Greedy quantization: new approach and applications to reflected backward SDE

Rancy El Nmeir

Quantification gloutonne: nouvelle approche et applications aux E.D.S. rétrogrades réfléchies

Greedy quantization: new approach and applications to reflected backward S.D.E.

Rancy El Nmeir

Thèse pour l'obtention du grade : **Docteur de l'université Sorbonne Université**
Docteur de l'université Saint-Joseph de Beyrouth

Laboratoire de Probabilités, Statistique et Modélisation - UMR 8001 - Sorbonne Université
Laboratoire de Mathématiques et Applications - Université Saint-Joseph de Beyrouth

Sous la direction de : Gilles Pagès
Rami El Haddad

Présentée devant un jury composé de: Idriss Kharroubi (Président)
Antoine Lejay (Rapporteur)
Martino Grasselli (Rapporteur)
Richard Aoun (Examinateur)
Aurélie Fischer (Examinateur)
Gihane Mansour Abou Jaoudeh (Examinateur)
Rami El Haddad (Directeur de thèse)
Gilles Pagès (Directeur de thèse)

Date de soutenance: 18 décembre 2020

École Doctorale de Sciences Mathématiques de Paris-Centre (ED386)
École Doctorale de Sciences, Ingénierie et Technologies (EDSIT)

Résumé

Cette thèse comporte deux parties dans lesquelles nous traitons la quantification vectorielle gloutonne avec des applications financières.

Dans la première partie, nous nous concentrons sur la quantification vectorielle gloutