approximation of $u(0,x) = \mathbb{E}\, f(X_T^{0,x})$ *i.e.* that $t=0$. At this stage, the idea is to replace the diffusion by its *Euler scheme* with step $\frac{T}{n}$, $n \geq 1$, starting at x: let $t_k^n = \frac{kT}{n}$, $k = 0, \ldots, n$ be the uniform mesh of $[0,T]$ with step $\frac{T}{n}$. It is recursively defined as follows (to alleviate notations, we drop the dependance in ^{0,x} of the Euler scheme):

$$\bar{X}_{t_{k+1}^n}^n = \bar{X}_{t_k^n}^n + \frac{T}{n} b(t_k^n, \bar{X}_{t_k^n}^n) + \sqrt{\frac{T}{n}} \sigma(t_k^n, \bar{X}_{t_k^n}^n) U_{k+1}^{(n)},\ k = 0, \ldots, n, \quad \bar{X}_0^n = x$$

where $(U_k^{(n)})_{k=1,\ldots,n}$ is an i.i.d. sequence of $\mathcal{N}(0; I_q)$-distributed random vectors representative of the Brownian increments *i.e.*

$$W_{t_k^n} - W_{t_{k-1}^n} = \sqrt{\frac{T}{n}} U_k^{(n)},\ k = 1, \ldots, n.$$

As $T = t_n^n$, the quantity $\mathbb{E} f(\bar{X}_T^n)$ is the counterpart of $\mathbb{E}\, f(X_T^{0,x})$ for the Euler scheme. Assume b and σ are Lipschitz continuous in (t,x) so that the regularity assumption of Theorem 1.2 is satisfied. Then,

$$\sup_{n \geq 1} \Big\| \max_{0 \leq k \leq n} |\bar{X}_{t_k^n}^n| \Big\|_{L^p(\mathbb{P})} + \Big\| \sup_{t \in [0,T]} |X_t^{0,x}| \Big\|_{L^p(\mathbb{P})} \leq \kappa_{p,b,\sigma,T} \big(1 + |x|\big) \tag{1.20}$$

and, on the other hand, the discrete time Euler scheme *strongly* converges to X for the sup norm in every $L^p(\mathbb{P})$ at rate $\sqrt{\frac{1}{n}}$ in the following sense

$$\Big\| \max_{k=0,\ldots,n} |\bar{X}_{t_k^n}^n - X_{t_k^n}^{0,x}| \Big\|_{L^p(\mathbb{P})} \leq C_{p,b,\sigma,T} \sqrt{\frac{T}{n}} \big(1 + |x|\big).$$

As a consequence, $\mathbb{E}\, f(\bar{X}_T^n) \to \mathbb{E}\, f(X_T^{0,x})$ with a $O\Big(\sqrt{\frac{1}{n}}\Big)$-rate as the step $\frac{T}{n}$ goes to 0 if f is Lipschitz continuous. This latter convergence still holds, without rate, if f is continuous with polynomial growth. It can be obtained under less stringent assumptions on b and σ (continuity in (t,x) and linear growth in x uniformly in t) since then a functional weak convergence holds.

By contrast, if b, σ and f are smooth enough then, the so-called *weak error* $\mathbb{E}\, f(\bar{X}_T^n) - \mathbb{E}\, f(X_T^{0,x})$ can be investigated *directly* by more analytic methods. As a result, a (faster) $O\big(\frac{1}{n}\big)$-rate can be established (see [74]). This rate can be extended to bounded Borel functions f, provided σ satisfies a uniform ellipticity property – or even a hypo-ellipticity assumption "à la Hörmander" for a modified Euler scheme – as proved in a celebrated Bally-Talay's paper (see [7]). This yields

$$u(0,x) = \mathbb{E}\, f(X_T^{0,x}) = \mathbb{E}\, f(\bar{X}_T^n) + O\Big(\frac{1}{n}\Big).$$

The point of interest at this stage is of course that the expectation $\mathbb{E}\, f(\bar{X}^n_{T})$ can be computed by *simulation* since the Euler scheme can be straightforwardly simulated as soon as b and σ are computable functions (and X_0 itself can be simulated). So, we can implement a Monte Carlo simulation to compute $\mathbb{E}f(\bar{X}^n_{T})$ *i.e.* simulate M i.i.d. copies $\big((\bar{X}^n_{T})_m\big)_{m=1,\ldots,M}$ of the above Euler scheme at time $T = t^n_n$ and approximate $\mathbb{E}\, f(\bar{X}^n_{T})$ by the strong Law of Large Numbers

$$\mathbb{E}\, f(\bar{X}^n_{T}) \approx \frac{1}{M}\sum_{m=1}^{M} f\big((\bar{X}^n_{T})_m\big)$$

since *a.s.* convergence holds as $M \to +\infty$. This second error (known as the Monte Carlo or the *statistical error*) is of order $O(\frac{1}{\sqrt{M}})$ owing to the Central Limit Theorem which provides (asymptotic) confidence intervals for an *a priori* prescribed given confidence level involving the asymptotic variance

$$\mathrm{Var}\big(f(\bar{X}^n_{T})\big) = \mathbb{E}\big(f(\bar{X}^n_{T}) - \mathbb{E}\, f(\bar{X}^n_{T})\big)^2 = \mathbb{E}f(\bar{X}^n_{T})^2 - \big(\mathbb{E}\, f(\bar{X}^n_{T})\big)^2.$$

This variance can be expressed by expectations of functions of $\bar{X}^n_{T}$, consequently it can be computed *on line* as a companion parameter of the original Monte Carlo simulation. By the way, note that one often has $\mathrm{Var}\big(f(\bar{X}^n_{T})\big) \approx \mathrm{Var}\big(f(X_{T})\big)$, either because f is continuous or because the diffusion is "elliptic enough", see above. For more details on these elementary aspects of the Monte Carlo method, we refer to classical textbooks devoted Monte Carlo simulation and Numerical Probability (see [43] for a more PDE oriented introduction to Monte Carlo method or [31, 56] for more connections with Finance, among many others).

The main asset of this approach is that it is dimension free, in the sense that its complexity grows linearly with the dimension d of the diffusion of interest, with little influence of the ellipticity of the function a, at least when the function f is regular as we just saw.

▷ Step 2b *(Quantization based cubature formula)*. If one has many computations to carry out with the same operator L, *e.g.* for various functions f, it may be interesting to replace the Monte Carlo simulation by a *cubature* formula based on an optimal quantization of $\bar{X}^n_{T}$. To perform this quantization, as it will be seen further on in Section 3, one can rely on a stochastic optimization procedure which can be viewed as a kind of *compressed Monte Carlo* simulation. In that perspective, one faces now the following chain of approximations

$$u(0,x) = \mathbb{E}\, f(X^{0,x}_{T}) \approx \mathbb{E}\, f(\bar{X}^n_{T}) \approx \mathbb{E}\, f\Big(\widehat{\bar{X}^n_{T}}^{\Gamma^{(N)}}\Big)$$

where $\Gamma^{(N)}$ is an optimal (quadratic) quantization grid for the random vector $\bar{X}^n_{T}$.

1.5. Toward automatic meshing.

An alternative to the direct quantization procedure is to consider the grid $\Gamma^{(N)}$ as a starting point to produce an optimized mesh for the numerical solving of the original *PDE* by deterministic schemes like finite element or finite volumes methods, etc. In such an approach, an optimal grid needs to be produced at each discretization time t^n_k. This approach has been widely investigated by Gunzberger's group in Florida (USA) (see *e.g.* [25] and the references therein). More recently, a new concept of quantization (dual quantization, see [64]) has refined this point of view by switching from Voronoi diagrams to a direct approach based on optimized Delaunay triangulations. The resulting grids are better adapted to deterministic numerical analysis methods in medium dimensions.

1.5.1. *From optimal stopping theory to variational inequalities*

▷ *Discrete time optimal stopping theory in a Markov framework.* We consider a standard discrete time Markovian framework: let $(X_k)_{0\le k\le n}$ be an $\mathbb{R}^d$-valued $(\mathcal{F}_k)_{0\le k\le n}$-Markov chain defined on the filtered probability space $(\Omega, \mathcal{A}, (\mathcal{F}_k)_{0\le k\le n}, \mathbb{P})$: the chain is $(\mathcal{F}_k)_{0\le k\le n}$-adapted, *i.e.* X_k is $\mathcal{F}_k$-measurable for every $k = 0, \ldots, n$, with transitions

$$P_k(x, dy) = \mathbb{P}\big(X_{k+1} \in dy \,|\, X_k = x\big),\ k = 0, \ldots, n-1,$$

so that for every bounded or non-negative Borel function $f:\mathbb{R}^d\to\mathbb{R}$, $P_kf(x)=\displaystyle\int_{\mathbb{R}^d}f(y)P_k(x,dy)$ and

$$\mathbb{E}\big(f(X_{k+1})\,|\,\mathcal{F}_k\big)=\mathbb{E}\big(f(X_{k+1})\,|\,X_k\big)=P_k(f)(X_k)\ a.s.$$

From now on, we denote by $\mathcal{F}$ the filtration $(\mathcal{F}_k)_{0\le k\le n}$. Intuitively, $\mathcal{F}_k$ is a σ-field of $\mathcal{A}$ which represents the *observable* (or *available*) *information* at time k. Let $Z=(Z_k)_{0\le k\le n}$ be an $\mathcal{F}$*-adapted* obstacle/payoff sequence of non-negative *integrable* random variables of the form

$$0\le Z_k=f_k(X_k)\in L^1(\Omega,\mathcal{F}_k,\mathbb{P}),\ k=0,\dots,n.$$

In term of modeling, this can be understood as follows: an agent plays a stochastic game. Each round of the game takes place at time $k\in\{0,\dots,n\}$. The random variable Z_k represents the reward when leaving the game at time k. The question: "Is there an optimal way to quit the game in order to maximize the gain?"

By "quitting the game", we mean leaving possibly at a *random* time $\tau:\Omega\to\{0,\dots,n\}$ but always honestly *i.e.* in such a way that, for every $\ell\in\{0,\dots,n\}$, the event

$$\{\tau=\ell\}=\big\{\omega\in\Omega\,|\,\tau(\omega)=\ell\big\}\in\mathcal{F}_\ell.$$

Thus, if the agent adopts this strategy τ, the available information that leads him/her to leave the game at time $\tau(\omega)$ is the following: if $\ell=\tau(\omega)$, for every $A\in\mathcal{F}_\ell$, he/she knows whether ω belongs or not to A. In particular the agent has observed the whole path $(X_k(\omega))_{0\le k\le\tau(\omega)}$ since the chain is $\mathcal{F}$*-adapted*. Such a random variable is called an $\mathcal{F}$*-stopping time*. In practice, one can imagine that reasonable strategies will involve or rely on the payoff sequence $Z_\ell)_\ell=f_\ell(X_\ell)$, $\ell=0,\dots,n$.

Imagine now that this agent enters the game at time $k\in\{0,\dots,n\}$. The aim of the agent is to attain the optimal possible mean gain given the available information at time k, namely

$$U_k=\mathbb{P}\text{-esssup}\Big\{\mathbb{E}\big(Z_\tau\,|\,\mathcal{F}_k\big),\ \tau:(\Omega,\mathcal{A})\to\{k,\dots,n\},\ \mathcal{F}\text{-stopping time}\Big\}\tag{1.21}$$

with an optimal mean gain given by $\mathbb{E}\,U_k$. The next question is to know whether there is an *optimal stopping time* (or equivalently an *optimal strategy*), when starting the game at time k, *i.e.* a $\{k,\dots,n\}$-valued $\mathcal{F}$-stopping time τ_k satisfying

$$U_k=\mathbb{E}\big(Z_{\tau_k}\,|\,\mathcal{F}_k\big).$$

For more details on this topic we refer to [53] or [42] (Chapter 2) or, more recently, [44].

The sequence $U=(U_k)_{0\le k\le n}$ is known as the $(\mathbb{P},\mathcal{F})$-*Snell envelope* of the sequence $(Z_k)_{0\le k\le n}$.

From a numerical point of view, we want to compute, or at least approximate, this Snell envelope, especially at time 0, and the related optimal stopping time τ_0 (if any).

The first important result of discrete time optimal stopping theory is the following Backward Dynamic Programming Principle (*BDPP*). Temporarily assume that $(Z_k)_{0\le k\le n}$ is an $\mathcal{F}$-adapted general sequence of non-negative integrable random variables.

Proposition 1.4. (*a*) *The* $(\mathbb{P},\mathcal{F})$*-Snell envelope* $(U_k)_{0\le k\le n}$ *satisfies the following BDDP:*

$$U_n=Z_n\ \ and\ \ U_k=\max\Big(Z_k,\mathbb{E}\big(U_{k+1}\,|\,\mathcal{F}_k\big)\Big),\ k=0,\dots,n-1,\tag{1.22}$$

and $\tau_k=\min\big\{\ell\in\{k,\dots,n\}\,|\,U_\ell=Z_\ell\big\}$ *is an optimal stopping time at time k i.e.*

$$U_k=\mathbb{E}\big(Z_{\tau_k}\,|\,\mathcal{F}_k\big).$$

(*b*) *Furthermore, if* $Z_k = f_k(X_k)$ *for every* $k \in \{0,\ldots,n\}$, *there exists a Borel function* $u_k : \mathbb{R}^d \to \mathbb{R}$ *such that*

$$U_k = u_k(X_k),\; k = 0,\ldots,n,$$

and

$$U_n = f_n(X_n) \quad \textit{and} \quad U_k = \max\Big(f_k(X_k), \mathbb{E}\big(U_{k+1}\,|\,X_k\big)\Big),\; k = 0,\ldots,n-1, \tag{1.23}$$

or, equivalently, the sequence $(u_k)_{0\le k\le n}$ *satisfies*

$$u_n = f_n \qquad \textit{and} \qquad u_k = \max\big(f_k, P_k u_{k+1}\big),\; k = 0,\ldots,n-1.$$

Proof. (a) We prove this claim by a backward induction on k. The fact that $U_n = Z_n$ is obvious since $\tau_n = n$ is the only $\{n\}$-valued stopping time (hence optimal at time n).

Now let $k \in \{0,\ldots,n-1\}$. Assume that $\tau_{k+1} = \min\big\{\ell \in \{k+1,\ldots,n\}\,|\,U_\ell = Z_\ell\big\}$ is an optimal $\mathcal{F}$-stopping time at time $k+1$ *i.e.*

$$U_{k+1} = \mathbb{E}\big(Z_{\tau_{k+1}}\,|\,\mathcal{F}_{k+1}\big).$$

As $\tau_{k+1} \ge k+1$ is in particular a $\{k,\ldots,n\}$-valued $\mathcal{F}$-stopping time, it follows that

$$\begin{aligned}\mathbb{E}\big(U_{k+1}\,|\,\mathcal{F}_k\big) &= \mathbb{E}\Big(\mathbb{E}\big(Z_{\tau_{k+1}}\,|\,\mathcal{F}_{k+1}\big)\,|\,\mathcal{F}_k\Big)\\ &= \mathbb{E}\big(Z_{\tau_{k+1}}\,|\,\mathcal{F}_k\big)\\ &\le U_k\end{aligned}$$

where we used to get the inequality in the last line the definition (1.21) of the Snell envelope. Since $U_k \ge Z_k$, by considering the deterministic stopping time $\tau = k$, we finally get

$$U_k \ge \max\Big(Z_k, \mathbb{E}\big(U_{k+1}\,|\,\mathcal{F}_k\big)\Big).$$

To prove the reverse inequality and establish the *BDPP* at time k, we consider a generic $\{k,\ldots,n\}$-valued $\mathcal{F}$-stopping time τ. Then, noting that $\{\tau \ge k+1\} = {}^c\{\tau \le k\} \in \mathcal{F}_k$,

$$\mathbb{E}\big(Z_\tau\,|\,\mathcal{F}_k\big) = Z_k \mathbf{1}_{\{\tau=k\}} + \mathbb{E}\Big(Z_{\tau\vee(k+1)}\,|\,\mathcal{F}_k\Big)\mathbf{1}_{\{\tau\ge k+1\}} \quad \mathbb{P}\text{-}a.s.$$

Now, using that $\tau\vee(k+1)$ is a $\{k+1,\ldots,n\}$-valued $\mathcal{F}$-stopping time,

$$\begin{aligned}\mathbb{E}\Big(Z_{\tau\vee(k+1)}\,|\,\mathcal{F}_k\Big) &= \mathbb{E}\Big(\mathbb{E}\Big(Z_{\tau\vee(k+1)}\,|\,\mathcal{F}_{k+1}\Big)\,|\,\mathcal{F}_k\Big) \quad \mathbb{P}\text{-}a.s.\\ &\le \mathbb{E}\Big(U_{k+1}\,|\,\mathcal{F}_k\Big) \quad \mathbb{P}\text{-}a.s.\end{aligned}$$

by the definition (1.21) of U_{k+1}. As a consequence,

$$\begin{aligned}\mathbb{E}\big(Z_\tau\,|\,\mathcal{F}_k\big) &\le Z_k \mathbf{1}_{\{\tau=k\}} + \mathbb{E}\Big(U_{k+1}\,|\,\mathcal{F}_k\Big)\mathbf{1}_{\{\tau\ge k+1\}}\\ &\le \max\Big(Z_k, \mathbb{E}\Big(U_{k+1}\,|\,\mathcal{F}_k\Big)\Big).\end{aligned}$$

Taking the $\mathbb{P}$-esssup over all such $\{k,\ldots,n\}$-valued $\mathcal{F}$-stopping times, we get

$$U_k \le \max\Big(Z_k, \mathbb{E}\Big(U_{k+1}\,|\,\mathcal{F}_k\Big)\Big) \quad \mathbb{P}\text{-}a.s.$$

which in turn implies that the Snell envelope satisfies (1.22) at time k. Let us deal now with the optimal stopping time. One checks from its definition that

$$\tau_k = k\mathbf{1}_{\{U_k=Z_k\}} + \tau_{k+1}\mathbf{1}_{\{U_k\neq Z_k\}}.$$

Using that both events $\{U_k = Z_k\}$ and $\{U_k \neq Z_k\}$ lie in $\mathcal{F}_k$ and that $\{U_k \neq Z_k\} \subset \big\{\mathbb{E}(U_{k+1}|\mathcal{F}_k) = U_k\big\}$, we get the following string of equalities

$$\begin{aligned}
\mathbb{E}\big(Z_{\tau_k}\,|\,\mathcal{F}_k\big) &= Z_k\mathbf{1}_{\{U_k=Z_k\}} + \mathbb{E}\Big(Z_{\tau_{k+1}}\,|\,\mathcal{F}_k\Big)\mathbf{1}_{\{U_k\neq Z_k\}} \quad \mathbb{P}\text{-}a.s.\\
&= U_k\mathbf{1}_{\{U_k=Z_k\}} + \mathbb{E}\Big(\mathbb{E}\big(Z_{\tau_{k+1}}\,|\,\mathcal{F}_{k+1}\big)\,|\,\mathcal{F}_k\Big)\mathbf{1}_{\{U_k\neq Z_k\}} \quad \mathbb{P}\text{-}a.s.\\
&= U_k\mathbf{1}_{\{U_k=Z_k\}} + \mathbb{E}\Big(U_{k+1}\,|\,\mathcal{F}_k\Big)\mathbf{1}_{\{U_k\neq Z_k\}} \quad \mathbb{P}\text{-}a.s.\\
&= U_k\mathbf{1}_{\{U_k=Z_k\}} + U_k\mathbf{1}_{\{U_k\neq Z_k\}} \quad \mathbb{P}\text{-}a.s.\\
&= U_k.
\end{aligned}$$

(b) This straightforwardly follows from the fact that, owing to the definition of the Markov transitions, if $U_{k+1} = u_{k+1}(X_{k+1})$ for $k\in\{0,\ldots,n-1\}$, then

$$\mathbb{E}\big(U_{k+1}\,|\,\mathcal{F}_k\big) = \mathbb{E}\big(U_{k+1}\,|\,X_k\big) = P_k u_{k+1}(X_k) \quad \mathbb{P}\text{-}a.s.$$

Hence, $U_k = \max\Big(f_k(X_k), \mathbb{E}\big(U_{k+1}\,|\,\mathcal{F}_k\big)\Big) = \max\big(f_k(X_k), P_k u_{k+1}(X_k)\big) = u_k(X_k).$ □

Remark. The above optimal stopping time τ_k may be not unique, but one shows that it is always the lowest stopping time for the game starting at time k: if $\tilde{\tau}_k$ is another optimal stopping time for the game starting at k, it satisfies $\tilde{\tau}_k \geq \tau_k$ *a.s.*. Moreover it follows from the above proof that the sequence of optimal stopping times $(\tau_k)_{0\leq k\leq n}$ satisfies the *dual backward dynamic programing principle*

$$\tau_k = k\mathbf{1}_{\{U_k=Z_k\}} + \tau_{k+1}\mathbf{1}_{\{U_k\neq Z_k\}}. \tag{1.24}$$

This second backward dynamic programming principle – sometimes called *dual* – is often used in regression methods to compute the Snell envelope (see *e.g.* Longstaff-Schwarz's paper [46]).

▷ *Approximation of the Snell envelope by a quantization tree.* The starting idea of the *quantization tree method*, originally introduced in [2], is to approximate the whole Markovian dynamics of the chain $X = (X_k)_{0\leq k\leq n}$ using a sequence of quantizations $(X_k)_{0\leq k\leq n}$ to produce a *skeleton* of the whole distribution of X, namely the tree quantization tree defined as the quantization grids $\Gamma_k = \{x_1^k, \ldots, x_{N_k}^k\}$ of X_k, $k = 0, \ldots, n$, "connected" by the transitions weights w_{ij}^k between states x_i^k and x_j^{k+1} defined for every $k\in\{0,\ldots,n-1\}$ by

$$w_{ij}^k = \mathbb{P}\big(\widehat{X}_{k+1} = x_j^{k+1}\,|\,\widehat{X}_k = x_i^k\big),\ 1 \leq i \leq N_k,\ 1 \leq j \leq N_{k+1}.$$

Although we will rely on these transitions below, it is important to keep in mind that the sequence of quantizations $(\widehat{X}_k)_{0\leq k\leq n}$ *is not a Markov chain.*

At this stage, the idea is to mimic the *BDPP* (1.23) satisfied by the Snell envelope $(U_k)_{0\leq k\leq n}$ by replacing X_k by a Γ_k-valued quantization $\widehat{X}_k = q(X_k)$ where $q : \mathbb{R}^d \to \Gamma_k$ is a Borel function. In what follows we will assume that $q = \pi_k$ is a nearest neighbor projection on Γ_k, so that $\widehat{X}_k$ is a Voronoi quantization, though not always necessary. Moreover, as already seen in the introduction one can also choose these grids Γ_k so as to optimize the L^p-mean quantization error criterion $\|X_k - \widehat{X}_k^\Gamma\|_{L^p(\mathbb{P})}$ (for algorithmic aspects, see section 2.3.4 further on).

Let $(f_k(\widehat{X}_k))_{0\le k\le n}$ be the sequence of quantized payoffs/obstacles. The key point, since the sequence $(\widehat{X}_k)_{0\le k\le n}$ is not a Markov chain, is *to force this Markov property* in the *BDPP*. Doing so leads to introduce the *pseudo-Snell envelope* $(\widehat{U}_k)_{0\le k\le n}$ defined by the following Quantized Backward (pseudo-)Dynamic Programming Principle:

$$(QBDPP)\ \equiv\quad \widehat{U}_n = f_n(\widehat{X}_n),\quad \widehat{U}_k = \max\Big(f_k(\widehat{X}_k), \mathbb{E}\big(\widehat{U}_{k+1}\,|\,\widehat{X}_k\big)\Big),\ k=0,\ldots,n-1. \tag{1.25}$$

The forcing of the Markov property is obtained by directly *conditioning by the single random vector* $\widehat{X}_k$ rather than by the σ-field $\widehat{\mathcal{F}}_k := \sigma(\widehat{X}_\ell,\, 0\le \ell\le k)$.

Then, it is straightforward still by a backward induction that, for every $k\in\{0,\ldots,n\}$, that there exists a Borel function $u_k:\mathbb{R}^d\to\mathbb{R}_+$, such that

$$\widehat{U}_k = \widehat{u}_k(\widehat{X}_k),\ k=0,\ldots,n-1.$$

From a computational point of view, (1.25) reads "in distribution",

$$\begin{aligned}\widehat{u}_n(x_i^n) &= f_n(x_i^n),\ 1\le i\le N_n,\\ \widehat{u}_k(x_i^k) &= \max\Big(f_k(x_i^k), \sum_{j=1}^{N_{k+1}} w_{ij}^k\,\widehat{u}_{k+1}(x_j^{k+1})\Big),\ 1\le i\le N_k,\ 1\le j\le N_{k+1},\ k=0,\ldots,n-1,\end{aligned} \tag{1.26}$$

where $\Gamma_k=\{x_1^k,\ldots,x_{N_k}^k\}$, $k=0,\ldots,n$. See subsection 2.3.4 for details on the practical implementation, including the computation by Monte Carlo simulation of the transition weights w_{ij}^k.

▷ *Error bounds.* The following theorem establishes the control on the approximation of the true Snell envelope $(U_k)_{0\le k\le n}$ by its quantized counterpart $(\widehat{U}_k)_{0\le k\le n}$ using the L^p-mean quantization errors $\|X_k-\widehat{X}_k\|_{L^p(\mathbb{P})}$.

Theorem 1.3 (see [2] (2001), [64] (2011))**.** *Assume that all functions $f_k:\mathbb{R}^d\to\mathbb{R}_+$, $k=0,\ldots,n-1$, are Lipschitz continuous and that the transitions $P_k(x,dy)=\mathbb{P}(X_{k+1}\in dy\,|\,X_k=x)$ propagate Lipschitz continuity i.e.*

$$[P_k]_{\rm Lip} = \sup_{[g]_{\rm Lip}\le 1}[P_kg]_{\rm Lip} < +\infty,\qquad k=0,\ldots,n.$$

Set $[P]_{\rm Lip} = \max\limits_{0\le k\le n-1}[P_k]_{\rm Lip}$ and $[f]_{\rm Lip}=\max_{0\le k\le n}[f_k]_{\rm Lip}$.

Let $p\in[1,+\infty)$. We assume that $\sum\limits_{k=1}^n \|X_k\|_{L^p(\mathbb{P})}+\|\widehat{X}_k\|_{L^p(\mathbb{P})}<+\infty$.

(a) For every $k\in\{0,\ldots,n\}$,

$$\|U_k-\widehat{U}_k\|_{L^p(\mathbb{P})}\le 2[f]_{\rm Lip}\sum_{\ell=k}^n\big([P]_{\rm Lip}\vee 1\big)^{n-\ell}\|X_\ell-\widehat{X}_\ell\|_{L^p(\mathbb{P})}.$$

(b) If $p=2$, for every $k\in\{0,\ldots,n\}$,

$$\|U_k-\widehat{U}_k\|_{L^2(\mathbb{P})}\le \sqrt{2}[f]_{\rm Lip}\left(\sum_{\ell=k}^n\big([P]_{\rm Lip}\vee 1\big)^{2(n-\ell)}\|X_\ell-\widehat{X}_\ell\|^2_{L^2(\mathbb{P})}\right)^{\frac12}. \tag{1.27}$$

Remark. In fact, the error bounds established in this theorem remain mathematically true as soon as the "quantizations" $\widehat{X}_k$ are simply $\sigma(X_k)$-measurable, *i.e.* when $\widehat{X}_k=q_k(X_k)$ where $q_k:\mathbb{R}^d\to\mathbb{R}^d$ is a Borel functions for every $k\in\{0,\ldots,n\}$, with at mot linear growth ($q_k(x)\le C_k(1+|x|)$). Thus, if the functions q_k

take countably many values, even (1.26) still makes sense; otherwise sums should be replaced by integrals with respect to the conditional distributions $\mathcal{L}(\widehat{X}_{k+1}\,|\,\widehat{X}_k)$ and the computational tractability is usually lost.

Proof. (b) Step 1. First, we control the Lipschitz constants of the functions u_k. It follows from the elementary inequality $|\sup_{i\in I} a_i - \sup_{i\in I} b_i| \le \sup_{i\in I} |a_i - b_i|$, a_i, $b_i \in \mathbb{R}$, $i \in I$, that

$$\begin{aligned}[u_k]_{\rm Lip} &\le \max\big([f_k]_{\rm Lip}, [P_k u_{k+1}]_{\rm Lip}\big)\\ &\le \max\big([f]_{\rm Lip}, [P_k]_{\rm Lip}[u_{k+1}]_{\rm Lip}\big)\end{aligned}$$

with the convention $[u_{n+1}]_{\rm Lip} = 0$. A straightforward backward induction yields

$$[u_k]_{\rm Lip} \le [f]_{\rm Lip}\big([P]_{\rm Lip} \vee 1\big)^{n-k}. \tag{1.28}$$

Step 2. We focus on claim (b) (quadratic case $p = 2$). First, we derive from Proposition 1.2(a) applied to $X = X_{k+1}$ and $Y = X_k$, $Q = P_k$ and $f = u_{k+1}$, $g = \widehat{u}_{k+1}$ and $h = \widehat{u}_k$ that

$$\Big\|\mathbb{E}\big(U_{k+1}|X_k\big) - \mathbb{E}\big(\widehat{U}_{k+1}|\widehat{X}_k\big)\Big\|^2_{L^2(\mathbb{P})} \le [P_k u_{k+1}]^2 \big\|X_k - \widehat{X}_k\big\|^2_{L^2(\mathbb{P})} + \big\|u_{k+1}(X_{k+1}) - \widehat{u}_{k+1}(\widehat{X}_{k+1})\big\|^2_{L^2(\mathbb{P})}. \tag{1.29}$$

Now, it follows by combining the original and the quantized dynamic programming formulas (1.22) and (1.25) that

$$|U_k - \widehat{U}_k| \le \max\Big(|f_k(X_k) - f_k(\widehat{X}_k)|, \big|\mathbb{E}\big(U_{k+1}|X_k\big) - \mathbb{E}\big(\widehat{U}_{k+1}|\widehat{X}_k\big)\big|\Big)$$

so that

$$|U_k - \widehat{U}_k|^2 \le |f_k(X_k) - f_k(\widehat{X}_k)|^2 + \big|\mathbb{E}\big(U_{k+1}|X_k\big) - \mathbb{E}\big(\widehat{U}_{k+1}|\widehat{X}_k\big)\big|^2.$$

Taking expectation and plugging (1.29) in the above inequality yields for every $k \in \{0, \ldots, n-1\}$,

$$\big\|U_k - \widehat{U}_k\big\|^2_{L^2(\mathbb{P})} \le \Big([f]^2_{\rm Lip} + [P]^2_{\rm Lip}[u_{k+1}]^2_{\rm Lip}\Big)\big\|X_k - \widehat{X}_k\big\|^2_{L^2(\mathbb{P})} + \big\|U_{k+1} - \widehat{U}_{k+1}\big\|^2_{L^2(\mathbb{P})}$$

still with the convention $[u_{n+1}]_{\rm Lip} = 0$. Now, using (1.28), we obtain

$$\begin{aligned}[f]^2_{\rm Lip} + [P]^2_{\rm Lip}[u_{k+1}]^2_{\rm Lip} &\le [f]^2_{\rm Lip} + [P]^2_{\rm Lip}\big(1 \vee [P]_{\rm Lip}\big)^{2(n-(k+1))}\\ &\le 2[f]^2_{\rm Lip}\big(1 \vee [P]_{\rm Lip}\big)^{2(n-k)}.\end{aligned}$$

Consequently

$$\begin{aligned}\big\|U_k - \widehat{U}_k\big\|^2_{L^2(\mathbb{P})} &\le 2\sum_{\ell=k}^{n-1}[f]^2_{\rm Lip}\big(1 \vee [P]_{\rm Lip}\big)^{2(n-\ell)}\big\|X_\ell - \widehat{X}_\ell\big\|^2_{L^2(\mathbb{P})} + [f]^2_{\rm Lip}\big\|X_n - \widehat{X}_n\big\|^2_{L^2(\mathbb{P})}\\ &\le 2[f]^2_{\rm Lip}\sum_{\ell=k}^{n}\big(1 \vee [P]_{\rm Lip}\big)^{2(n-\ell)}\big\|X_\ell - \widehat{X}_\ell\big\|^2_{L^2(\mathbb{P})}\end{aligned}$$

which completes the proof.

Claim (a) is established following the above lines of the proof, relying now on Claim (b) of Proposition 1.2. and Minkowski's Inequality instead of the Pythagorus like Theorem □

Example of application: the Euler scheme. Let $(\bar{X}^n_{t^n_k})_{0\le k\le n}$ be the Euler scheme with step $\frac{T}{n}$ of the d-dimensional diffusion $(X^{0,x}_t)_{t\in[0,T]}$, solution to the *SDE* (1.18). It defines a homogeneous Markov chain with transition

$$\bar{P}^n_k g(x) = \mathbb{E}\, g\left(x + \frac{T}{n} b(t^n_k, \bar{X}^n_{t^n_k}) + \sigma(t^n_k, \bar{X}^n_{t^n_k})\sqrt{\frac{T}{n}}Z\right), \quad Z \overset{\mathcal{L}}{\sim} \mathcal{N}(0, I_q).$$

If f is Lipschitz continuous,

$$\begin{aligned}\big|\bar P^n_k g(x)-\bar P^n_k g(x')\big|^2 &\le [g]^2_{\rm Lip}\mathbb{E}\Big|x-x'+\frac{T}{n}\big(b(t^n_k,x)-b(t^n_k,x')\big)+\sqrt{\frac{T}{n}}\big(\sigma(t^n_k,x)-\sigma(t^n_k,x')\big)Z\Big|^2\\ &\le [g]^2_{\rm Lip}\left(\Big|x-x'+\frac{T}{n}\big(b(t^n_k,x)-b(t^n_k,x')\big)\Big|^2+\Big\|\sigma(t^n_k,x)-\sigma(t^n_k,x')\Big\|^2\frac{T}{n}\right)\\ &\le [g]^2_{\rm Lip}|x-x'|^2\left(1+\frac{T}{n}[\sigma]^2_{\rm Lip}+\frac{2T}{n}[b]_{\rm Lip}+\frac{T^2}{n^2}[b]^2_{\rm Lip}\right)\end{aligned}$$

where $\|A\| = \sqrt{{\rm Tr}(AA^*)}$, Tr stands for the trace of a square matrix, A^* stands for the transpose of the $d\times q$-matrix A. The coefficient $[\sigma]_{\rm Lip}$ should be understood as the Lipschitz coefficient of $\sigma:[0,T]\times\mathbb{R}^d\to(\mathbb{M}(d,q,\mathbb{R}),\|\,.\|)$ in x with respect to the Euclidean norm $|\,.\,|$ uniformly in $t\in[0,T]$ ($[b]_{\rm Lip}$ is defined accordingly). As a consequence

$$[\bar P^n_k g]_{\rm Lip}\le\left(1+\frac{C_{b,\sigma,T}T}{n}\right)[g]_{\rm Lip},\;\; k=0,\ldots,n-1,$$

where $C_{b,\sigma,T}=[b]_{\rm Lip}+\dfrac12\Big([b]^2_{\rm Lip}T+[\sigma]^2_{\rm Lip}\Big)$ *i.e.*

$$[\bar P^n]_{\rm Lip}\le 1+\frac{C_{b,\sigma,T}T}{n}.$$

Let $(\bar U_k)_{0\le k\le n}$ denote the $(\mathcal{F}^W_{t^n_k})_{0\le k\le n}$-Snell envelope of the payoff process $\big(f_k(\bar X^n_{t^n_k})\big)_{0\le k\le n}$ and let $(\widehat U_k)_{0\le k\le n}$ be the pseudo-Snell envelope associated by (1.25) to a quantized version $(\widehat{\bar X^n_{t^n_k}})_{0\le k\le n}$ of this payoff process. Applying the control established in claim (b) of the above theorem yields with obvious notations

$$\begin{aligned}\big\|\bar U_k-\widehat U_k\big\|_{L^2(\mathbb{P})} &\le \sqrt2[f]_{\rm Lip}\left(\sum_{\ell=k}^n\Big(1+\frac{C_{b,\sigma,T}T}{n}\Big)^{2(n-\ell)}\big\|X_\ell-\widehat X_\ell\big\|^2_{L^2(\mathbb{P})}\right)^{\frac12}\\ &\le \sqrt2[f]_{\rm Lip}\left(\sum_{\ell=k}^n e^{2C_{b,\sigma,T}(T-t^n_\ell)}\big\|X_\ell-\widehat X_\ell\big\|^2_{L^2(\mathbb{P})}\right)^{\frac12} \qquad (1.30)\\ &\le \sqrt2[f]_{\rm Lip}e^{C_{b,\sigma,T}T}\left(\sum_{\ell=k}^n\big\|X_\ell-\widehat X_\ell\big\|^2_{L^2(\mathbb{P})}\right)^{\frac12}. \qquad (1.31)\end{aligned}$$

The fact to be emphasized concerning the upper bound (1.31) is that the real constants on the left hand side only depend on b, σ and T *but not on* n (except for the range of the sum itself of course) whereas in the sharper (1.30) the constants do not explode with n. We will see further on that (1.30) can be used to calibrate the sizes of the quantization grids associated to the quantizations $\widehat X_k$ (see section 2.3.4). (1.30) and Finally note that the above computations hold more generally for the Euler scheme of a diffusion driven by a Lévy processes Z with Lévy measure ν satisfying $\nu(z^2)<+\infty$ *i.e.* $Z_t\in L^2$ for every $t\in[0,T]$.

▷ *Connection with parabolic variational inequalities.* We consider the parabolic variational inequality on $[0,T]\times\mathbb{R}^d$ defined by

$$\max\left(f-u,\frac{\partial u}{\partial t}+Lu\right)(t,x)=0,\;(t,x)\in[0,T)\times\mathbb{R}^d,\quad u(T,\,.\,)=f(T,\,.\,) \qquad (1.32)$$

where L is the operator introduced in (1.17) (which is for the probabilist the generator of the diffusion (1.18)). We assume that $f:[0,T]\times\mathbb{R}^d\to\mathbb{R}_+$ is (at least) continuous with polynomial growth in the space variable x, uniformly in $t\in[0,T]$.

From now on, we will switch to a completely heuristic reasoning in order to highlight in a simpler way the connection between the above variational inequality and optimal stopping theory in continuous time. This connection holds through a probabilistic representation formula involving the diffusion process (1.18) in the same spirit as that which holds for parabolic PDEs through the Feynman-Kac formula. A probabilistic representation of this variational inequality (1.32) is provided, under appropriate conditions that we will not detail here (see [5,6] or, more recently, [44] and the references therein for a rigorous presentation in various settings), by the continuous time optimal stopping problem, related to the diffusion process $(X^{x,t})_{s\in[t,T]}$ solution to Equation (1.18) and the *obstacle process* $Z_t := f(t, X^{x,0}_t) \geq 0$, $t \in [0,T]$. This obstacle process is $\mathcal{F}^W$-adapted, non-negative and continuous (hence predictable) where $\mathcal{F}^W$ denotes the augmented filtration of the Brownian motion W. The function f having polynomial growth in x uniformly in $t \in [0,T]$, then $\sup_{t\in[0,T]} Z_t \in L^1(\mathbb{P})$. We define the $\mathbb{P}$-Snell envelope $(U_t)_{t\in[0,T]}$ by

$$U_t = \mathbb{P}\text{-esssup}\Big\{\mathbb{E}\big(f(\tau, Z_\tau)\,|\,\mathcal{F}_t\big),\ \tau \in \mathcal{T}^W_{t,T}\Big\} \tag{1.33}$$

where $\mathcal{T}^W_{t,T}$ denotes the set of $\mathcal{F}^W$-stopping times $\tau : (\Omega, \mathcal{A}, \mathbb{P}) \to [t,T]$ *i.e.* $[t,T]$-valued random times satisfying

$$\forall\, s \in [t,T],\ \{\tau \leq s\} \in \mathcal{F}^W_s.$$

(This definition implies that for every $s \in [t,T]$, $\{\tau = s\} \in \mathcal{F}^W_s$ but the converse – which is required for technical reasons – is usually not true since $[0,T]$ is not countable.) One shows (see *e.g.* [73]) that under these conditions, there exists a function $u : [0,T] \times \mathbb{R}^d \to \mathbb{R}_+$ such that

$$U_t = u(t, X^{x,0}_t),\ t \in [0,T], \tag{1.34}$$

and, in terms of *réduite*,

$$u(t,x) = \sup\Big\{\mathbb{E}\, f(\tau, X^{x,t}_\tau),\ \tau \in \mathcal{T}^W_{t,T}\Big\},\ t \in [0,T],\ x \in \mathbb{R}^d.$$

Unfortunately, even in simple frameworks, this function u is not smooth enough, say $\mathcal{C}^{1,2}([0,T] \times \mathbb{R}^d, \mathbb{R})$, to apply Itô's formula.

We consider again the uniform mesh of $[0,T]$ of step $\frac{T}{n}$, $t^n_k = \frac{kT}{n}$, $k = 0, \ldots, n$. We can approximate the sequence $(U_{t^n_k})_{0\leq k\leq n}$ by the sequence $(U^n_{t^n_k})_{0\leq k\leq n}$ defined by replacing in (1.33) the set $\mathcal{T}_{t,T}$ by its subset $\mathcal{T}^n_{k,n}$ of $\mathcal{F}^W$-stopping times taking values in $\{t^n_\ell,\ \ell = k, \ldots, n\}$ of $[0,T]$. Stopping times of $\mathcal{T}^n_{k,n}$ are of discrete nature and are subsequently characterized by the simpler property

$$\tau \in \mathcal{T}^n_{k,n} \ \text{ if and only if } \ \{\tau = t^n_\ell\} \in \mathcal{F}^W_{t^n_\ell},\ \ell = k, \ldots, n.$$

As a consequence, $U^n_{t^n_k}$ is defined for every $k \in \{0, \ldots, n\}$ by

$$U^n_{t^n_k} = \mathbb{P}\text{-esssup}\Big\{\mathbb{E}\big(f(\tau, Z_\tau)\,|\,\mathcal{F}_{t^n_k}\big),\ \tau \in \mathcal{T}^n_{k,n}\Big\}.$$

Hence, the sequence $(U^n_{t^n_k})_{0\leq k\leq n}$ is nothing but the $\mathbb{P}$-Snell envelope of $(Z_{t^n_k})_{0\leq k\leq n}$ viewed as a discrete time optimal stopping problem associated to the Markov chain $(X^{x,0}_{t^n_k})_{0\leq k\leq n}$ with transitions $P_k(\xi, dy) = \mathbb{P}(X_{t^n_{k+1}} \in dy\,|\,X_{t^n_k} = \xi)$, $k = 0, \ldots, n-1$. They are usually not explicit and, more important in practice, even not simulatable: more generally, exact simulation of time samples of a d-dimensional diffusion process is impossible, at least at a reasonable cost, as soon as $d \geq 2$ ([4]).

[4] When $d = 1$ an exact (and efficient) simulation method has been devised for diffusion processes in [9]; unfortunately it deeply depends on the scalar feature of the diffusion.

Then, keeping in mind that $U^n_{t^n_n} = U_{_T} = f(T, X^{x,0}_{_T})$, the sequence $(U^n_{t^n_k})_{0\le k\le n}$ satisfies the Markovian *BDPP* (1.23) reading

$$U^n_{t^n_n} = f(T, X^{x,0}_{_T}),\;\; U^n_{t^n_k} = \max\Big(f\big(X^{0,x}_{t^n_k}\big), \mathbb{E}\big(U^n_{t^n_{k+1}}\,|\,X^{0,x}_{t^n_k}\big)\Big),\; 0\le k\le n-1.$$

Consequently, one may consider that $U_{t^n_k}$ is close enough to $U^n_{t^n_k}$ so that the sequence $(U_{t^n_k})_{0\le k\le n}$ approximately satisfies

$$U_{_T} = f(T, X^{x,0}_{_T}),\;\; \max\Big(f(X^{x,0}_{t^n_k}) - U_{t^n_k}, \mathbb{E}\big(U_{t^n_{k+1}}\,|\,X^{x,0}_{t^n_k}\big) - U_{t^n_k}\Big)\approx 0$$

or equivalently ,since $U_{t^n_k} = u(t^n_k, X_{t^n_k})$, $k = 0,\dots,n$, owing to (1.34),

$$u(T, X^{x,0}_{_T}) = f(T, X^{x,0}_{_T}),\;\; 0\approx \max\Big(f(X^{0,x}_{t^n_k}) - u(t^n_k, X^{x,0}_{t^n_k}), \mathbb{E}\big(u(t^n_{k+1}, X^{x,0}_{t^n_{k+1}})\,|\,X^{0,x}_{t^n_k}\big) - u(t^n_k, X^{x,0}_{t^n_k})\Big).$$

If, in a second step, we assume that the function u is $\mathcal{C}^{1,2}([0,T]\times\mathbb{R})$ and that $\nabla_x u$ has polynomial growth. Then we can apply Itô's Lemma between t^n_k and t^n_{k+1} to get

$$u(t^n_{k+1}, X^{x,0}_{t^n_{k+1}}) = u(t^n_k, X^{x,0}_{t^n_k}) + \int_{t^n_k}^{t^n_{k+1}}\left(\frac{\partial u}{\partial t} + Lu\right)(s, X^{x,0}_s)ds + \int_{t^n_k}^{t^n_{k+1}}\big(\nabla_x u\,|\,\sigma\big)(s, X^{x,0}_s)dW_s$$

where the stochastic integral with respect to W is a true martingale since σ has linear growth, $\nabla_x u$ has polynomial growth and $\big\|\sup_{t\in[0,T]}|X^{x,0}_t|\big\|_p < +\infty$ by (1.20). As a consequence,

$$\begin{aligned}\mathbb{E}\big(u(t^n_{k+1}, X^{x,0}_{t^n_{k+1}})\,|\,X_{t^n_k}\big) - u(t^n_k, X^{x,0}_{t^n_k}) &= \mathbb{E}\left(\int_{t^n_k}^{t^n_{k+1}}\left(\frac{\partial u}{\partial t} + Lu\right)(s, X^{x,0}_s)ds\right) \qquad \mathbb{P}\text{-}a.s.\\ &\approx \left(\frac{\partial u}{\partial t} + Lu\right)(t^n_k, X^{x,0}_{t^n_k})\frac{T}{n} \qquad \mathbb{P}\text{-}a.s.\end{aligned}$$

Finally, this leads to

$$u(T, X^{x,0}_{_T}) = f(T, X^{x,0}_{_T}),\;\; \max\left(f(X_{t^n_k}) - u(t^n_k, X^{x,0}_{t^n_k}), \left(\frac{\partial u}{\partial t} + Lu\right)(t^n_k, X^{x,0}_{t^n_k})\right)\approx 0 \qquad \mathbb{P}\text{-}a.s.$$

Letting $n\to+\infty$, one may reasonably guess, switching back to continuous time that, $\mathbb{P}$-*a.s.*,

$$u(T, X^{x,0}_{_T}) = f(T, X^{x,0}_{_T}),\;\; \max\left(f(X_t) - u(t, X^{x,0}_t), \left(\frac{\partial u}{\partial t} + Lu\right)(t, X^{x,0}_t)\right) = 0.$$

If the support of $X^{x,0}_t$ is the whole space $\mathbb{R}^d$ for every $t\in(0,T]$, this suggests that u is a solution to the above variational inequality (1.32).

The simplest non trivial setting in which the above heuristic reasoning can be made rigorous is when $X^{x,0}_t = x + W_t$, relying on distribution theory and extended versions of Itô's Lemma (see [44] for a rigorous variational approach using distribution theory).

In practice, this approach is usually used in a backward way as a verification theorem: starting from a solution u to (1.32) (possibly in a viscosity sense) with enough regularity, one checks that $(u(t, X^{x,0}_t))_{t\in[0,T]}$ is the $\mathbb{P}$-Snell envelope of the payoff/obstacle process $Z_t = f(t, X^{x,0}_t)$ (see again [44] and [6]).

▷ *Numerical aspects.* From a numerical point of view, this connection suggests numerical methods to solve (1.32), at least when deterministic schemes become inefficient due to the *curse of dimensionality*: any probabilistic method devised to compute the *réduite* of a discrete time optimal stopping problem with a Markov structure process can be viewed as a way to solve (1.32). Having in mind that, in these probabilistic methods, the discrete time Markov process needs to be simulated at a moment or another, we cannot rely on a sample $(X_{t_k^n})_{0\le k\le n}$ of the diffusion process itself at least when $d \ge 2$ as mentioned beforehand. A noticeable exception, of course, is the case of diffusion processes of the form $X_t^{x,0} = \varphi(t,x,W_t)$ like *e.g.* the multi-dimensional correlated Black-Scholes model.

Consequently, as developed above, for numerical purposes, we do need to introduce as above the discrete time Euler scheme $\bar X^{n,x,0} = (\bar X^{n,x,0}_{t_k^n})_{0\le k\le n}$ with step $\frac{T}{n}$ (with obvious notations) of the diffusion $(X_t^{x,0})_{t\in[0,T]}$ and the Snell envelope $\bar U^n = (\bar U_{t_k^n})_{0\le k\le n}$ related to the payoff/obstacle process $\bar Z^n_{t_k^n} = f(t_k^n, \bar X^{n,x,0}_{t_k^n})$, $k = 0,\dots,n$.

All these successive approximations of the "true" original Snell envelope have been investigated in full details in [3], depending on the regularity of the payoff function and of the coefficients b and σ (or $a = \sigma\sigma^*$).

Note that for the sake of simplicity, we did not take into account a zero order term in (1.32). This is done in the cited literature and corresponds to a discounting factor appearing in the payoff/obstacle process.

The resulting convergence rates are of the form $O(n^{-\frac{1}{2}})$ between $(\bar U_{t_k^n})_{0\le k\le n}$ and $(U^n_{t_k^n})_{0\le k\le n}$ but, if f is quasi-convexity in the space variable x uniformly in $t\in[0,T]$, it is $O(n^{-1})$ this time between $(U^n_{t_k^n})_{0\le k\le n}$ and $(U_{t_k^n})_{0\le k\le n}$. These rates holds for various type of convergence (on compacts set with respect to x or in $L^p(\mathbb{P})$).

Conversely, we must of course have in mind that in low dimension (say $d\le 2$ or sometimes 3), the numerical strategy is the exact converse of what precedes: one relates the optimal stopping theorem to the parabolic variational inequality and implement a deterministic solver based on finite difference or finite elements/volume methods. So is the case for the seminal paper on the computation of American Put option in a Black-Scholes model, see [39] (see also [76] for exchange options). For an overview of *PDE* methods for option pricing, we refer to [1].

2. Optimal vector quantization

As emphasized in what precedes, whatever the application fields are, we need to have "good" and, if possible, "optimal", quantization grids at various levels N. This section is devoted to the existence of optimal quantizers (*via* the proof of Proposition 1.1(b)) and their properties as well as the numerical methods to compute them.

2.1. Existence of optimal quantization grids (proof of Proposition 1.1(a)-(b))

We will proceed by induction on the level N. First note that, if $p\ge 1$, for every level $N\ge 1$, the L^p-mean quantization error function defined on $(\mathbb{R}^d)^N$ by

$$(x_1,\dots,x_{_N}) \longmapsto \Big\| \min_{1\le i\le N} |X - x_i| \Big\|_{L^p(\mathbb{P})}$$

is clearly 1-Lipschitz continuous with respect to the ℓ^∞-norm on $(\mathbb{R}^d)^N$ defined by $|(x_1,\dots,x_{_N})|_{\ell^\infty} := \max_{1\le i\le N} |x_i|$. This is a straightforward consequence of Minkowski's inequality combined with the more elementary inequality $|\min_{1\le i\le N} a_i - \min_{1\le i\le N} b_i| \le \max_{1\le i\le N}|a_i - b_i|$. As a consequence, it implies the continuity of its p^{th} power, the L^p-distortion function $\mathcal{G}_{p,N}$.

If $p\in(0,1]$, one shows directly that $\mathcal{G}_{p,N}$ is p-Hölder continuous.

$\triangleright$ $N=1$. The non-negative continuous function $\mathcal{G}_{p,1}$ clearly goes to $+\infty$ as $|x_1|\to+\infty$. Hence, $\mathcal{G}_{p,1}$attains a minimum at a so-called L^p-median $x_1^{(1)}$ (which is clearly unique when $p>1$ by a strict convexity argument). So $\{x_1^{(1)}\}$ is an optimal quantization grid at level 1.

$\triangleright$ $N \Rightarrow N+1$. Assume there exists $x^{(N)} \in (\mathbb{R}^d)^N$ such that $\mathcal{G}_{p,N}(x^{(N)}) = \min_{(\mathbb{R}^d)^N}\mathcal{G}_{p,N}$. Set $\Gamma^{(N)} = \{x_i^{(N)},\ i=1,\dots,N\}$. Then, either $\mathrm{supp}(\mu)\setminus\Gamma^{(N)} = \emptyset$ and any $(N+1)$-tuple of $(\mathbb{R}^d)^{N+1}$ which "exhausts" the grid $\Gamma^{(N)}$ makes the function $\mathcal{G}_{p,N+1}$ equal to 0, its lowest possible value, or there exists $\xi_{N+1}\in\mathrm{supp}(\mu)\setminus\Gamma^{(N)}$.

In this second case, let $\Gamma^* = \Gamma^{(N)} \cup \{\xi_{N+1}\}$ and let $(C_i(\Gamma^*))_{1\le i\le N+1}$ be a Voronoi partition of Γ^* where $\overset{\circ}{C}_{N+1}(\Gamma^*)$ is the Voronoi cell of ξ_{N+1}. As $\xi_{N+1}\notin\Gamma^{(N)}$, it is clear that $\overset{\circ}{C}_{N+1}(\Gamma^*)\neq\emptyset$ and that $|X-\xi_{N+1}| < \min_{1\le i\le N}|X-x_i^{(N)}|$ on the interior of this cell. Furthermore, $\mathbb{P}(X\in C_{N+1}(\Gamma^*)) = \mu\big(\overset{\circ}{C}_{N+1}(\Gamma^*)\big) > 0$ since $\xi_{N+1}\in\overset{\circ}{C}_{N+1}(\Gamma^*)$ and $\xi_{N+1}\in\mathrm{supp}(\mu)$. Note that, everywhere on $(\mathbb{R}^d)^N$, one has $|X-\xi_{N+1}|\wedge\min\limits_{1\le i\le N}|X-x_i^{(N)}| \le \min\limits_{1\le i\le N}|X-x_i^{(N)}|$, so that, combining both inequalities yields

$$\lambda_{N+1} = \mathbb{E}\left(|X-\widehat{X}^{\Gamma^*}|^p\right) = \mathbb{E}\big(|X-\xi_{N+1}|^p\wedge\min_{1\le i\le N}|X-x_i^{(N)}|^p\big) < \mathbb{E}\big(\min_{1\le i\le N}|X-x_i^{(N)}|^p\big) = \mathcal{G}_{p,N}(x^{(N)} = e_{p,N}(X)^p.$$

In particular $\mathrm{card}(\Gamma^{(N)}) = N$, otherwise $\mathrm{card}(\Gamma^*)\le N$ which would contradict that $\mathcal{G}_{p,N}$ is minimum at $x^{(N)}$. Hence, the set $K_{N+1} = \Big\{x\in(\mathbb{R}^d)^{N+1} : \mathcal{G}_{p,N+1}(x)\le\lambda_{N+1}\Big\}$ is non-empty since it contains all the $(N+1)$-tuples which "exhaust" the elements of Γ^*. It is closed since $\mathcal{G}_{p,N+1}$ is continuous. In fact, we will show that it is also a bounded subset of $(\mathbb{R}^d)^{N+1}$. Let $x(k) = (x(k)_1,\ldots,x(k)_{N+1})$, $k\in\mathbb{N}$, be a K_{N+1}-valued sequence of $(N+1)$-tuples. Up to at most $N+1$ extractions, one may assume without loss of generality that there exists a subset $I\subset\{1,\ldots,N+1\}$ such that for every $i\in I$, $x_i(k)\to x_i(\infty)\in\mathbb{R}^d$ and for every $i\notin I$, $|x_i(k)|\to+\infty$. By a straightforward application of Fatou's Lemma

$$\liminf_k \mathcal{G}_{p,N+1}(x(k)) \ge \liminf_k \Big\|\min_{i\in I}|X-x_i(\infty)|\Big\|^p_{L^p(\mathbb{P})} \ge e_{p,|I|}(X)^p.$$

The sequence $\big(x(k)\big)_{k\in\mathbb{N}}$ being K_{N+1}-valued, one has $e_{p,|I|}(X)^p\le\lambda_{N+1} < e_{p,N}(X)^p$. In turn, this implies that $|I| = N+1$ *i.e.* the sequence of $(N+1)$-tuples $\big(x(k)\big)_{k\ge0}$ is bounded. As a consequence, the set K_{N+1} is compact and the function $\mathcal{G}_{p,N+1}$ attains a minimum over K_{N+1} at an $x^{(N+1)}$ which is obviously its absolute minimum and has pairwise components such that, with obvious notation, $\mathrm{card}(\Gamma^{(N+1)}) = N+1$ and $\mathbb{P}\big(X\in C_i(\Gamma^{(N+1)})\big) > 0$, $i = 1,\ldots,N+1$. $\square$

Remarks. • The proof of Claim (b) and (c) are as follows: if $N\ge2$ and $\mathbb{P}\big(X\in C_i(\Gamma^{(N)}\big) = 0$ for some index i, then $e_{p,N}(X) = e_{p,N-1}(X)$ which is impossible by the above proof if $\mathrm{supp}(\mu)$ has at least N points. Moreover, if $(z_N)_{N\ge1}$ is an everywhere dense sequence in $\mathbb{R}^d$, then $0\le e_{p,N}(X)\le e_p\big(\{z_1,\ldots,z_N\}\big)\to0$ as $N\to+\infty$ by the Lebesgue dominated convergence theorem.

• This existence result admits many extensions, in particular in infinite dimension when $\mathbb{R}^d$ is replaced by a separable Hilbert space or, more generally, a reflexive Banach space. It holds true for L^1-spaces as well (see [35] or Section 2.3.2 for a brief introduction to *functional quantization*).

• The set $\mathrm{argmin}\,\mathcal{G}_{p,N}$ is never reduced to a single N-tuple when $N\ge2$, simply because $\mathrm{argmin}\,\mathcal{G}_{p,N}$ is left stable under the action of the $N!$ permutations of $\{1,\ldots,N\}$. Even from a geometrical viewpoint, as soon as $d\ge2$, uniqueness of the resulting optimal quantization *grid* usually fails, *e.g.* because of the various transformations that may leave the distribution μ of X invariant. Thus, the normal distribution $\mathcal{N}(0;I_d)$ is invariant under all orthogonal transforms. But there are also examples (see [33]) for which optimal grids at level N do not even make up a "connected" set.

However, in 1-dimension, it has been proved (see *e.g.* [40]) that, as soon as μ is absolutely continuous with a log-concave density, there exists exactly one optimal quantization grid at level N. This grid has full size N and is characterized by its stationarity (see the next section) so that $\mathrm{argmin}\,\mathcal{G}_{p,N}$ is made of the $N!$ resulting N-tuples. Such distributions are called *strictly unimodal*.

2.2. Stationary quantization grid

2.2.1. *Definition*

Definition 2.1. *Let $X \in L^1_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ with distribution μ. A (quantization) grid $\Gamma = \{x_1, \ldots, x_{_N}\} \subset \mathbb{R}^d$ of size N is* μ-stationary *if the following two conditions are satisfied*

$$\begin{cases} (i)\ \mu\text{-negligibility of the boundary of the Voronoi diagram}: \quad \mu\Big(\bigcup_{1\le i\le N} \partial C_i(\Gamma)\Big) = 0. \\ (ii)\ \text{Self-consistency of the centroids}: \forall\, i \in \{1,\ldots,N\},\ x_i = \dfrac{\int_{C_i(\Gamma)} \xi \mu(d\xi)}{\mu(C_i(\Gamma))} = \mathbb{E}\big(X \,|\, X \in C_i(\Gamma)\big) \ \text{if}\ \mu(C_i(\Gamma)) \neq 0. \end{cases}$$

Note that item (ii) also reads

$$\widehat{X}^{\Gamma} = \mathbb{E}\big(X \,|\, \widehat{X}^{\Gamma}\big). \tag{2.35}$$

We will see further on in Section 3.1 (see Corollary 3.1 and the comments that follow) that L^2-*optimal grids at level N of full size N are always stationary*: they always satisfy condition (i) (see Theorem 4.3, p.38, in [33]) so that the related N-tuples turn out to be critical points of the quadratic distortion function $\mathcal{G}_{2,N}$, where which is differentiable at such points. All such critical points induce stationary grids which satisfy condition (ii) (see Proposition 3.1 in Section 3 further on).

Note that almost all numerical algorithms devised to search for (at least locally) optimal quantizers at a given level N are based on this stationary property (see Section 3 devoted to algorithmic aspects).

Remarks. • The above μ-stationarity property is also known in the statistical literature as the *self-consistency* property and such quantizers as *self-consistent prototypes* (see [75]). Moreover, this also gave rise to a new terminology (borrowed from statistics) to emphasize that each x_i is the μ-centre of mass of its Voronoi cell: one speaks of the points x_i as the *centroids* of their Voronoi cells.

• Stationarity is a key property for both applications and numerical methods t compute quadratic grids algorithms. Unfortunately, it is satisfied only by few quantizers. A new notion of quantization, called *dual quantization* has been recently developed (see [64] for a theoretical introduction and [63, 65] for applications to Numerical Probability) in which a reverse stationarity property is satisfied by all dual quantizers. Typically for dual quantization, one has

$$X = \mathbb{E}\big(\widehat{X}^{\Gamma,dual} \,|\, X\big).$$

This dual quantization requires the introduction of an exogenous noise and relies on the Delaunay triangulation rather than on the Voronoi diagram (hence its name "dual").

2.2.2. *Application to quantization based cubature formulas (II): second order*

Proposition 2.1. *Let Γ be a μ-stationary quantizer and let $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^d$ be a μ-distributed random vector such that $\mathbb{E}\,|X|^{1+\alpha} < +\infty$. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{C}^1$ function with an α-Hölder gradient ∇F. We denote by $[\nabla F]_\alpha$ its α-Hölder ratio. Then*

$$\big\|E\big(F(X) \,|\, \widehat{X}^{\Gamma}\big) - F(\widehat{X}^{\Gamma})\|_1 \le [\nabla F]_\alpha \big\|X - \widehat{X}^{\Gamma}\big\|^{1+\alpha}_{L^{1+\alpha}(\mathbb{P})}$$

so that, by taking expectation,

$$\big|\mathbb{E}\,F(X) - \mathbb{E}\,F(\widehat{X}^{\Gamma})\big| \le [\nabla F]_\alpha \big\|X - \widehat{X}^{\Gamma}\big\|^{1+\alpha}_{L^{1+\alpha}(\mathbb{P})}. \tag{2.36}$$

In particular, if ∇F is Lipschitz continuous

$$\big|\mathbb{E}\,F(X) - \mathbb{E}\,F(\widehat{X}^{\Gamma})\big| \le [\nabla F]_{\rm Lip} \big\|X - \widehat{X}^{\Gamma}\big\|^{2}_{L^{2}(\mathbb{P})}.$$

Remark. Since we know from Claim (c) in Proposition 1.1 that $e_{2,N}(X)\downarrow 0$ as $N\to+\infty$, it is clear that if we consider optimal quadratic (hence stationary) grids $\Gamma^{(N)}$ satisfying $e_{2,N}(X)=\big\|X-\widehat{X}^{\Gamma^{(N)}}\big\|_{L^2(\mathbb{P})}\to 0$ which implies that $\big\|X-\widehat{X}^{\Gamma^{(N)}}\big\|^{1+\alpha}_{L^2(\mathbb{P})}=o\big(\big\|X-\widehat{X}^{\Gamma^{(N)}}\big\|_{L^2(\mathbb{P})}\big)$ as $N\to+\infty$. So, the error bounds in the above cubature formulas become infinitely smaller than those obtained in Section 1.3.2 when N is large enough. Moreover, it is proved in Section 2.3 below (Theorem 2.1 known as Zador's Theorem) that a sharp convergence rate of decay $e_{2,N}(X)\sim c_X N^{-\frac{1}{d}}$ holds.

Proof. It follows from a second order Taylor expansion that

$$\big|F(X)-F(\widehat{X}^{\Gamma})-\big(\nabla F(\widehat{X}^{\Gamma})|X-\widehat{X}^{\Gamma}\big)\big|\le[\nabla F]_\alpha\big|X-\widehat{X}^{\Gamma}\big|^{1+\alpha}.$$

The stationarity property reads $\widehat{X}^{\Gamma}=\mathbb{E}(X\,|\,\widehat{X}^{\Gamma})$ and, as $\nabla F(\widehat{X}^{\Gamma})$ is $\sigma(\widehat{X}^{\Gamma})$-measurable, we obtain

$$\mathbb{E}\Big(\big(\nabla F(\widehat{X}^{\Gamma})|X-\widehat{X}^{\Gamma}\big)\,|\,\widehat{X}^{\Gamma}\Big)=\Big(\nabla F(\widehat{X}^{\Gamma})\big|\underbrace{\mathbb{E}\big(X-\widehat{X}^{\Gamma}\,|\,\widehat{X}^{\Gamma}\big)}_{=0}\Big)=0. \qquad \square$$

2.2.3. *Application to convex functions*

If $F:\mathbb{R}^d\to\mathbb{R}$ is convex and $X,\ F(X)\in L^1(\mathbb{P})$, then for every stationary quantization grid Γ, one has

$$\mathbb{E}\,F(\widehat{X}^{\Gamma})=\mathbb{E}\,F\Big(\mathbb{E}\big(X\,|\,\widehat{X}^{\Gamma}\big)\Big)\le\mathbb{E}\Big(\mathbb{E}\big(F(X)\,|\,\widehat{X}^{\Gamma}\big)\Big)=\mathbb{E}\,F(X).$$

This shows that quantization based cubature formulas provide lower bounds for $\mathbb{E}\,F(X)$ when F is convex. Note that one has the reverse inequality by the *dual quantization* based cubature formula

$$\mathbb{E}\,F(X)\ge\mathbb{E}F\big(\widehat{X}^{\Gamma,dual}\big)$$

since it holds for *for any grid* (see [64]). These provides a two-sided bound can be numerically implemented.

2.3. Optimal quantization rate

2.3.1. *Rate results*

This rate is ruled by two results: the first one is asymptotic and is known as Zador's Theorem (see [33]), the second one is non-asymptotic and, in some way, universal.

Theorem 2.1. (a) Zador's Theorem (asymptotic rate, see [33]). *Let $X\in L^{p+\eta}_{\mathbb{R}^d}(\mathbb{P})$ for some $\eta>0$, with distribution $\mathbb{P}_X=\mu$. Let $|\,.\,|$ denote (here) any norm on $\mathbb{R}^d$. We define $\varphi=\frac{d\mu}{d\lambda_d}$ (where λ_d denotes the Lebesgue measure on $(\mathbb{R}^d,\mathcal{B}or(\mathbb{R}^d))$) the density of the non-singular part of μ. Then*

$$\lim_N N^{\frac{1}{d}}e_{p,N}(X)=\widetilde{J}_{p,d}\|\varphi\|^{\frac{1}{p}}_{L^{\frac{p}{p+d}}(\lambda_d)}$$

where $\widetilde{J}_{p,d}=\inf_{N\ge1}N^{\frac{1}{d}}e_{p,N}\big(U([0,1]^d)\big)\in(0,+\infty)$ depends on the current norm. When this norm is the canonical Euclidean norm on $\mathbb{R}^d$, one has $\widetilde{J}_{1,2}=\sqrt{\frac{2+3\log\sqrt{3}}{3^{\frac{7}{4}}\sqrt{2}}}$, $\widetilde{J}_{2,2}=\sqrt{\frac{5}{18\sqrt{3}}}$ and ([5])

$$\widetilde{J}_{p,d}\sim\left(\frac{d}{2\pi e}\right)^{\frac{1}{2}}\qquad as\quad d\to+\infty. \tag{2.37}$$

[5] where $a_k\sim b_k$ means that $a_k=u_kb_k$ with $\lim_k u_k=1$.

(b) Extended Pierce's Lemma (non-asymptotic rate, see [49]). *Let $p, \eta > 0$. There exists a real constant $\kappa_{d,p,\eta} \in (0,+\infty)$ such that, for every $\mathbb{R}^d$-valued random vector $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^d$,*

$$\forall\, N \geq 1, \quad e_{p,N}(X) \leq \kappa_{d,p,\eta}\, \sigma_{p+\eta}(X) N^{-\frac{1}{d}} \tag{2.38}$$

where, for every $r \in (0,+\infty)$, $\sigma_r(X) = \inf_{a\in\mathbb{R}^d} \|X-a\|_{L^r(\mathbb{P})}$ denotes the L^r-standard deviation of X.

Claim (a), known as Zador's Theorem, was first established in [77] for the uniform distribution on the unit hypercube $[0,1]^d$, using a self-similarity argument "à la Hammersley". The extension to more general distributions is due to Bucklew and Wise in [15], with a gap in the proof for non-compactly supported distributions. This gap was filled in [33], Chapter II, Theorem 6.2 (p.78).

When μ is purely singular, the above result remains true but the above asymptotics is not the right one since the limit in this normalization is 0. Other asymptotics can emerge *e.g.* for (uniform) distributions on fractal sets, leading to the definition of a *quantization dimension* (see again [33], Chapter III, for an introduction to quantization on fractal sets).

2.3.2. *Additional results and first applications*

▷ *Square grids for $U([0,1]^d)$.* A natural question is to compare optimal quantization methods with "standard" square grid methods on $[0,1]^d$ when $N = m^d$, $m \in \mathbb{N}^*$ (these values of N are the only ones for which such square grids exist). Let

$$\Gamma_d(m) = \prod_{\ell=1}^{d} \left(\frac{2i_\ell - 1}{2m}\right)_{1\leq i_\ell \leq m}$$

be a hyper-cubic "product quantization" grid. When $p = 2$, this grid induces, owing to Pythagorus' Theorem and Formula (1.4) for the uniform distribution $U([0,1])$, a mean quadratic quantization error given by

$$e_{2,N}\big(\Gamma_d(m), U([0,1]^d)\big)^2 = d\left(\frac{1}{2\sqrt{3}\,m}\right)^2 = \frac{d}{12m^2},$$

so that

$$e_{2,N}\big(\Gamma_d(m), U([0,1]^d)\big) = \sqrt{\frac{d}{12}}\, N^{-\frac{1}{d}}.$$

If we consider that d is large enough so that, by (2.37), $\widetilde{J}_{p,d} \approx \big(\frac{d}{2\pi e}\big)^{\frac{1}{2}}$, one derives by the above Zador Theorem,

$$e_{2,N}\big(U([0,1]^d)\big) \approx \sqrt{\frac{d}{2\pi e}} N^{-\frac{1}{d}} \approx \sqrt{\frac{d}{17.0795}}\, N^{-\frac{1}{d}}.$$

This shows that, *as d grows to infinity*, the ratio between true optimal and optimal product quantization remains (asymptotically) bounded and, in fact, very close to 1 since $\sqrt{\frac{2\pi e}{12}} \approx 1.1930\ldots$! But the main difference between these two points of view is that optimal quantization grids *do exist for every size/level* N whereas product grids become increasingly sparse in higher dimensions.

▷ *Empirical measure theorem.* If $\Gamma^{(N)} = \{x_1^{(N)}, \ldots, x_N^{(N)}\}$, $N \geq 1$, denotes a sequence of L^p-optimal N-quantizers of the distribution μ, we saw in Equation (1.13) that the weighted distribution $\hat{\mu}_N$ of $\widehat{X}^{\Gamma^{(N)}}$ weakly converges to the original distribution μ of X. But what about the *regular* empirical measures

$$\widetilde{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i^{(N)}},\ N \geq 1?$$

The answer is provided by the following *empirical measure theorem* established in [33] (see Theorem 7.5, p.96) and [22].

Theorem 2.2. *If $X \in L^{p+\eta}(\mathbb{P})$ for some $\eta > 0$, then, with the notations of Zador's Theorem, if $\varphi \not\equiv 0$,*

$$\widetilde{\mu}_{N} \xrightarrow{(w)} \widetilde{\mu} = \frac{\varphi^{\frac{d}{d+p}}}{\int_{\mathbb{R}^d} \varphi^{\frac{d}{d+p}} d\lambda_d}.\lambda_d \quad \textit{as} \quad N \to +\infty$$

where $\xrightarrow{(w)}$ denotes weal convergence of probability measures.

Remark. In fact the theorem holds true for (sub-)sequences of *asymptotically optimal quantizers* in the sense that they satisfy Zador sharp rate.

▷ *A brief look back at Information Theory.* Let $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^d$ be a (static) square integrable random vector/signal X with a non-singular distribution μ. We consider an optimal quadratic quantization grid $\Gamma^{(N)}$ of size N for μ. Let $\hat{\mu}_{N}$ denote the distribution of its optimal quadratic quantization $\widehat{X}^{\Gamma^{(N)}}$ We saw in Section 1.2 that the mean transmission cost $\kappa_{\hat{\mu}_{N}}$, when coded by an appropriate instantaneous code, satisfies the two-sided inequality (1.6)-(1.7) involving the dyadic entropy of $H_2(\hat{\mu}_{N})$ of $\hat{\mu}_{N}$.

▷ *A universal asymptotic bound.* Let $w_i^{(N)} = \hat{\mu}_{N}\big(x_i^{(N)}\big)$, $i = 1, \ldots, N$, be the weights attached to the distribution $\hat{\mu}_{N}$. We know, by classical optimization arguments, that

$$H_2(\hat{\mu}_{N}) = -\sum_{i=1}^{N} w_i^{(N)} \log_2 w_i^{(N)} \leq \log_2 N$$

so that we derives that $2^{\kappa_{\hat{\mu}_N}-1} \leq N \leq 2^{\kappa_{\hat{\mu}_N}}$. Now, let $\kappa > 0$ be a prescribed mean transmission cost. Plugging this universal bound in (2.1) with $N = \lceil 2^{\kappa-1} \rceil$, we derive that the quadratic mean transmission error $\big\|X - \widehat{X}^{\Gamma_N}\big\|_{L^2(\mathbb{P})}$ satisfies

$$L^2\text{-mean transmission error}(\kappa) \precsim \widetilde{J}_{2,d} \, \big\|\varphi\big\|^{\frac{1}{2}}_{L^{\frac{d}{d+2}}(\lambda_d)} 2^{-\frac{\kappa-1}{d}} \quad \text{as} \quad \kappa \to +\infty.$$

In particular, expressed in terms of quantization level, we get this coarser equivalent

$$-\log_2\Big(L^2\text{-mean transmission error}(N)\Big) \sim \frac{\kappa(N)}{d} \quad \text{as} \quad N \to +\infty.$$

This provides a more precise "flavor" of Shannon's Source Coding Theorem (for a general formulation and a proof in a more general framework, we refer to [23]).

▷*A sharper, but heuristic, bound.* It is commonly shared in the Information Theory community (see [30]) that an absolutely distribution μ with a continuous bounded density $\varphi > 0$ satisfies ($\varphi \log \varphi$ is bounded and)

$$w_i^{(N)} \sim c_{N} \frac{\varphi^{\frac{d}{d+2}}(x_i^{(N)})}{N} \quad \text{as} \quad N \to +\infty$$

uniformly over the elementary quantizers $x_i^{(N)}$ lying in a fixed compact set, where c_N is a normalizing constant. Partial results in that direction have been recently established in [37] for a wide class of absolutely continuous distributions (including among many others normal distributions, gamma distributions, hyper-exponential distributions).

It follows from the empirical measure theorem (2.2) that

$$\lim_{N} c_N = \left(\int_{\mathbb{R}^d} \varphi^{\frac{d}{d+2}}(\xi) d\xi\right)^{-1}.$$

If we assume a sharper result, namely $w_i^{(N)} - c_{_N} \frac{\varphi^{\frac{d}{d+2}}(x_i^{(N)})}{N} \to 0$ as $N \to +\infty$, uniformly when $x_i^{(N)}$ lies a compact set, elementary computations based on repeated applications of Theorem 2.2 (empirical mean theorem) yield

$$C_\mu = -\lim_N \Big(H_2(\hat{\mu}_{_N}) - \log_2 N\Big) = \log\left(\int_{\mathbb{R}^d} \varphi^{\frac{2}{d+2}} d\lambda_d\right) + \frac{d}{d+2}\int_{\mathbb{R}^d} \varphi \log_2 \varphi \, d\lambda_d.$$

Noting that $\int_{\mathbb{R}^d} \varphi^{\frac{2}{d+2}} d\lambda_d = \int_{\mathbb{R}^d} \varphi^{-\frac{d}{d+2}} d\mu$ and applying Jensen's Inequality to the distribution μ and the concave function log imply that $C_\mu > 0$. This is in accordance with the fact that the uniform distribution $\mu_{_N}^{Unif}$ over $\{1, \ldots, N\}$ or any set of size N has the highest possible entropy equal to $\log_2 N$. This finally yields the slightly sharper bound for the quadratic mean signal transmission error

$$L^2\text{-mean transmission error}(\kappa) \precsim \widetilde{J}_{2,d} \, \|\varphi\|^{\frac{1}{2}}_{L^{\frac{d}{d+2}}(\lambda_d)} 2^{\frac{1-C_\mu}{d}} 2^{-\frac{\kappa}{d}} \quad \text{as} \quad \kappa \to +\infty.$$

▷ *From vector to functional quantization.* Most issues raised in the former sections can be formulated in an infinite dimensional setting where $\mathbb{R}^d$ is replaced by a separable Hilbert space or more generally a (reflexive or not) separable Banach space $(E, |\,.\,|_{_E})$. From a more probabilistic point of view E is often a functional space like $L^2_{\mathbb{R}}([0,T], \lambda)$ or $\mathcal{C}([0,T], \mathbb{R})$. This corresponds to the path space of a bi-measurable stochastic process $(X_t)_{t\in[0,T]}$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ so that for every $\omega \in \Omega$, $X(\omega) = (t \mapsto X_t(\omega)) \in E$: viewing things that way makes X an E-valued (infinitely dimensional) random vector. For this reason, the quantization problem in infinite dimension is often known as *functional quantization*. Thus, existence of optimal quantizers can be established in the case of reflexive Banach spaces like $L^p_{\mathbb{R}}([0,T], dt)$, $1 < p < +\infty$, following the proof of Proposition 1.1, using this time a *weak continuity lower semi-continuity* argument (see *e.g.* [21,35,47]). It also holds true for the $L^1_{\mathbb{R}}(dt)$ space, but not in full generality for the space $\mathcal{C}([0,T], \mathbb{R})$.

However, no such general rate result as Zador's Theorem is available, but optimal quantization rates, sometimes sharp, can be established in this framework for various classes of stochastic processes, including the (fractional) Brownian motion and the Brownian bridge, scalar Brownian diffusions processes, etc. We refer to [24,35,47–49] and the references therein for results in that direction. It turns out that in many situations these convergence rates for optimal functional quantization error "live" in a $(\log N)^{-r}$ scale ($r = \frac{1}{2}$ for the standard Brownian motion and bridge, diffusion processes, $r = H$ for the fractional Brownian motion with Hurst constant H, $r = m + \frac{1}{2}$ for the m-fold integrated Brownian motion, etc).

More generally, when dealing with Gaussian processes this quantization rate is closely related to the rate of decay of the eigenvalues of its covariance operator (Karhunen-Loève eigensystem). For general stochastic processes, a connection has been made between the mean $L^p([0T), dt)$-functional quantization rate and the regularity of the mapping $t \mapsto X_t$ from $[0,T]$ to $L^p([0,T], dt)$ (see [49]).

Numerical applications of functional quantization have also been devised, based on optimized grids. Packages of such grids available on the website `www.quantize.maths-fi.com`, see Section 3.5. They mostly rely on cubature formulas which can be straightforwardly extended to Hilbert or Banach spaces. See *e.g.* [59] for applications to the pricing and hedging of path-dependent options. Let us cite as well applications to variance reduction by universal stratified sampling (see [18]). But we will not go further in that direction in this paper.

2.3.3. *Numerical integration (III): rates*

Assume $X \in L^2_{\mathbb{R}^d}(\mathbb{P})$ (still with a distribution μ). The above results show that, provided one has access to quadratic optimal quantization grids $\Gamma_{_N}$ for various levels/sizes, it is possible to produce cubature formulas with respect to μ which have the following properties.

▷ *α-Hölder functions, $\alpha \in (0,1]$.* An α-Hölder function $F : \mathbb{R}^d \to \mathbb{R}$ can be integrated using the N-point cubature formula (1.11) induced by a grid Γ with an accuracy upper-bounded by $\|X - \widehat{X}^\Gamma\|_{L^1(\mathbb{P})}$ owing to (1.13).

Let $\Gamma^{(N)}$ be an optimal quadratic quantization grid at level N for the distribution μ and that $X\in L^{2+\eta}_{\mathbb{R}^d}(\mathbb{P})$ for some $\eta>0$. We use that $\|X-\widehat{X}^{\Gamma^{(N)}}\|_{L^1(\mathbb{P})}\le\|X-\widehat{X}^{\Gamma^{(N)}}\|_{L^2(\mathbb{P})}=e_{2,N}(\mu)\le C_{d,\,\mu}N^{-\frac{1}{d}}$ where $C_{d,\,\mu}$ is a real constant only depending on the dimension d and on the $(2+\eta)$-pseudo standard deviation of μ given by (2.38) in Theorem 2.1(b).

▷ *Locally α-Hölder functions.* Let us come back to the general error bound (1.12) for α-Höder functions with $\alpha+\beta$-polynomial growth at infinity. If $\alpha+\beta\ge1$ then $\frac{\beta}{1-\alpha}\ge1$ so that, combining stationarity and Jensen's inequality, we get

$$\|\widehat{X}^{\Gamma^{(N)}}\|_{L^{\frac{\beta}{1-\alpha}}(\mathbb{P})}=\big\|\mathbb{E}\big(X\,|\,\widehat{X}^{\Gamma^{(N)}}\big)\big\|_{L^{\frac{\beta}{1-\alpha}}(\mathbb{P})}\le\big\|X\big\|_{L^{\frac{\beta}{1-\alpha}}(\mathbb{P})}$$

which makes the right hand side of (1.12) only depending on (L^r-norms of) X, namely

$$\big|\mathbb{E}\,F(X)-\mathbb{E}\,F(\widehat{X}^{\Gamma^{(N)}})\big|\le[F]_{\alpha,\beta}C_{d,\mu}\left(1+2\|X\|^{\beta}_{L^{\frac{\beta}{1-\alpha}}(\mathbb{P})}\right)N^{-\frac{\alpha}{d}}.$$

▷ *Functions with α-Hölder gradient.* If F is continuously differentiable on $\mathbb{R}^d$ with an α-Hölder gradient $\alpha\in(0,1]$ and $X\in L^{2+\eta}_{\mathbb{R}^d}(\mathbb{P})$, note that in the cubature formula (2.36), one has

$$\|X-\widehat{X}^{\Gamma^{(N)}}\|_{L^{1+\alpha}(\mathbb{P})}\le\|X-\widehat{X}^{\Gamma^{(N)}}\|_{L^2(\mathbb{P})}=e_{2,N}(\mu)=O\big(N^{-\frac{1}{d}}\big)$$

since the grid $\Gamma^{(N)}$ is stationary (*i.e.* satisfies (2.35)) so that

$$\big|\mathbb{E}\,F(X)-\mathbb{E}\,F(\widehat{X}^{\Gamma_N})\big|\le C_{d,\mu}[\nabla F]_{\alpha}N^{-\frac{1+\alpha}{d}}. \tag{2.39}$$

This last result somewhat extends the one established in [52] for compactly supported absolutely continuous distributions μ on $\mathbb{R}^d$.

▷ *Numerical tests.* Various numerical tests have been carried out on these cubature formulas in [59] and in the survey [62]. They include very performing results obtained using a spatial Richardson-Romberg extrapolation, not developed here, which consists in a linear combination of several quantization based cubature formulas with appropriate sizes.

2.3.4. *Application to optimal stopping problems: design of a quantization tree*

Assume that the characteristics of the *quantization tree* $(\Gamma_k,w^k)_{0\le k\le n}$ related to the Markov chain $(X_k)_{0\le k\le n}$ is made up with grids such that $\Gamma_k=\Gamma_k^{(N_k)}$ of size N_k is L^2-optimal for the marginal distribution μ_k of X_k at every time $k\in\{0,\ldots,n\}$. Let

$$N=N_0+\cdots+N_n$$

denote the total number of points (or nodes) used in the quantization tree.

If one sets $N_k=\bar{N}:=\frac{N}{n+1}$, or $\frac{N-1}{n}$ if X_0 is deterministic so that $N_0=1$, it follows from the upper-bound (1.27) that resulting approximation provided by the quantization tree of the Snell envelope related to the optimal stopping problem (1.33) and their related *réduites* satisfy

$$\big|\mathbb{E}\,\widehat{u}_0(\widehat{X}_0^{\Gamma_0})-\mathbb{E}\,u_0(X_0)\big|\le\|\widehat{U}_0-U_0\|_{L^2(\mathbb{P})}\le C_X\sqrt{n}\bar{N}^{-\frac{1}{d}}.$$

In the special case where the Markov chain is the Euler scheme $(\bar{X}^n_{t^n_k})_{0\le k\le n}$ with step $\frac{T}{n}$ of a diffusion process with drift b and diffusion coefficient σ satisfying appropriate Lipschitz assumptions, then the constant $C_X=C_{[b]_{\rm Lip},[\sigma]_{\rm Lip},T}$ *does not depend on the time discretization parameter* n as shown beforehand in the error bound (1.31). In particular, if $X_0=x_0\in\mathbb{R}^d$ (with the obvious notations u^n and $\widehat{u}^n$) we have

$$\big|\widehat{u}_0^n(x_0)-u_0^n(x_0)\big|\le C_{b,\sigma,T}(N-1)^{-\frac{1}{d}}n^{\frac{1}{d}+\frac{1}{2}}.$$

Although we will not discuss this point here in details, we must have in mind that the time discretization error bounds established in [2] behave under a Lipschitz continuity assumption on b, σ and f, in x uniformly in $t\in[0,T]$, is $O(n^{-\alpha})$ with $\alpha=\frac{1}{2}$ when the discrete time approximating Markov chain is the Euler scheme. When the discrete time approximating Markov chain is the sampled process itself $(X_{t_k^n})_{0\le k\le n}$ and the function f is *quasi-convex* with linear growth in x uniformly in $t\in[0,T]$, the rate holds with $\alpha=1$ (this class of functions includes convex functions f with linear growth in x and functions having a Lipschitz continuous gradient $\nabla_x f$ in x uniformly in t in both cases).

In fact, especially in a diffusion framework, it is important for practical purposes to optimize the shape or the *design* of the quantization tree, *i.e.* the sizes N_k of the grids, subject to the above global constraint $\sum_{0\le k\le n} N_k = N$. The idea is to take advantage of the non asymptotic bound (2.38) to bound the mean quadratic quantization error of each marginal *i.e.*

$$\min_{N_0+\cdots+N_n=N}\sum_{k=0}^{n} N_k^{-\frac{2}{d}}\big\|\bar X^{n,x}_{t_k^n}\big\|^2_{L^{2+\eta}(\mathbb{P})}$$

subject to the additional constraint that all the N_ks are non-zero positive integers. In practice, $\big\|\bar X^{n,x}_{t_k^n}\big\|_{L^{2+\eta}(\mathbb{P})}$ is not known explicitly and one may replace this quantity by its classical upper bound, $\big\|\bar X^{n,x}_{t_k^n}\big\|_{2+\eta}\le e^{C_{b,\sigma,T}t_k^n}(1+\|X_0\|_{L^{2+\eta}(\mathbb{P})})$ which holds true as soon as both functions b and σ satisfy a linear growth assumption (in x, uniformly in $t\in[0,T]$): this follows from (1.20) after integrating the starting value with respect to the distribution of X_0 (see [11, 56] among others for details). This (approximately) leads to

$$N_k=\left\lfloor\frac{a_k^{\frac{2d}{d+2}}}{\sum_{\ell=0}^{n}a_\ell^{\frac{2d}{d+2}}}N\right\rfloor\vee 1$$

where a_k is a known upper-bound of $\big\|\bar X^{n,x}_{t_k^n}\big\|_{L^{2+\eta}(\mathbb{P})}$ for every $k=0,\ldots,n$. The resulting bound is of the form

$$\big|\widehat u_0^n(x_0)-u_0^n(x_0)\big|\le C_{b,\sigma,T}N^{-\frac{1}{d}}\left(\sum_{k=0}^{n}a_k^{\frac{2d}{d+2}}\right)^{\frac{1}{2}+\frac{1}{d}}.$$

▷ *Numerical tests*: Extensive numerical experiments on quantization based schemes for pricing American options, but also non-linear filtering, stochastic control have been carried out, including the introduction and analysis of a Richardson-Romberg extrapolation method which again dramatically improves the convergence rate. For more details we refer to [2–4, 65] (the last reference is devoted to both Voronoi and dual vector quantizations applied to the the pricing of American style derivatives) and [19, 62, 68] (for non-linear filtering, stochastic control applied to Finance) and the references therein.

3. How to get optimal (quadratic) quantization?

The foundation of almost every numerical method devoted to the computation of optimal quantizers is to establish the differentiability of the L^p-distortion function $\mathcal{G}_{p,N}$ (at level N). In what follows $\partial A=\bar A\setminus \overset{\circ}{A}$ denotes the boundary of $A\subset\mathbb{R}^d$.

3.1. Differentiability of the (quadratic) distortion function

Proposition 3.1. *Let $X\in L^2_{\mathbb{R}^d}(\mathbb{P})$ with distribution $\mathbb{P}_X=\mu$. If $x=(x_1,\ldots,x_N)\in(\mathbb{R}^d)^N$ has pairwise distinct components (i.e. $x_i\ne x_j$, $i\ne j$) and $\mathbb{P}\big(X\in\bigcup_{1\le i\le N}\partial C_i(x)\big)=0$ (with an obvious abuse of notation concerning* *the Voronoi cells), the L^2-distortion function $\mathcal{G}_{2,N}$ is differentiable at x and*

$$\nabla\mathcal{G}_{2,N}(x) = 2\left(\int_{C_i(x)}(x_i-\xi)\mu(d\xi)\right)_{1\le i\le N} = 2\Big(\mathbb{E}\mathbf{1}_{\{X\in C_i(x)\}}(x_i-X)\Big)_{1\le i\le N}. \tag{3.40}$$

Proof. First note that, as the N-tuple x has pairwise distinct components, all the interiors $\overset{\circ}{C}_i(x)$, $i=1,\ldots,N$, of the Voronoi cells induced by x are non-empty. For every $\xi\notin\bigcup_{1\le i\le N}\partial C_i(x)$, *i.e* $\mu(d\xi)$-*a.s.*, one has

$$\forall\, i\in\{1,\ldots,N\},\quad \frac{\partial}{\partial x_i}\Big(\min_{1\le j\le N}|x_j-\xi|^2\Big) = \mathbf{1}_{\{\xi\in\overset{\circ}{C}_i(x)\}}\frac{\partial|x_i-\xi|^2}{\partial x_i} = 2\,\mathbf{1}_{\{\xi\in\overset{\circ}{C}_i(x)\}}(x_i-\xi).$$

On the other hand, for every $x,x'\in(\mathbb{R}^d)^N$, the function $\mathcal{G}_{2,N}$ is locally Lipschitz continuous since

$$\begin{aligned}|\mathcal{G}_{2,N}(x')-\mathcal{G}_{2,N}(x)| &\le \int_{\mathbb{R}^d}\big|\min_{1\le j\le N}|x_j-\xi|-\min_{1\le j\le N}|x'_j-\xi|\big|\big(\min_{1\le j\le N}|x_j-\xi|+\min_{1\le j\le N}|x'_j-\xi|\big)\mu(d\xi)\\ &\le \max_{1\le j\le N}|x_j-x'_j|\int_{\mathbb{R}^d}\Big(\max_{1\le j\le N}\big(|x_j|+|\xi|\big)+\big(\min_{1\le j\le N}|x_j|+|\xi|\big)\Big)\mu(d\xi)\\ &\le C_\mu|x-x'|_\infty\Big(1+|x|_\infty+|x'|_\infty\Big).\end{aligned}$$

As a consequence, $\mathcal{G}_{2,N}$ is differentiable at x by the local interchange Lebesgue differentiation Theorem. □

Remarks. In fact, when $p>1$, the L^p-distortion function with respect to an Euclidean norm is also differentiable at N-tuple having pairwise distinct components with gradient

$$\nabla\mathcal{G}_{p,N}(x) = p\left(\int_{C_i(x)}\frac{x_i-\xi}{|x_i-\xi|}|x_i-\xi|^{p-1}\mu(d\xi)\right)_{1\le i\le N} = p\left(\mathbb{E}\Big(\mathbf{1}_{\{X\in C_i(x)\}}\frac{x_i-X}{|x_i-X|}|x_i-X|^{p-1}\Big)\right)_{1\le i\le N}.$$

An extension to the case $p\in(0,1]$ does exist under appropriate continuity and integrability assumptions on the distribution μ so that $\mu(\{a\})=0$ for every a and the function $a\mapsto\int_{\mathbb{R}^d}|\xi-a|^{p-1}\mu(d\xi)$ remains bounded on compact sets of $\mathbb{R}^d$. A more general differentiation result exists for *strictly convex smooth* norms (see Lemma 2.5, p.28 in [33]).

bigskip This leads to the following corollary which is the starting point of all numerical methods to compute optimal (or at least locally optimal) quantizers.

Corollary 3.1. *Let $X\in L^2_{\mathbb{R}^d}(\mathbb{P})$ with distribution μ such that* $\mathrm{card}\big(\mathrm{supp}(\mu)\big)\ge N$.

(a) Any grid $\Gamma^{(N)}$ attached to an N-tuple $x^{(N)}$ which minimizes the quadratic distortion function $\mathcal{G}_{2,N}$ is a stationary quantizer of full size N in the sense of Definition 2.1. ("attached" means $\Gamma^{(N)}=\{x_i^{(N)},\,1\le i\le N\}$.)

(b) Any grid Γ_N attached to a critical point $x^{(N)}$ of $\mathcal{G}_{2,N}$, hence having a Voronoi partition with μ-negligible boundary, is a stationary quantizer at level (at most) N.

Proof. (a) As $\mathrm{card}(\mathrm{supp}(\mu))\ge N$, any optimal grid has full size N by Proposition 1.1(b) and can be reduced, up to a permutation, to an N-tuple $x^{(N)}$. Moreover, owing to Theorems 4.1 and 4.2 in [33] (p.37-38), we know that $\mu\big(\bigcup_{1\le i\le N}\partial C_i(x)\big)=0$ and $\mu\big(C_i(x^{(N)})\big)>0$, $i=1,\ldots,N$, *i.e.* the boundaries of its Voronoi partition is μ-negligible. Then, one easily checks that $x^{(N)}$ is in fact a both global and local minimum of $\mathcal{G}_{2,N}$ and $\mathcal{G}_{2,N}$ is differentiable at $x^{(N)}$ so that $\nabla\mathcal{G}_{2,N}\big(x^{(N)}\big)=0$ *i.e.*

$$\forall\, i\in\{1,\ldots,N\},\quad \int_{C_i(x^{(N)})}(x_i-\xi)\mu(d\xi)=0.$$

Or, equivalently,

$$x_i^{(N)} = \frac{\int_{C_i(x^{(N)})} \xi\mu(d\xi)}{\mu(C_i(x^{(N)}))} = \mathbb{E}\big(X \mid X \in C_i(x^{(N)})\big)$$

which in turn can be rewritten

$$\widehat{X}^{\Gamma^{(N)}} = \mathbb{E}\big(X \mid \widehat{X}^{\Gamma^{(N)}}\big) \quad \text{where} \quad \Gamma^{(N)} = \Big\{x_i^{(N)},\, i = 1, \ldots, N\Big\}.$$

Claim (b) is obvious. □

3.2. Competitive Learning Vector Quantization

The Competitive Learning Vector Quantization ($CLVQ$) is a *stochastic gradient descent* deriving from the quadratic distortion $\mathcal{G}_{2,N}$, viewed as a potential function to be minimized. Let us start by its *mean field* version.

3.2.1. *The deterministic "batch" algorithm.*

First let us consider the classical – deterministic – gradient descent associated to $\mathcal{G}_{2,N}$, also known as "batch" $CLVQ$ to k-means. It is a recursive zero search procedure of its gradient $\nabla\mathcal{G}_{2,N}$ formally reading

$$x(k+1) = x(k) - \gamma_{k+1}\nabla\mathcal{G}_{2,N}\big(x(k)\big), \quad x(0) \in \Big(\mathrm{Hull}\big(\mathrm{supp}(\mu)\big)\Big)^N, \tag{3.41}$$

where $\mathrm{Hull}\big(\mathrm{supp}(\mu)\big)$ denotes the *closed* convex hull of the support of the distribution μ. The sequence $(\gamma_k)_{k\geq 1}$ is a sequence of positive *step parameters* satisfying the so-called *decreasing step assumptions* $\sum_{k\geq 1}\gamma_k = +\infty$ and $\sum_{k\geq 1}\gamma_k^2 < +\infty$. We will also assume in the present case that $\gamma_k \in (0,1)$, $k \geq 1$, for reasons to be explained further on. When $\nabla\mathcal{G}_{2,N}$ is itself differentiable (enough) one could replace this step γ_{k+1} by the inverse of the Hessian of $H\big(\mathcal{G}_{2,N}\big)$, leading to the classical Newton-Raphson procedure. It opens the door to many of its variants like Levenberg-Marquardt algorithm, etc.

The procedure (3.1) is well-defined if the distribution μ *assigns no mass to hyperplanes and* $x(0)$ *has pairwise distinct components*. Indeed, when these conditions are fulfilled, one shows that $x_i(k+1)$ lies in the interior of the Voronoi cell of $x_i(k)$ which prevents the components to get stuck as a result of an iteration. Thus, the existence of $\nabla\mathcal{G}_{2,N}$ for the next iteration is preserved.

Unfortunately $\liminf_{\max_i |x_i|\to+\infty}\mathcal{G}_{2,N}(x) = \mathbb{E}\,|X|^2 < +\infty$ and one only has that $\liminf\limits_{\min_i |x_i|\to+\infty}\mathcal{G}_{2,N}(x) = +\infty$. Hence, the function $\mathcal{G}_{2,N}$ *is not* a standard *potential function* for optimization purposes and the classical convergence results from optimization theory do not apply: the boundedness of the sequence $(x(k))_{k\geq 0}$ cannot be established, except when μ itself has a compact support, so that its closed convex hull is compact too.

3.2.2. *The $CLVQ$ as a stochastic gradient descent.*

In fact, there is an upstream problem: beyond dimension $d \geq 2$ or 3, an accurate computation of the components of $\mathcal{G}_{2,N}(x)$ for a given $x \in (\mathbb{R}^d)^N$ becomes too costly to be used in such a recursive procedure as soon as μ has an infinite support (especially if μ has no atom). When computing simple integrals on $\mathbb{R}^d$, it is well-known that, as the dimension d increases, one has to switch from cubature formulas to Monte Carlo simulations, provided μ can be simulated. For the same reason and, in the same spirit, we will switch from the above deterministic gradient descent to a *stochastic gradient descent*, taking advantage of the representations (3.40) of $\nabla\mathcal{G}_{2,N}$ as an expectation. We also assume that the distribution μ can be simulated at reasonably low cost.

Such a *stochastic gradient descent* formally reads in our setting (the factor 2 is dropped on purpose)

$$x(k+1) = x(k) - \gamma_{k+1}\left(\mathbf{1}_{\{X_{k+1}\in C_i(x)\}}(x_i(k) - X_{k+1})\right)_{1\leq i\leq N},\ x(0) \in \Big(\mathrm{Hull}\big(\mathrm{supp}(\mu)\big)\Big)^N, \tag{3.42}$$

$x(0)$ having pairwise distinct components. We simply replaced *mutatis mutandis* the expectation operator $\mathbb{E}$ by a sequence $(X_k)_{k\geq 1}$ of independent copies of X defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

The iterates $(x(k))_{k\geq 0}$ are now random vectors (though we will still denote them by script letters). When μ assigns no mass to hyperplanes, one shows that the algorithm is well defined *i.e.* no components of $x(k+1)$ get stuck as a result of an iteration. As could be expected from the deterministic framework, and for the same reason induced by the behavior of $\mathcal{G}_{2,N}$ at infinity, it is all the more hopeless to apply the classical counterpart convergence theorems for stochastic gradient like those established, *e.g.* in [8, 27, 41]. Of course, its main asset is that it can be implemented very easily when μ is simulatable. In fact, under various names (*k-means*, *CLVQ*, *nuées dynamiques*, etc), it has been widely implemented for years in the communities of Artificial Neural Networks, Data Mining and, more recently, Machine Learning, as a an *unsupervised clustering procedure* producing prototypes and Voronoi cells (classifiers) to perform automatic classification (see below).

3.2.3. *The special case of an empirical measure (Data mining).*

Note however that, when μ is the empirical measure of a dataset $(\xi_k)_{1\leq k\leq n}$ given by (1.9), the deterministic gradient descent (3.41) at a given level $N \ll n$ can always be implemented since each computation of $\nabla\mathcal{G}_{2,N}$ will require n computations of the known function appearing inside the expectation operator in (3.40). In this framework, the procedure is also known as *Forgy's algorithm* or *batch k-means* procedure by practitioners in Data mining. However, even in that case, it can be too much time consuming in practice when the size n of the dataset is too large. So, practitioners often come back to the above stochastic gradient by *sampling* at random uniformly in the dataset the input datum at each iteration which can be seen as bootstrapping.

When implemented at level $N\geq 1$ for optimal classification purposes, the resulting procedure is designed to – hopefully - converge toward a(n at least local) minimum of $\mathcal{G}_{2,N}$. If so, it produces N *prototypes* of the dataset (the resulting N-quantizer). The classification is finally obtained by assigning to each datum of the dataset the *label* of its Voronoi cell among the prototypes.

3.2.4. *Stochastic Gradient Descent formula: the CLVQ step by step.*

At this stage, it is time et us describe more precisely what this *CLVQ* algorithm (3.42) really does from a geometric point of view at each iteration.

Still starting from a starting N-tuple $x(0) = \big(x_1(0),\dots,x_N(0)\big)$ in $\big(\mathrm{Hull}\big(\mathrm{supp}(\mu)\big)\big)^N$, we update the procedure from k to $k+1$ in a two fold procedure which can be analyzed and interpreted as follows: let $x(k) := \big(x_1(k),\dots,x_N(k)\big)\in(\mathbb{R}^d)^N$ be the value of the state vector of the algorithm at the k^{th} iteration (the running vector of "prototypes").

▷ *Competition phase*: This phase is also known as *the winner selection* stage; it amounts to solving the following *nearest neighbour search*

$$i_{win}(k+1)\in\mathrm{argmin}_{i\in\{1,\dots,N\}}\big|x_i(k)-X_{k+1}\big|.$$

When the input X_{k+1} falls in a *general position* (*i.e.* not on a median hyperplane), $i_{win}(k+1)$ is uniquely defined. In case of conflict, when X_{k+1} falls on a median hyperplane and has subsequently (at least) two nearest neighbours, a rule to determine the winner has to be defined. Picking up $i_{win}(k+1)$ at random among all the nearest neighbours seems the most natural rule. Note that when μ assigns no mass to hyperplanes (*e.g.* because it is absolutely continuous) this situation $\mathbb{P}$-*a.s.* never occurs.

▷ *Learning phase*: During this second phase, the current N-tuple set of elementary quantizers (or prototypes) $x(k) = (x_i(k))_{i=1,\dots,N}$ is updated to incorporate the information provided by the input X_{k+1}. As a result the nearest neighbour $x_{i_{win}(k+1)}(k)$ is moved closer to $k+1$ by a dilatation centered at X_{k+1} with a ratio in $1-\gamma_{k+1}\in(0,1)$ since $\gamma_{k+1}\in(0,1)$. To be more precise

$$\begin{cases} x_{i_{win}(k+1)}(k+1) = \mathrm{Dilat}(X_{k+1},1-\gamma_{k+1})(x_{i_{win}(k+1)}(k)), \\ \\ x_i(k+1) = x_i(k) \qquad \text{for every } i\neq i_{win}(k+1) \end{cases}$$

where $\mathrm{Dilat}(\xi,\rho)$ denotes the dilation centered at ξ with ratio $\rho\in(0,1)$ defined by

$$\mathrm{Dilat}(\xi,\rho)(u) = \xi+\rho(u-\xi),\ u\in\mathbb{R}^d.$$

Note that, still because the step $\gamma_k \in (0,1)$, $x_{i_{win}}(k+1)$ is a *convex combination* of $x_{i_{win}}(k)$ and X_{k+1}. Consequently, as the pairwise distinct components of $x(0)$ lie in $\mathrm{Hull}\big(\mathrm{supp}(\mu)\big)^N$, a straightforward induction shows that so will be the case of all the iterations $x(k)$, $k \geq 0$.

One also deduces that, if the N-tuple $x(k)$ has pairwise distinct components, this feature is preserved by the *learning* phase. So that the above procedure is well-defined, up to the convention to be made in case of conflict between several components $x_j(k)$ in the competitive phase.

The name of this stochastic optimization algorithm – Competitive Learning Vector Quantization – is a synthesis of these two phases.

The heuristics is that the $CLVQ$ procedure *a.s.* converges at least toward an local minimum of $\mathcal{G}_{2,N}$. It has been rigorously proved in few situations (*e.g.* when $d = 1$ and μ has a log-concave density on a bounded interval). But one must have in mind that, as soon as the structural dimension d increases, the "landscape" of the potential function $\mathcal{G}_{2,N}$ – the quadratic distortion – has many local minima and various kinds of saddle points. Then the convergence toward a the true minimum is not granted, as confirmed by numerical experiments. For partial results on the convergence of the $CLVQ$ algorithm, we may refer to [55] in the special case where X has a compactly supported distribution. For a weak rate of convergence, namely a Central Limit Theorem under the assumption that the algorithm converges toward the global minimum, we refer to [69] f.

Several L^p-variants have also been investigated more recently, in particular the k-median case corresponding to $p = 1$ (see [16]).

3.2.5. *Adaptive on line computation of the "companion parameters".*

Assume that the above $CLVQ$ procedure *a.s.* converges toward an N-tuple $x^{(N,*)} = \big(x_1^{(N,*)}, \ldots, x_N^{(N,*)}\big) \in \mathrm{Hull}\big(\mathrm{supp}(\mu)\big)^N$. One can compute *on line* the associated weights and the resulting mean quadratic quantization error as follows:

▷ *Weights* $w_i^{(N,*)} = \mathbb{P}(\widehat{X}^{x^{(N),*}} = x_i^{(N,*)})$, $i = 1, \ldots, N$: for every $i = 1, \ldots, N$,

$$w_i(k+1 := (1-\gamma_{k+1})w_i(k) + \gamma_{k+1}\mathbf{1}_{\{i_{win}=i\}} \xrightarrow{a.s.} w_i^{(N,*)} \quad \text{on the event } \Big\{x(k) \to x^{(N,*)}\Big\}.$$

▷ *Quadratic Distortion error* $\mathcal{G}_{2,N}(x) = \mathbb{E}\big|\min_{1\leq i\leq N}|X - x_i|\big|^2$:

$$\mathcal{G}_N(k+1) := (1-\gamma_{k+1})\mathcal{G}_N(k) + \gamma_{k+1}|x_{i_{win}(k+1)}(k) - X_{k+1}|^2 \xrightarrow{a.s.} \mathcal{G}_{2,N}(x^{(N,*)}) \;\text{ on the event } \Big\{x(k) \to x^{(N,*)}\Big\}.$$

Note that, since the "ingredients" involved in the above companion procedures are those involved in the competition and learning phases, they add (almost) no extra C.P.U. time cost, especially if one has in mind that the costly part of the algorithm, as well as that of the Lloyd procedure presented below, lies in the *nearest neighbour search* of the "competition phase". For a proof of the *a.s.* the convergence of the above online adaptive version of the companion procedures, we refer to [2].

In some way the CLVQ algorithm can be seen as a Non Linear Monte Carlo Simulation devised to design an optimal skeleton of the distribution μ of X, this skeleton being the distribution of $\widehat{X}^{x^{(N,*)}}$ (with an obvious abuse of notation), that is the N-tuple $x^{(N,*)}$ itself and its companion weight vector $\big(w_i^{(N,*)}\big)_{1\leq i\leq N}$.

3.3. A fixed point algorithm: Lloyd's procedures

3.3.1. *The batch Lloyd procedure*

This time, we start directly from the stationary Equation (2.35). The idea is to use this identity to devise a fixed point procedure. Let $\Gamma(0) \subset \mathrm{Hull}\big(\mathrm{supp}(\mu)\big)$ (with N pairwise distinct elements). Then, for every integer

$k \geq 0$, one updates the current grid $\Gamma(k)$ as follows:

$$\begin{cases} (i)\, \textit{Centroid updating:} & \widetilde{X}^{\Gamma(k+1)} = \mathbb{E}\big(X \,|\, \widehat{X}^{\Gamma(k)}\big) = \frac{\mathbb{E}\left(X\mathbf{1}_{\{X\in C_i(\Gamma(k)\}}\right)}{\mathbb{P}(X\in C_i(\Gamma(k)))}, \\ (ii)\, \textit{Voronoi cell updating:} & \widehat{X}^{\Gamma(k+1)} \leftarrow \widetilde{X}^{\Gamma(k+1)}. \end{cases} \tag{3.43}$$

This procedure clearly lives in $\mathrm{Hull}\big(\mathrm{supp}(\mu)\big)$ by a convexity argument derived form (i). Note that (i) can be re-written (with obvious notations) as

$$x_i(k+1) = \mathbb{E}\big(X \,|\, \widehat{X}^{\Gamma(k)} = x_i(k)\big),\ i = 1, \ldots, N.$$

As set, when μ has an infinite support, this procedure is a pseudo-algorithm as it requires to be implemented the computation of the the quantities $\mathbb{E}\big(X\mathbf{1}_{\{X\in C_i(\Gamma(k))\}}\big)$ and $\mathbb{P}\big(X\in C_i(\Gamma(k))\big)$ at each iteration k.

3.3.2. *Randomized Lloyd's procedure.*

There are basically two cases of application for the batch Lloyd procedure:

– When μ is the empirical measure of a dataset given by (1.9), the former batch Lloyd procedure can be implemented directly, at least when this dataset is not too large: one computes at each iteration the expectation and the probability appearing in the ratio (3.43)(i).

– If the distribution μ has an infinite support, the situation becomes more involved. When $d = 1$, it is often possible to compute the two quantities of interest which appear in the conditional expectation of (3.43)(i) since both are terms of the ratio are 1-dimensional integrals. This is sometimes still possible in low dimension, say $d = 2$, by using appropriate cubature formulas for numerical integration on convex sets, see *e.g.* the algorithms available on the website

www.qhull.org

and the references therein.

Nevertheless, as d increases, this becomes rapidly intractable. In order to compute both expectations in (3.43)(i) there is no alternative to the Monte Carlo (or Quasi-Monte Carlo) methods. One replaces the distribution μ of X by the empirical measure $\mu_{_M}$ of a large M-sample of (independent copies) $(X_m)_{1\leq m\leq M}$ of X defined a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, namely

$$\mu(d\xi) = \mathbb{P}_{_X}(d\xi) \leftarrow \mu_{_M}(\omega, d\xi) = \frac{1}{M}\sum_{m=1}^{M} \delta_{X_m(\omega)}(d\xi),$$

considering that, as M increases, both distributions $\mu(d\xi)$ and $\mu_{_M}(\omega, d\xi)$ are $\mathbb{P}(d\omega)$-*a.s.* get close enough. Then phase (i) (centroid updating) becomes $\Gamma(k+1) = \big\{x_i(k+1, M),\ i = 1, \ldots, N\big\}$ where

$$(i)_{_M} \ \equiv\ x_i(k+1, M) = \frac{\sum_{m=1}^{M} X_m \mathbf{1}_{\{X^m\in C_i(\Gamma(k))\}}}{\mathrm{card}\big(\{1 \leq m \leq M,\, X_m \in C_i(\Gamma(k))\}\big)},\ i = 1, \ldots, N. \tag{3.44}$$

This algorithm is often called *randomized Lloyd's algorithm*. Like for the $CLVQ$ algorithm, the convergence results for the Lloyd procedure are still partial, even in its original (deterministic) form (3.43) (see *e.g.* [67]).

Two examples of (nearly) optimal quadratic quantizations are reproduced in Figures 2 and 3. They have been obtained by a hybrid stochastic optimization procedure mixing the randomized Lloyd algorithm and the $CLVQ$ algorithm. Furthermore, it relies on a "splitting approach" that is a level-by-level computation of the grids (see [59] as concerns the normal distribution).

For more details about these numerical stochastic optimization procedures used to produce optimal quantizers at level N, we refer *e.g.* to [8, 59] for $CLVQ$, [25, 40, 67] for (randomized) Lloyd's procedure or more applied textbooks like [30].

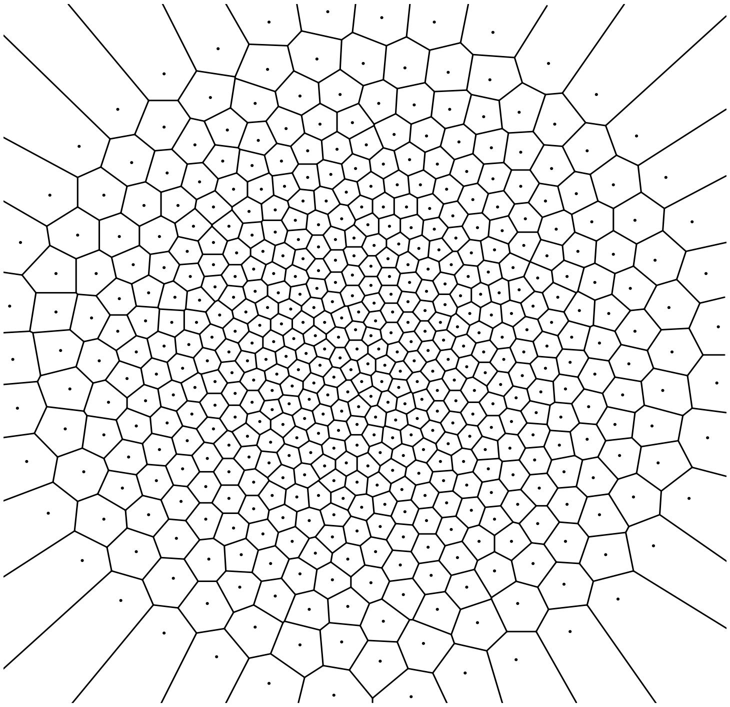

FIGURE 2. *Optimal N-quantizer ($N = 500$) of the bi-variate Normal distribution depicted with its Voronoi tessellation (with J. Printems).*

3.4. Nearest neighbor search: how to speed it up...

In both above described procedures – $CLVQ$ and randomized Lloyd's procedures – the most time consuming phase is by far the nearest neighbour search which determines the Voronoi cell in which the new input X^{k+1} falls at step $k+1$. This nearest neighbour search is well-known to have an exploding complexity when d and N are large. Its reduction or, at least, its control is known as a highly challenging problem in computer science.

There are several methods to speed it up, at least in medium dimension. We consider in what follows, either an M-sample $(X^m(\omega))_{m=1,\ldots,M}$ obtained by simulation of the distribution μ or a true dataset denoted in both cases by $(\xi_m)_{m=1,\ldots,M}$ to alleviate notations.

$\triangleright$ *Partial Distance Search principle (PDS, Chen).* This idea is quite simple and very efficient: a nearest neighbour search amounts to check, in a canonical Euclidean framework, whether a squared norm $|\xi|^2 = |(\xi^1,\ldots,\xi^d)|^2 = (\xi^1)^2 + \cdots + (\xi^d)^2$ is lower than a record value, say $\delta^2_{\rm rec}$.

To test the inequality $|\xi|^2 < \delta^2_{\rm rec}$, one proceeds by induction on the coordinates as described in **Algorithm 1** below.

The testing procedure is stopped as soon as the partial computation of the squared norm is higher than the record value. This procedure is known as Chen's Partial Distance Search protocol. Adaptation to other common norms is straightforward. It can still be improved in medium dimensions, say up to $d = 10$, by storing the data in a tree structure as described below.

$\triangleright$ *Space partitioning tree (see [29]).* As a first step, one divides the dataset into two subsets of size $\frac{M}{2}$ according to the median of the first coordinate. Then, one divides again each sub-sample following the second coordinate and so on. At each step, the procedure is applied either to a new coordinate or, when the dimension is exhausted,

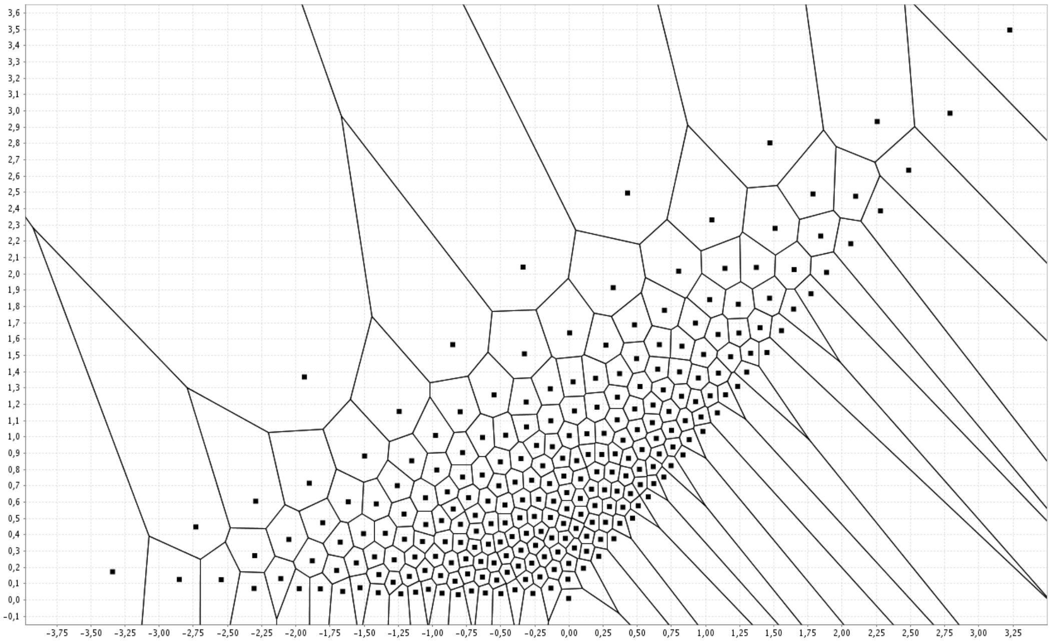

FIGURE 3. *Optimal N-quantization ($N = 250$) of $\big(W_1, \sup_{t\in[0,1]} W_t\big)$ depicted with its Voronoi tessellation, W standard Brownian motion (with B. Wilbertz).*

Algorithm 1 (Chen's Partial Distance Search protocol)

```
$record = true$;
$aux = 0$;
$\ell = 1$
while $aux + (\xi^\ell)^2 < \delta^2_{\rm rec}$ do
    $\ell := \ell + 1$;
end while
if $\ell = d + 1$ then
    $record = false$;
end if
```

to the one with the highest index. Each leaf of the resulting tree is labelled. The data are coded in the tree by their leaf label and a local index inside its assigned leaf. After this pre-processing phase, whose complexity is $O(M \log M)$, the search cost of a single datum in the tree is $O(\log M)$. It is commonly shared by users that the depth of the tree should be fixed to 7 to get the best compromise. In many situations of interest, this choice seems more or less independent of the size of the dataset and its structure.

▷ *Principal Axis Tree (McNames, see [51]).* The additional improvement brought by this approach lies in a preliminary Principal Component Analysis (PCA) inducing a change of coordinates which makes the search more efficient, especially in view of applying the PDS principle: in the new basis induced by the PCA the absolute value of coordinates tends to decrease when the coordinate index increases implying that Chen's PDS procedure stops sooner in average.

▷ *Rough Self-Quantization pre-processing (Corlay, see [17], chapter 1).* The underlying idea is completely different and competes with the PAT approach when designing the search tree. It consists of a cascade of *rough* *pre-quantizations* of the dataset with N_0 ($N_0 \approx 7$) prototypes (or classes) at each level. This again speeds up Chen's PDS principle of course.

A different approach, not detailed here, is to devise *approximate nearest neighbor search procedures*, often based on some *anchor points* (see [30] for examples and the references therein).

3.5. Where to download optimal quantization grids?

Optimized quadratic quantization grids of the d-dimensional normal distributions $\mathcal{N}(0; I_d)$, computed on the occasion of various numerical experiments (pricing of multi-asset American style options, portfolio management, nonlinear filtering, swing option pricing, etc) by the hybrid "splitting method" described above can be downloaded from the website

www.quantize.maths-fi.com

for $N = 1$ up to 10^4 and $d = 1, \ldots, 10$. The precise procedure is described in a companion file which can be downloaded. On this website are also made available *functional quantization grids* of the standard Brownian motion over the interval $[0, 1]$, the Brownian bridge, as well as a detailed procedure to compute grids for the (normalized) Ornstein-Uhlenbeck process and its bridge ([6]).

4. Greedy quantization

4.1. What is greedy quantization?

We still consider $p \in (0, +\infty)$ and $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^d$ an L^p-integrable random vector. The starting idea of *optimal greedy quantization* is to determine a *sequence* $(a_N)_{N\geq 1}$ of points of $\mathbb{R}^d$ which is recursively optimal *step by step* or *level by level* with respect to the L^p-mean quantization criterion. We mean that, if we denote $a^{(N)} = \{a_1, \ldots, a_{_N}\}$, $N \geq 1$, then the points $a_{_N}$ are recursively defined by $a^{(0)} = \emptyset$ and

$$a_{N+1} \in \operatorname{argmin}_{\xi\in\mathbb{R}^d} e_p\big(a^{(N)} \cup \{\xi\}, X\big), \quad N \geq 0. \tag{4.45}$$

Note that this section is no longer a survey but is made of more recent material, developed in an extended version in [50]. *A priori* the sequence is not uniquely defined since the above function may attain its minimum at several points of $\mathbb{R}^d$. Note that if $N = 0$, a_1 is simply the L^p-median of the distribution μ of X and is subsequently unique whenever $p > 1$ by a strict convexity argument. This idea to design not only optimal N-tuples but an optimal sequence which, hopefully, will produce N-tuples with a rate optimal behaviour as $N \to +\infty$ is very natural and can be compared to sequences with low discrepancy implemented in Quasi-Monte Carlo methods. In fact such sequences have already been investigated but only in an L^1 setting when X has a compact support as a model of short term planning (by contrast with long term planning, see [12]). Our aim in this section is to solve this *greedy* optimization problem for general distributions $\mu = \mathbb{P}_{_X}$ in any L^p-space in two directions: first establish the existence of such *L^p-optimal greedy sequences* and then evaluate the rate of decay of $e_p(a^{(N)}, X)$ to 0 as the quantization level N goes to infinity.

For convenience, we will introduce for every subset $A \subset \mathbb{R}^d$ and every $\xi \in \mathbb{R}^d$, the notation $\operatorname{dist}(\xi, A) = \inf_{a\in A} |\xi - a|$ for the distance between ξ and A, where $|\,.\,|$ denotes the canonical Euclidean norm in what follows. However, note that except for algorithmic aspects the results of this section are true with any norm on $\mathbb{R}^d$. $B(\xi, r)$ will denote the closed ball centered at $\xi \in \mathbb{R}^d$ with radius $r > 0$.

Proposition 4.1. *(a) If $X \in L^p_{\mathbb{R}^d}(\mathbb{P})$, then the sequence of optimization problems (4.45) admits at least one solution $(a_N)_{N\geq 1}$ where a_1 is the L^p-median of the distribution μ. Moreover, the sequence $\big(e_p(a^{(n)}, X)\big)_{1\leq n\leq N}$ is (strictly) decreasing as long as $N \leq |\operatorname{supp}(\mu)|$ (so that $a_n \notin a^{(n-1)}$, $1 \leq n \leq N$).*

[6] All downloads for scientific and non-commercial purposes are free of charges.

(*b*) *Any solution* $(a_N)_{N\geq 1}$ *to* (4.45) *satisfies* $\lim_N e_p(a^{(N)},X)=0$ *i.e.*

$$\lim_{N\to+\infty}\int_{\mathbb{R}^d}\min_{1\leq i\leq N}|\xi-a_i|^p\mu(d\xi)=0.$$

Such a solution is called an L^p-optimal greedy quantization sequence*.*

Proof. (*a*) We proceed by induction. When $N=1$, the existence of a_1 obviously follows from the fact that $a\mapsto\mathbb{E}(|X-a|^p)$ is continuous and goes to infinity as $|a|\to+\infty$. Assume there exists $a_1,\dots a_{_N}$ such that $e_p(a^{(k)},X)=\min_{a\in\mathbb{R}^d}e_p(a^{(k-1)}\cup\{a\},X)$ for $k=2,\dots,N$.

If $\operatorname{supp}(\mu)\subset\{a_1,\dots,a_{_N}\}$ then for every $a\in\mathbb{R}^d$, $e_p(a^{(N)}\cup\{a\},X)=e_p(a^{(N)},X)$. Otherwise, let $\xi^*\in\operatorname{supp}(\mu)\setminus\{a_1,\dots,a_{_N}\}$. It is clear that $|\xi-\xi^*|<\operatorname{dist}(\xi,a^{(N)})$ on the ball $B\big(\xi^*,\frac14\operatorname{dist}(\xi^*,a^{(N)})\big)$ which satisfies $\mu\big(B\big(\xi^*,\frac14\operatorname{dist}(\xi^*,a^{(N)})\big)\big)>0$. Consequently, $e_p\big(a^{(N)}\cup\{\xi^*\},X\big)<e_p\big(a^{(N)},X\big)$. Now let

$$K^0_{N+1}=\big\{\xi\in\mathbb{R}^d\,|\,e_p(a^{(N)}\cup\{\xi\},X)\leq e_p(a^{(N)}\cup\{\xi^*\},X)\big\}.$$

This is a closed non-empty set of $\mathbb{R}^d$. Now let $(\xi_k)_{k\geq1}$ be a sequence of elements of K^0_{N+1} such that $|\xi_k|\to+\infty$. as $k\to+\infty$. It follows from Fatou's Lemma that

$$\begin{aligned}\liminf_k e_p(a^{(N)}\cup\{\xi_k\})^p &\geq \int_{\mathbb{R}^d}\liminf_k\min\big(\operatorname{dist}(\xi,a^{(N)})^p,|\xi-\xi_k|^p\big)\mu(d\xi)\\ &= \int_{\mathbb{R}^d}\operatorname{dist}(\xi,a^{(N)})^p\mu(d\xi)\\ &= e_p(a^{(N)},X)^p>e_p(a^{(N)}\cup\{\xi^*\},X)^p.\end{aligned}$$

This yields a contradiction which in turn implies that K^0_{N+1} is a compact set. On the other hand, $\xi\mapsto e_p(a^{(N)}\cup\{\xi\},X)$ is clearly Lipschitz continuous on $\mathbb{R}^d$, hence it attains its minimum on K^0_{N+1} which is clearly its absolute minimum.

(*b*) It is clear that, for every $\xi\in\mathbb{R}^d$, $\min_{1\leq i\leq N}|\xi-a_i|$ is non-increasing and converges toward $\inf_{N\geq1}|\xi-a_{_N}|$ so that by the Lebesgue dominated convergence theorem ($|\xi-a_1|\in L^1(\mu)$), one has

$$e_p(a^{(N)},X)^p\downarrow\ell_\infty:=\int_{\mathbb{R}^d}\inf_{N\geq1}|\xi-a_N|^p\mu(d\xi).$$

Let $a^{(\infty)}=\{a_{_N},\,N\geq1\}$. If $\ell_\infty\neq0$, then there exists $\xi_0\in\operatorname{supp}(\mu)$ such that $\varepsilon_0=\operatorname{dist}(\xi_0,a^{(\infty)})>0$ and, for every $\xi\in B(\xi_0,\frac{\varepsilon_0}{4})$, $\operatorname{dist}(\xi,a^{(N)})\geq\frac34\varepsilon_0$ so that

$$\int_{B(\xi_0,\frac{\varepsilon_0}{4})}\operatorname{dist}\big(\xi,a^{(\infty)}\big)^p\mu(d\xi)\geq\eta_0\quad\text{with}\quad\eta_0=\Big(\frac{3\varepsilon_0}{4}\Big)^p\mu\big(B(\xi_0,\frac{\varepsilon_0}{4})\big).$$

Let N_0 be a positive integer such that,

$$\int_{\mathbb{R}^d}\operatorname{dist}\big(\xi,a^{(N_0)}\big)^p\mu(d\xi)\leq\ell_\infty+\frac{\eta_0}{2}\Big(1-\frac{1}{3^p}\Big).$$

We consider the (N_0+1)-quantizer $a^{(N_0)}\cup\{\xi_0\}$. On the one hand,

$$\int_{B(\xi_0,\frac{\varepsilon_0}{4})}\operatorname{dist}\big(\xi,a^{(N_0)}\cup\{\xi_0\}\big)^p\mu(d\xi)\leq\Big(\frac{\varepsilon_0}{4}\Big)^p\mu\Big(B(\xi_0,\frac{\varepsilon_0}{4})\Big)=\frac{\eta_0}{3^p}$$

and, on the other hand,

$$\begin{aligned}\int_{{}^cB(\xi_0,\frac{\varepsilon_0}{4})}\mathrm{dist}\big(\xi,a^{(N_0)}\cup\{\xi_0\}\big)^p\mu(d\xi) &\le \int_{{}^cB(\xi_0,\frac{\varepsilon_0}{4})}\mathrm{dist}\big(\xi,a^{(N_0)}\big)^p\mu(d\xi)\\ &\le \int_{\mathbb{R}^d}\mathrm{dist}\big(\xi,a^{(N_0)}\big)^p\mu(d\xi)-\int_{B(\xi_0,\frac{\varepsilon_0}{4})}\mathrm{dist}\big(\xi,a^{(N_0)}\big)^p\mu(d\xi)\\ &\le \ell_\infty+\frac{\eta_0}{2}\Big(1-\frac{1}{3^p}\Big)-\eta_0\end{aligned}$$

so that

$$\int_{\mathbb{R}^d}\mathrm{dist}\big(\xi,a^{(N_0)}\cup\{\xi_0\}\big)^p\mu(d\xi)\le \ell_\infty+\frac{\eta_0}{2}\Big(1-\frac{1}{3^p}\Big)-\eta_0+\frac{\eta_0}{3^p}<\ell_\infty$$

which yields a contradiction. Hence $\ell_\infty=0$ which completes the proof. □

4.2. Greedy quantization is rate optimal

4.2.1. *The main result*

In this section we answer the question whether a greedy sequence may be rate optimal *i.e.* that $e_p(a^{(N)},X)$ goes to zero at the same rate as a sequence of *optimal quantizers* at level N when $N\to+\infty$. Following [36], we define for every $b\in(0,+\infty)$ the *b-maximal function* associated to a greedy sequence $(a_N)_{N\ge1}$ by

$$\forall\,\xi\in\mathbb{R}^d,\quad \Psi_b(\xi)=\sup_{N\ge1}\frac{\lambda_d\big(B(\xi,b\,\mathrm{dist}(\xi,a^{(N)}))\big)}{\mu\big(B(\xi,b\,\mathrm{dist}(\xi,a^{(N)}))\big)}\in[0,+\infty].$$

It is clear that $\Psi_b(\xi)>0$ for every $\xi\neq a_1$ (L^p-median).

Note that this notion of maximal function (originally introduced in [36]) can be naturally defined with respect to a sequence of grids $(\Gamma_N)_{N\ge1}$ where Γ_N has size N. The theorem below yields a criterion based on the integrability of a maximal function Ψ_b which implies that an L^p-optimal greedy quantization sequence is L^p-rate optimal (in the sense of Zador's theorem). Practical easy-to-check criterions are given further on.

Theorem 4.1. *Let $p\in(0,+\infty)$ and let $\mu=\mathbb{P}_X$ be such that $\int_{\mathbb{R}^d}|\xi|^p\mu(d\xi)<+\infty$. Assume that there exists $b\in(0,\frac12)$ such that $\Psi_b\in L^{\frac{p}{p+d}}(\mu)$. Then*

$$\limsup_N N^{\frac1d}e_p(a^{(N)},X)<+\infty.$$

Proof. First, if μ is a Dirac mass δ_a for some $a\in\mathbb{R}^d$, then $a_1=a$ and $e_p(a^{(N)},X)=0$ for every integer $N\ge1$. Otherwise, we rely on the following micro-macro inequality established in [36] (see Equation (3.4) in the proof of Theorem 2).

$$\forall\,\xi\in\mathbb{R}^d,\qquad \mathrm{dist}(\xi,a^{(N)})^p\le\frac{C_{p,b}}{\mu\big(B(y,bd(y,a^{(N)}))\big)}\Big(e_p(a^{(N)},X)^p-e_p(a^{(N)}\cup\{\xi\},X)^p\Big)$$

where $b\in(0,\frac12)$ and $C_{p,b}$ is a positive real constant depending on p and b. Then, it follows that

$$e_p(a^{(N)}\cup\{\xi\},X)^p\le e_p(a^{(N)},X)^p-\frac{1}{C_{p,b}}\frac{\mu\big(B(\xi,b\,\mathrm{dist}(\xi,a^{(N)}))\big)}{\lambda_d\big(B(\xi,b\,\mathrm{dist}(\xi,a^{(N)}))\big)}b^d\mathrm{dist}(\xi,a^{(N)})^{p+d}V_d$$

where V_d denotes the hyper-volume of the unit ball with respect to the current norm on $\mathbb{R}^d$ *i.e.* $V_d = \lambda_d\big(B_{|\,.\,|}(0;1)\big)$. This implies that

$$e_p(a^{(N)}\cup\{\xi\},X)^p \le e_p(a^{(N)},X)^p - \frac{1}{\widetilde{C}_{p,b,d}}\frac{1}{\Psi_b(\xi)}\mathrm{dist}(\xi,a^{(N)})^{p+d}$$

where $\widetilde{C}_{p,b,d} = C_{p,b}/(b^dV_d)\in(0,+\infty)$. Note that $\mu(\{a_1\})<1$ since μ is not a Dirac mass, so that

$$\int_{\mathbb{R}^d}\Psi_b^{\frac{p}{p+d}}d\mu > 0.$$

Consequently, as $\Psi_b\in L^{\frac{p}{p+d}}(\mu)$, we can define the probability distribution $\nu = \kappa_{b,p,d}\,\Psi_b^{\frac{p}{p+d}}.\mu$ (where $\kappa_{b,p,d} = \Big(\int_{\mathbb{R}^d}\Psi_b^{\frac{p}{p+d}}d\mu\Big)^{-1}\in(0,+\infty)$ is a normalizing real constant). Then, integrating the above inequality with respect to ν yields

$$\int_{\mathbb{R}^d}e_p\big(a^{(N)}\cup\{\xi\},X\big)^p\nu(d\xi)\le e_p(a^{(N)},X)^p-\frac{1}{\widetilde{C}_{p,b,d}}\int_{\mathbb{R}^d}\mathrm{dist}(\xi,a^{(N)})^{p+d}\frac{\nu(d\xi)}{\Psi_b(\xi)}.$$

Jensen's Inequality applied to the convex function $u\mapsto u^{1+\frac{d}{p}}$ yields

$$\begin{aligned}\int_{\mathbb{R}^d}\mathrm{dist}(\xi,a^{(N)})^{p+d}\frac{\nu(d\xi)}{\Psi_b(\xi)} &\ge \left(\int_{\mathbb{R}^d}\mathrm{dist}(\xi,a^{(N)})^p\frac{\nu(d\xi)}{\Psi_b(\xi)^{\frac{p}{p+d}}}\right)^{1+\frac{d}{p}}\\ &= \kappa_{b,p,d}^{1+\frac{d}{p}}\left(\int_{\mathbb{R}^d}\mathrm{dist}(\xi,a^{(N)})^p\mu(d\xi)\right)^{1+\frac{d}{p}}\\ &= \kappa_{b,p,d}^{1+\frac{d}{p}}\,e_p\big(a^{(N)},X\big)^{p+d}.\end{aligned}$$

On the other hand, it is clear that

$$e_p\big(a^{(N+1)},X\big)^p\le\int_{\mathbb{R}^d}\nu(d\xi)e_p\big(a^{(N)}\cup\{\xi\},X)\big)^p$$

so that, finally, if we set $A_{N} = e_p(a^{(N)},X)^p$, $N\ge1$, this sequence satisfies for every integer $N\ge1$, the recursive inequality

$$A_{N+1}\le A_{N}-\widetilde{\kappa}_{b,p,d}A_{N}^{1+\frac{d}{p}}$$

where $\widetilde{\kappa}_{b,p,d} = \kappa_{b,p,d}^{1+\frac{d}{p}}/\widetilde{C}_{r,b,d}$. The sequence $(A_{N})_{N\ge1}$ being non-negative, one classically derives the announced conclusion. □

Remark. When μ has no absolutely continuous part with respect to the Lebesgue measure, it is likely that, like for standard optimal vector quantization in Zador's Theorem, this rate is not optimal. The natural conjecture should be that greedy quantization sequence(s) go to 0 at the same rate as that obtained for sequences of optimal quantizers which is not $N^{-\frac{1}{d}}$ when the distribution μ is singular (see *e.g.* [33]).

We produce below easy-to-check criterions on μ borrowed (and slightly adapted) from [36] that ensure that Ψ_b has the required integrability property so that any greedy sequence is rate optimal.

▷ *The case of compactly supported distributions.*

Lemma 4.1 (see Lemma 1 in [36])**.** *If $X \in L^p_{\mathbb{R}^d}(\mathbb{P})$ has a distribution μ having an absolutely continuous component and if $(\Gamma_N)_{N\geq 1}$ is a sequence of N-quantizers such that $\int_{\mathbb{R}^d} \mathrm{dist}(\xi, \Gamma_{\scriptscriptstyle N})^p \mu(d\xi) \to 0$ then the maximal functions Ψ_b lie in all $L^r_{loc}(\mu)$ for every $r \in (0,1)$ i.e.*

$$\forall\, r \in (0,1),\ \forall\, b \in (0,+\infty),\ \forall R \in (0,+\infty), \qquad \int_{\{|\xi|\leq R\}} \psi_b(\xi)^r \mu(d\xi) < +\infty.$$

By combining this result (with $r = \frac{p}{p+d}$) with Proposition 4.1, we derive the following proposition which generalizes that stated in [12] for absolutely continuous distributions μ having a compact convex support on $\mathbb{R}^d$.

Proposition 4.2 (Compact support)**.** *If X has a distribution μ with compact support, then any greedy L^p-optimal sequence $(a_{\scriptscriptstyle N})_{N\geq 1}$ is rate optimal i.e. satisfies*

$$\limsup_N N^{\frac{1}{d}} e_p(X, a^{(N)}) < +\infty.$$

▷ *The case of non-compactly supported distributions with radial density functions*

Lemma 4.2 (see Corollary 3 in [36])**.** *Assume $X \in L^{p+\eta}_{\mathbb{R}^d}(\mathbb{P})$ for some $\eta > 0$ with an essentially radial distribution $\mu = \varphi.\lambda_d$ in the sense that*

$$\varphi = h(|\,.\,|_0) \text{ on } B_{|\,.\,|_0}(0,R)^c \quad \text{with } h : (R,+\infty) \to \mathbb{R}_+, \text{ non-increasing and } |\,.\,|_0 \text{ a norm on } \mathbb{R}^d. \tag{4.46}$$

Let $(\Gamma_N)_{N\geq 1}$ be a sequence of N-quantizers such that $\int_{\mathbb{R}^d} \mathrm{dist}(\xi, \Gamma_{\scriptscriptstyle N})^p \mu(d\xi) \to 0$. If there exists a real constant $c > 1$ such that

$$\int_{\mathbb{R}^d} \varphi(c\,\xi)^{-\frac{p}{p+d}} \mu(d\xi) = \int_{\mathbb{R}^d} \varphi(c\,\xi)^{-\frac{p}{p+d}} \varphi(\xi) d\xi < +\infty \tag{4.47}$$

then $\Psi_b \in L^{\frac{p}{p+d}}(\mu)$.

In fact Corollary 3 in [36] is stated to be used only with L^p-optimal quantizers so the above formulation includes minor modifications. Combining this lemma with Theorem 4.1 yields the following proposition.

Proposition 4.3 (Non-compact support with radial density)**.** *Assume $X \in L^{p+\eta}_{\mathbb{R}^d}(\mathbb{P})$ for some $\eta > 0$ with an essentially radial distribution in the sense of (4.46). If φ satisfies (4.47), then any greedy L^p-optimal sequence $(a_{\scriptscriptstyle N})_{N\geq 1}$ is rate optimal i.e. satisfies*

$$\limsup_N N^{\frac{1}{d}} e_p(X, a^{(N)}) < +\infty.$$

This case includes *e.g.* all the (centered) *hyper-exponential distributions* of the form $\mu = \varphi.\lambda_d$ with

$$\varphi(\xi) = \kappa_{a,b,c} |\xi|_0^c\, e^{-a|\xi|_0^b},\ \xi \in \mathbb{R}^d,\ a, b > 0,\ c > -d$$

and $|\,.\,|_0$ denotes any norm on $\mathbb{R}^d$ and subsequently all hyper-exponential distributions distributions since L^p-mean-quantization errors is invariant by translation of the random vector X. In particular, this includes all normal and Laplace distributions

Remark. In one dimension, (4.46) can be replaced *mutatis mutandis* by the following one-sided variant: there exists $R_0, R_0' \in \mathbb{R}$, $R_0' \geq R_0$ such that

$$\mathrm{supp}(\mu) \subset [R_0, +\infty) \ \text{ and } \ f_{|[R_0',+\infty)} \text{ is non-increasing.} \tag{4.48}$$

This criterion is *e.g.* satisfied by the gamma distributions on $\mathbb{R}_+$ (including the exponential distributions).

▷ *The case of non-compactly supported distributions with possibly non-radial density functions.*

A criterion for non-radial density functions can also be derived from Corollary 4 in [36], see [50] for details.

4.2.2. *Some additional remarks and two questions.*

From now on, we denote by $\big(a_{N,p}\big)_{N\geq 1}$ an $L^p(\mu)$-optimal greedy quantization sequence and by $\big(a^{(N),p}\big)_{N\geq 1}$ the resulting sequence of N-quantizers.

$\triangleright$ *Optimal rate of decay.* If $\mathbb{E}|X|^{p+\eta} < +\infty$ for some $\eta > 0$ and $\mu = \mathbb{P}_X$ has a non-zero absolutely continuous component $\varphi.\lambda_d$ with respect to the Lebesgue measure, Zador's Theorem implies

$$\liminf_N N^{\frac{1}{d}} e_p(a^{(N)}, X) \geq \lim_N N^{\frac{1}{d}} e_p(X) = \widetilde{J}_{p,d} \|\varphi\|^{\frac{1}{p}}_{L^{\frac{p}{p+d}}(\lambda_d)} > 0.$$

In that setting, we derive that $e_p(a^{(N)}, X) \asymp N^{-\frac{1}{d}}$ as $N \to +\infty$.

$\triangleright$ *About the sharpness of optimal greedy sequences.* It is pointed out in [12] (see Theorem 4.10) that, by contrast with what is observed with global optimization at level N (Zador's Theorem), $N^{\frac{1}{d}} e_p(a^{(N)}, X)$ usually does not converge when $(a_N)_{N\geq 1}$ is a greedy L^p-optimal sequence. The counter-example is exhibited in the one-dimensional framework for the uniform distribution $U([0,1])$ where an L^1-optimal greedy sequence $(a_{N,1})_{N\geq}$ is analyzed. It is shown that for this sequence $N e_p\big(a^{(N),1},\, U([0,1])\big)$ does not converge as N goes to infinity and $1 < \liminf_N \frac{e_1\left(a^{(N),1},\, U([0,1])\right)}{e_{1,N}\left(U([0,1])\right)} < \limsup_N \frac{e_1\left(a^{(N),1}, U([0,1])\right)}{e_{1,N}\left(U([0,1])\right)} < +\infty$. Other numerical experiments reproduced below in the quadratic case $p = 2$, still with $U([0,1])$, on the one hand and, on the other hand, in [50], with the scalar and bivariate normal distributions $\mathcal{N}(0;1)$ and $\mathcal{N}(0;I_2)$ (which have an unbounded support), strongly suggest that, for more general absolute continuous distributions μ on $(\mathbb{R}^d, \mathcal{B}or(\mathbb{R}^d))$, $L^p(\mu)$-optimal greedy sequence(s) $(a_{N,p})_{N\geq 1}$ satisfy

$$\liminf_N \frac{e_p\big(a^{(N),p}, \mu\big)}{e_{p,N}\big(\mu\big)} > 1.$$

However, no proof of this fact is known to us so far. As a consequence, owing to Theorem 2.2 (and the remark that follows), this would prove that *no subsequence extracted from an $L^p(\mu)$-greedy optimal sequence $(a_{N,p})_{N\geq 1}$ can produce a (subsequence) $\big(a^{(N'),p}\big)_{N\geq 1}$ of asymptotically L^p-optimal N'-quantizers for μ.*

$\triangleright$ *Rate optimality of non-greedy sequences.* Another natural question arises at this stage: "Are there rate optimal sequences for the L^p-mean quantization error which are not solution to the greedy problem?"

To answer – positively – to this question, let us consider the celebrated *dyadic Van der Corput* (*VdC*) sequence, viewed as a quantization sequence rather than a sequence with low discrepancy. Let us recall that the *VdC* sequence is defined by

$$\forall\, N \geq 1, \quad \xi_N = \sum_{k=0}^{r} \frac{n_k}{2^{k+1}} \quad \text{where} \quad N = n_r 2^r + \cdots + n_0,\; n_i \in \{0,1\},\; i = 1, \ldots, r. \tag{4.49}$$

$\to$ *L^1-mean quantization problem.* Elementary computations carried out with the L^1-mean quantization error modulus, not reproduced here, show that

$$\liminf_N N e_1(\xi_1, \ldots, \xi_N, [0,1]) = \frac{1}{4} = \widetilde{J}_{1,1} \;\text{ and }\; \limsup_N N e_1(\xi_1, \ldots, \xi_N, [0,1]) = \frac{9}{32} = \frac{9}{8} \times \widetilde{J}_{1,1}$$

where $\widetilde{J}_{1,1} = \lim_N e_{1,N}\big(U([0,1])\big)$. This $\liminf$ is achieved by the subsequence $N' = 2^{n-1}$, $n \geq 1$, and the $\limsup$ with the subsequence $N = \frac{3}{2}.2^n = 3.2^{n-1}$. So we can answer to the first question: *it does exist rate optimal sequences for the L^1-mean quantization error which are not solutions to the greedy problem* (4.45). In fact this example shows that it even exists rate optimal sequences $(\xi_N)_{N\geq 1}$ containing subsequence of quantizers $(\xi^{(N')})_{N\geq 1}$ which are asymptotically L^1-rate optimal quantizers: so is the case of the Van der Corput sequence with the above subsequence $N' = 2^{n-1}$.

On the other hand the sequence $\big(a^{(N),1}\big)_{N\geq 1}$ has been investigated in [12], numerical computations carried out in this paper still suggest that

$$\liminf_N Ne_1(a^{(N),1}, U([0,1])) \approx 1,02 \times \widetilde{J}_{1,1} \ \text{ and } \ \limsup_N Ne_1(a^{(N),1}, U([0,1])) \approx 1,09 \times \widetilde{J}_{1,1}.$$

The second approximate asymptotics shows that the sequence $\big(a_{N,1}\big)_{N\geq 1}$ outperforms the VdC sequence from the $\limsup$ criterion since $1.09 < 1.125 = 9/8$.

→ *L^2-mean quantization problem.* The same phenomenons are confirmed in the quadratic case since

$$\liminf_N Ne_2(\xi_1,\ldots,\xi_{_N},[0,1]) = \frac{1}{2\sqrt{3}} = \widetilde{J}_{2,1} \ \text{ and } \ \limsup_N Ne_2(\xi_1,\ldots,\xi_{_N},[0,1]) = \frac{3\sqrt{5}}{4} \times \widetilde{J}_{2,1}$$

where $\widetilde{J}_{2,1} = \lim_N Ne_{2,N}(U([0,1])) = \inf_N Ne_{2,N}(U([0,1]))$. Using the greedy Lloyd procedure briefly described in the next Section 4.3, we observe numerically that

$$\liminf_N Ne_2(a^{(N),2}, U([0,1])) \approx 1.02732 \times \widetilde{J}_{2,1} > \widetilde{J}_{2,1} \ \text{ and } \ \limsup_N Ne_2(a^{(N),2}, U([0,1])) \approx 1.13401 \times \widetilde{J}_{2,1}.$$

As for the $\liminf$, one checks again that no subsequence of $\big(a^{(N),2}\big)_{N\geq 1}$ can be asymptotically L^2-optimal. As for the $\limsup$, one checks again that the quadratic optimal greedy sequence $\big(a_{N,2}\big)_{N\geq 1}$ outperforms the VdC sequence from the $\limsup$ criterion since $1.13401 < 1.67706 = \frac{3\sqrt{5}}{4}$.

Both these results and numerical experiments naturally lead to two open theoretical questions, left for further investigations:

(1) Does it exist distributions μ (with a non-countable support) for which an L^p-optimal greedy sequence $(a_{_{N,p}})_{N\geq 1}$ produces a sequence of asymptotically optimal N-quantizers$(a^{(N),p})_{N\geq 1}$ *i.e.* such that $\frac{e_p(a^{(N),p},\mu)}{e_{p,N}(\mu)} \to 1$ as N goes to $+\infty$; or at least subsequences $(a^{(N'),p})_{N\geq 1}$ having this property?

(2) Does an $L^p(\mu)$-optimal greedy sequence for a distribution μ on $\mathbb{R}^d$ produce the lowest possible value for $\limsup_N N^{\frac{1}{d}} e_{p,N}\big(a^{(N),p},\mu\big)$ among all $\mathbb{R}^d$-valued sequences $(\xi_N)_{N\geq 1}$?

The answer to the first question is probably negative (as suggested by further numerical experiments carried out in [50]), whereas the second one remains more uncertain.

4.3. **Algorithmic aspects**

We adopt notations of Sections 3. Practical computation of an optimal greedy sequence of quantizers relies on obvious variants of the historical algorithms ($CLVQ$ and Lloyd) implemented recursively: to switch from level N to $N+1$, one first adds a $(N+1)^{th}$ point (sampled from the support of the distribution μ) to the N-tuple $(a_1,\ldots,a_{_N})$ computed during the first N stages of the optimization procedure. This makes up the starting $(N+1)$-tuple for the variants of the $CLVQ$ and Lloyd procedures. Then, one implements one of these two optimization procedures with the following restriction: *all formerly computed components a_i, $1\leq i\leq N$, are frozen*, and only the new point is moved, following the standard rules. Thus, assume we implement a greedy $CLVQ$ algorithm, based on an i.i.d. sequence $(X_m)_{m\geq 1}$ of copies of X. When the $(N+1)^{th}$ component is the "winner" in the competition phase (*i.e.* this $(N+1)^{th}$ component at the m^{th} iteration is the nearest neighbour to the new input stimulus, say X_{m+1}). For Lloyd's procedure, this $(N+1)^{th}$ component is the only one to be updated following (3.44), the other N components remaining frozen as well.

For more details about these greedy variants, especially Lloyd's, we refer to [50] where convergence proofs are provided (complete in 1-dimension, partial in higher dimension).

We reproduce in Figure 4 the graph of the sequence $N \mapsto Ne_{2,N}\big(a^{(N)}, U([0,1])\big)$ where $(a_N)_{N\geq 1}$ is an L^2-optimal greedy quantization sequence for the uniform distribution $U([0,1])$.

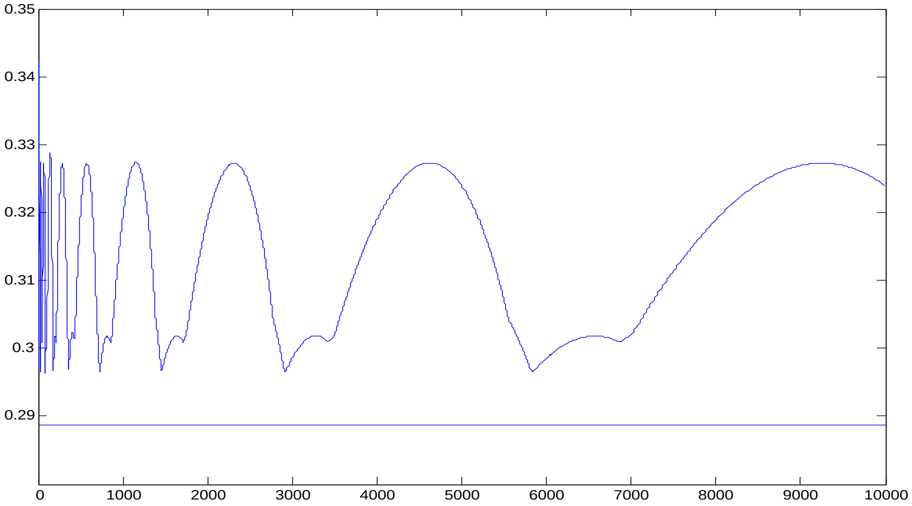

FIGURE 4. $N \mapsto Ne_2\big(a^{(N),2}, U([0,1])\big)$, $N = 1,\ldots,10\,000$, $(a_{_{N,2}})_{N\geq 1}$ *L^2-optimal greedy sequence; flat solid line* (——) $\equiv \widetilde{J}_{2,1} = \sqrt{\frac{1}{2\sqrt{3}}}$.

4.4. Greedy quantization versus Quasi-Monte Carlo?

Of course, for every integer $N \geq 1$, the *weights* induced by the μ-mass of the Voronoi cells associated to $a^{(N)}$ define canonically a sequence of N-tuples which usually cannot be "arranged" into a sequence, even up to a re-scaling. When considering the unit hypercube $[0,1]^d$ as a state space in d dimension, it is easy natural to compare an optimal greedy sequence with respect to the uniform distribution $U([0,1]^d)$ and the so-called *uniformly distributed sequences* usually implemented in the Quasi-Monte Carlo method.

Let us recall that a sequence $(\xi_{_N})_{N\geq 1}$ is uniformly distributed over $[0,1]^d$ if the empirical measures $\nu_{_N} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \delta_{\xi_i}$, $N \geq 1$, weakly converges toward the *Lebesgue measure* λ_d on $[0,1]^d$. In particular, for every bounded λ_d-*a.s.* continuous function $f : [0,1]^d \to \mathbb{R}$, $\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} f(\xi_i) \to \int_{[0,1]^d} f d\lambda_d = \int_{[0,1]^d} f(u)du$. This means that the weights associated to a uniformly distributed sequence are by definition equal to $\frac{1}{N}$ (*i.e.* are equal to 1 up to the normalization factor $1/N$). We will see that the cost induced by considering these uniform weights $\frac{1}{N}$ is essentially $\log N$. This follows from Proinov's Theorem (see [70]) recalled below, which evaluates precisely the rate of convergence of these sequences on Lipschitz continuous functions.

In the Quasi-Monte Carlo (*QMC*) method, the performance of an N-tuple $(\xi_1,\ldots,\xi_{_N}) \in ([0,1]^d)^N$ is measured by the Kolmogorov-Smirnov distance between the d-dimensional cumulative distribution function of its empirical measure $\nu_{_N}$ and the uniform distribution $U([0,1]^d)$, namely the so-called *star discrepancy* defined by

$$D^*_N(\xi_1\ldots,\xi_{_N}) = \sup_{u\in[0,1]^d} \left| \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{\xi_i\in[\![0,u]\!]\}} - \lambda_d\big([\![0,u]\!]\big)\right|$$

where $[\![0,u]\!] = \prod_{\ell=1}^{d}[0,u^\ell]$, $u = (u^1,\ldots,u^d) \in [0,1]^d$.

Several sequences $\xi = (\xi_N)_{N\geq 1}$ have been exhibited (see [54]) whose star discrepancy at the origin satisfy for a real constant $C(\xi)\in(0,+\infty)$,

$$\forall\, N\geq 1,\quad D^*_N(\xi_1\ldots,\xi_{_N})\leq C(\xi)\frac{(1+\log N)^d}{N}. \tag{4.50}$$

Among them, in 1-dimension, the dyadic VdC sequence as defined by (4.49) (and its p-adic generalizations $VdC(p)$, $p\geq 2$) and, in d-dimension, the Halton sequences (made up with $VdC(p_i)$ sequences where the bases p_i, $i=1,\ldots,d$ are the first d prime numbers), the Faure sequences, the Sobol' sequences and the unifying framework developed by Niederreiter, see [54]. For definitions of these sequences and numerical tests on various numerical integration problems, we refer to [11, 54, 56, 66]. Although such a rate has never been proved to be the lowest possible, its optimality is a commonly shared opinion in the QMC community (however see [54] for a review of existing lower bounds).

The striking fact with these sequences satisfying (4.50), called *sequences with low discrepancy*, is that, when implemented on the class of functions with *finite variation* on $[0,1]^d$, the Koksma-Hlawka inequality implies that, for every such function $f:[0,1]^d\to\mathbb{R}$

$$\left|\int_{[0,1]^d} f d\lambda_d - \frac{1}{N}\sum_{i=1}^N f(\xi_i)\right| \leq V(f)D^*_N(\xi_1\ldots,\xi_{_N})$$

where $V(f)$ denotes the variation of the function f. So it induces for this specific class of functions a rate of numerical integration of order $O\Big(\frac{(\log N)^d}{N}\Big)$. In one dimension ($d=1$), the above notion of finite variation coincides with the standard definition of finite variation in real analysis. When $d\geq 2$, several definitions can be given, the most popular being the finite variation in the Hardy & Krause sense (as described *e.g.* in [54]). Another – slightly less general but more elementary – being the *finite variation in the signed measure* sense developed in [11] (see also [56]). Unfortunately, as the dimension d increases, the set of functions with finite variation (in any of the above senses) becomes somewhat "sparse" among the set of all real-valued Borel functions defined on $[0,1]^d$. So, one may have doubts about this striking performance when dealing with practical simulation problems. This is confirmed by their behaviour on the more natural space of Lipschitz continuous functions which is ruled by the following Proinov Theorem.

Theorem 4.2. *(Proinov, [70]) Assume $\mathbb{R}^d$ is equipped with the ℓ^∞-norm $|(\xi^1,\ldots,\xi^d)|_\infty = \max\limits_{1\leq i\leq d}|\xi^i|$. For every continuous function $f:[0,1]^d\to\mathbb{R}$, let*

$$w(f,\delta) := \sup_{\xi,\,\xi'\in[0,1]^d,\,|\xi-\xi'|_\infty\leq\delta}|f(\xi)-f(\xi')|,\quad \delta\in(0,1).$$

denote the related uniform continuity modulus of f (with range δ).
(a) Let $(\xi_1,\ldots,\xi_{_N})\in([0,1]^d)^N$. For every continuous function $f:[0,1]^d\to\mathbb{R}$ and every integer $N\geq 1$,

$$\left|\int_{[0,1]^d} f d\lambda_d - \frac{1}{N}\sum_{i=1}^N f(\xi_i)\right| \leq C_d\, w\big(f, D^*_{_N}(\xi_1,\ldots,\xi_{_N})^{\frac{1}{d}}\big)$$

where $C_d\in(0,+\infty)$ is a universal optimal real constant only depending on d. In particular, if the function $f:[0,1]^d f\to\mathbb{R}$ is Lipschitz continuous with coefficient $[f]_{\rm Lip} := \sup_{x,y\in[0,1]^d}\frac{|f(x)-f(y)|}{|x-y|_\infty}$, then

$$\left|\int_{[0,1]^d} f d\lambda_d - \frac{1}{N}\sum_{i=1}^N f(\xi_i)\right| \leq C_d\,[f]_{\rm Lip} D^*_{_N}(\xi_1,\ldots,\xi_{_N})^{\frac{1}{d}}.$$

If $d = 1$, $C_d = 1$ and if $d \geq 2$, $C_d \in [1, 4]$.
(b) In particular if $(\xi_N)_{N\geq 1}$ is a sequence with low discrepancy in the above sense, then for every integer $N \geq 1$,

$$\left|\int_{[0,1]^d} f d\lambda_d - \frac{1}{N}\sum_{i=1}^{N} f(\xi_i)\right| \leq C_d\,[f]_{\rm Lip} C(\xi)\frac{1+\log N}{N^{\frac{1}{d}}}.$$

Remark. As the function $f_\xi : u \mapsto \min_{1\leq i\leq N} |u - \xi_i|$ defined on $[0, 1]$ is 1-Lipschitz continuous and equal to 0 on $\{\xi_1, \ldots, \xi_N\}$, Proinov's inequality in (a) implies that $e_1\big(\xi_1, \ldots, \xi_N, U([0, 1])\big) \leq D^*_N(\xi_1, \ldots, \xi_N)$.

The above claim (b) emphasizes the fact that considering uniform weights $\frac{1}{N}$ induces the loss of a $\log N$ factor compared to an $(L^1, U([0, 1]))$-optimal greedy (or simply rate optimal) sequence since, for such a greedy sequence $(a_N)_{N\geq 1}$, one has, for every $N \geq 1$,

$$\left|\int_{[0,1]^d} f d\lambda_d - \sum_{i=1}^{N} w_i^{(N)} f(a_i)\right| \leq \kappa(a)\,[f]_{\rm Lip}\frac{1}{N^{\frac{1}{d}}}.$$

Of course the practical implementation of such greedy sequences is more demanding since one needs to have access to the attached N-tuples of weights $w^{(N)} = (w_1^{(N)}, \ldots, w_N^{(N)})$. However, by contrast, one checks that cubature formulas based on optimal quantization turn out to be efficient for much lower values of N than sequences with low discrepancy (see *e.g.* [59] for the pricing of European derivatives).

References

[1] Achdou Y., Pironneau O. (2005). Computational Methods for Option Pricing, *SIAM series: Frontiers in Applied Mathematics*, 297p.

[2] Bally V., Pagès G. (2003). A quantization algorithm for solving discrete time multidimensional optimal stopping problems, *Bernoulli*, **9**(6):1003-1049.

[3] Bally V., Pagès G. (2003). Error analysis of the quantization algorithm for obstacle problems, *Stoch. Proc. & their Appl.*, **106**(1):1-40.

[4] V. Bally, G. Pagès and J. Printems (2005): A quantization method for pricing and hedging multi-dimensional American style options, *Mathematical Finance*, **15**(1):119-168.

[5] Bensoussan A. (1984). On the theory of option pricing, *Acta Appl. Math.*, **2**:139-158.

[6] Bensoussan A., Lions J.-L. (1978). *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 525p.

[7] Bally V., Talay D. (1996). The distribution of the Euler scheme for stochastic differential equations: I. Convergence rate of the distribution function, *Probab. Theory Related Fields*, **104**(1): 43-60.

[8] Benveniste A., Métivier M. and Priouret P. (1987). *Algorithmes adaptatifs and approximations stochastiques.* Masson, Paris, 367p. English updated translation (by S.S. Wilson), (2012). *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag Berlin and Heidelberg, 380p.

[9] Beskos A., Roberts G.O. (2005). Exact simulation of diffusions, *Ann. Appl. Prob.*, **15**(4): 2422-2444.

[10] Biau G., Devroye L. and Lugosi G. (2008). On the performance of clustering in Hilbert spaces, *IEEE Transactions on Information Theory*, **54**:781-790.

[11] Bouleau N., Lépingle D. (1994). *Numerical methods for stochastic processes*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 359p.

[12] Brancolini A., Butazzo G., Santambrogio F. and Stepanov E. (2009). Long-term planning *versus* short term planning in the asymptotical location problem, *ESAIM: COCVV*, **15**, 509-524.

[13] Brandejsky A., de Saporta B. and Dufour F. (2013). Optimal stopping for partially observed piecewise-deterministic Markov processes, *Stochastic Processes and their Applications*, **123**(8): 3201-3238.

[14] Brandejsky A., de Saporta B. and Dufour F. (2012). Numerical method for impulse control of piecewise deterministic Markov processes, *Automatica*, **48**(5): 779-793.

[15] Bucklew J.A., Wise G.L. (1988). Multidimensional Asymptotic Quantization Theory with r^{th} Power Distortion Measures, *IEEE Transactions on Information Theory*, Vol. IT-**28**(2): 239-247.

[16] Cardot H., Cénac P. and Monnez J.-M. (2012). A fast and recursive algorithm for clustering large dataset with k-medians, *Computational Statistics & Data Analysis*, **56**(6):1431-1449.

[17] Corlay S. (2011). *Some aspects of optimal quantization and applications to finance.* Thèse de l'UPMC. http://tel.archives-ouvertes.fr/tel-00626445/

[18] Corlay S., Pagès G. (2014). *Functional quantization based stratified sampling methods*, to appear in *Monte Carlo Methods and Appl.*. arXiv: 1008.4441v2.

[19] Corsi M., Pham H. and Runggaldier W. (2008). Numerical approximation by quantization of control problems in finance under partial observation, chapter from *Mathematical Modeling and Numerical Methods in Finance* (special volume, A. Bensoussan, Q. Zhang guest eds.), *Handbook of Numerical Analysis* **15** (P.G. Ciarlet Editor), North Holland, 325-360.

[20] Cover T., Thomas J. (2006). *Elements of Information Theory.* Second edition. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, xxiv+748p.

[21] Cuesta-Albertos J.A., Matrán C. (1988). The strong law of large numbers for k-means and best possible nets of Banach valued random variables. *Probab. & Theory Related Fields*, **78**:523-534.

[22] Delattre S., Graf S., Luschgy H. and Pagès G. (2004). Quantization of probability distributions under norm-based distortion measures, *Statistics & Decision*, **22**:261-282.

[23] Dembo A., Zeitouni O. (2010). *Large deviations techniques and applications.* Corrected reprint of the second (1998) edition. Stochastic Modeling and Applied Probability, **38**. Springer-Verlag, Berlin, 2010. xvi+396p.

[24] Dereich S. (2007). The coding complexity of diffusion processes under $L^p([0,1])$-norm distortion, *Stochastic Process. Appl.*, **118**(6):938-951.

[25] Du Q., Faber V. and Gunzburger M. (1999). Centroidal Voronoi Tessellations: Applications and Algorithms, *SIAM Review*, **41**: 637-676.

[26] Du Q., Emelianenko M. and Ju L. (2006). Convergence of the Lloyd algorithm for computing centroidal Voronoi tessellations, *SIAM Journal on Numerical Analysis*, **44**:102-119.

[27] Duflo M. (1997). *Algorithmes stochastiques*, coll. Mathématiques & Applications, **23**, Springer-Verlag, Berlin, 319p.

[28] Fournié É., Lasry J.M., Lebuchoux J., Lions P.-L., Touzi N. (1999). Some applications of Malliavin calculus to Monte Carlo methods in finance, *Finance and Stochastics*, **3**:391-412.

[29] Friedman J. H., Bentley J.L. and Finkel R.A. (1977). An Algorithm for Finding Best Matches in Logarithmic Expected Time, *ACM Transactions on Mathematical Software*, **3**(3):209-226.

[30] Gersho A., Gray R.M. (1991). *Vector Quantization and Signal Compression*, Springer International Series in Engineering and Computer Science, **159**, Springer, 732p.

[31] Glasserman P. (2003). Monte Carlo Methods in Financial Engineering, Springer-Verlag, New York, 596p.

[32] Gobet E., Pagès G., Pham H. and Printems J. (2007). Discretization and simulation of the Zakai Equation, *SIAM J. on Numerical Analysis*, **44**(6):2505-2538.

[33] Graf S., Luschgy H. (2000). *Foundations of Quantization for Probability Distributions.* LNM 1730. Springer, Berlin.

[34] Graf S., Luschgy H. (2002). Rates of convergence for the empirical quantization error, *Ann. Probab.*, **30**(2):874-897.

[35] Graf S., Luschgy H. and Pagès G. (2007). Optimal quantizers for Radon random vectors in a Banach space. *J. Approx. Theory* , **144**(1):27-53.

[36] Graf S., Luschgy H. and Pagès G. (2008). Distortion mismatch in the quantization of probability measures, *ESAIM P&S*, **12**:127-154.

[37] Graf S., Luschgy H. and Pagès G. (2012). The local quantization behaviour of absolutely continuous probabilities, *Annals of Probab.* **40**(4):1795-1828.

[38] Huffman D.A. (1952). A method for the construction of minimum redundancy codes, *Proc. IRE*, **40**: 1098-1101.

[39] Jaillet P., Lamberton D. and Lapeyre B. (1990). Variational inequalities and the pricing of American options, *Acta Appl. Math.*, **21**:263-289.

[40] Kieffer J.C. (1982). Exponential rate of convergence for Lloyd's method I, *IEEE Trans. on Inform. Theory*, Special issue on quantization, IT-**28**(2):205-210.

[41] Kushner H.J., Yin G.G. (2003). Stochastic approximation and recursive algorithms and applications, 2^{nd} edition, *Applications of Mathematics, Stochastic Modeling and Applied Probability*, **35**, Springer-Verlag, New York, xxii+474p.

[42] Lamberton D., Lapeyre B. (1996). *Introduction to stochastic calculus applied to finance.* Chapman & Hall, London, 185p.

[43] Lapeyre B., Pardoux É., Sentis R. (2003). *Introduction to Monte-Carlo Methods for Transport and Diffusion Equations*, coll. Oxford Texts in Applied and Engineering Mathematics, (translated from French by A. Craig and F. Craig), Oxford University Press, USA, 176p.

[44] Lamberton D. (2009). *Optimal stopping and American options*, Ljubljana Summer School on Financial Mathematics, http://www.fmf.uni-lj.si/finmath09/ShortCourseAmericanOptions.pdf.

[45] Lejay A., Reutenauer V. (2012). A variance reduction technique using a quantized Brownian motion as a control variate, *J. Comput. Finance*, **16**(2):61-84.

[46] Longstaff F.A., Schwarz E.S. (2001). Valuing American options by simulation: a simple least-squares approach, *Review of Financial Studies*, **14**:113-148.

[47] Luschgy H., Pagès G. (2002). Functional quantization of stochastic processes, *J. Funct. Anal.*, **196**:486-531.

[48] Luschgy H., Pagès G. (2004). Sharp asymptotics of the functional quantization problem for Gaussian processes, *Ann. Probab.* **32**:1574-1599.

[49] Luschgy H., Pagès G. (2008). Functional quantization rate and mean regularity of processes with an application to Lévy processes, *Annals of Applied Probability*, **18**(2):427-469.

[50] Luschgy H., Pagès G. (2014). Greedy vector quantization, pré-pub. LPMA 1624, submitted for publication. arXiv:1409.0732.

[51] McNames J. (2001). A Fast Nearest-Neighbor Algorithm Based on a Principal Axis Search Tree, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(9):964-976.

[52] Mirebeau J.M., Cohen A. (2012). Greedy bisection generates optimally adapted triangulations, *Math. Comp.*, **81**(278):811-837.

[53] Neveu J. (1972). *Martingales à temps discret*, Masson, 1972, 218p. English translation: *Discrete-parameter martingales*, North-Holland, New York, 1975, 236p.

[54] Niederreiter H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF regional conference series in Applied mathematics, SIAM, Philadelphia, 241p.

[55] Pagès G. (1998). A space vector quantization method for numerical integration, *J. Computational and Applied Mathematics*, **89**, 1-38 (Extended version of "Voronoi Tessellation, space quantization algorithms and numerical integration", in: *Proceedings of the ESANN' 93*, M. Verleysen (Ed.), Bruxelles, Quorum Editions, (1993), 221-228.

[56] Pagès G. (2014). *Introduction to Numerical Probability and Applications to Finance*, to appear, coll. Universitext, Springer. Preliminary version available at www.proba.jussieu.fr/pageperso/pages.html.

[57] Pagès G., Pham H. (2005). Optimal quantization methods for nonlinear filtering with discrete-time observations, *Bernoulli*, **11**(5):893-932.

[58] Pagès G., Pham H. and Printems J. (2004). An Optimal Markovian Quantization Algorithm for Multidimensional Stochastic Control Problems, *Stochastics and Dynamics*, **4**(4):501-545.

[59] Pagès G., Printems J. (2003). Optimal quadratic quantization for numerics: the Gaussian case, *Monte Carlo Methods and Appl.* **9**(2):135-165.

[60] Pagès G., Printems J. (2005). Functional quantization for numerics with an application to option pricing, *Monte Carlo Methods & Applications*, **11**(4):407-446.

[61] Pagès G., Printems J. (2009). Optimal quantization for finance: from random vectors to stochastic processes, chapter from *Mathematical Modeling and Numerical Methods in Finance* (special volume, A. Bensoussan, Q. Zhang guest eds.), coll. *Handbook of Numerical Analysis* (P.-G. Ciarlet Editor) **15**, North Holland, 595-649.

[62] Pagès G., Pham H. and Printems J. (2005). Optimal quantization methods and applications to numerical problems in finance, *Handbook on Numerical Methods in Finance* (S. Rachev, ed.), Birkhauser, Boston, 253-298.

[63] Pagès G., Wilbertz B. (2012). Dual Quantization for random walks with application to credit derivatives, Journal of Computational Finance, **16**(2):33-60.

[64] Pagès G., Wilbertz B. (2012). Intrinsic stationarity for vector quantization: Foundation of dual quantization, *SIAM J. on Numerical Analysis*, **50**:747-780.

[65] Pagès G., Wilbertz B. (2012). Optimal Delaunay et Voronoi quantization methods for pricing American options, *Numerical methods in Finance* (R. Carmona, P. Hu, P. Del Moral, N. Oudjane eds.), Springer, 171-217.

[66] Pagès G., Xiao Y.J. (1997). Sequences with low discrepancy and pseudo-random numbers: theoretical results and numerical tests, *Journal of Statistical Computation and Simulation*, **56**:163-183.

[67] Pagès G., Yu J. (2013). Pointwise convergence of the Lloyd algorithm in higher dimension, pré-pub PMA 1604, submitted, arXiv:1401.0192v.

[68] Pham H., Runggaldier W., Sellami A. (2004). Approximation by quantization of the filter process and applications to optimal stopping problems under partial observation, *Monte Carlo Methods and Applications*, **11**(1):57-81.

[69] Pollard D. (1982). A central limit theorem for k-means clustering, *Ann. Probab.*, **10**:919-926.

[70] Proinov P.D. (1988). Discrepancy and integration of continuous functions, *J. of Approx. Theory*, **52**:121-131.

[71] Shannon C.E. (1948). A mathematical theory of communication, *bell Syst. Tech. J.*, **27**:376-423; 623-656. Discrepancy and integration of continuous functions, *J. of Approx. Theory*, **52**:121-131.

[72] Shannon C.E., Weaver W. (1949). *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, 144p.

[73] Shiryaev A.N. (2007). *Optimal Stopping rules*, 2^{nd} edition (translated from the Russian), Applications of Mathematics, Springer Verlag, Berlin, 232p.

[74] Talay D., Tubaro L. (1990). Expansion of the global error for numerical schemes solving stochastic differential equations, *Stoch. Anal. Appl.*, **8**:94-120.

[75] Tarpey T., Li L. and Flury B. (1995). Principal Points and Self-Consistent Points of Elliptical Distributions. *Annals of Statistics*, **23**(1):103-112.

[76] Villeneuve S., Zanette A. (2002). Parabolic A.D.I. methods for pricing american option on two stocks, *Mathematics of Operation Research*, **27**(1):121-149.

[77] Zador P.L. (1982). Asymptotic quantization error of continuous signals and the quantization dimension, *IEEE Trans. Inform. Theory*, IT-**28**(2):139-14.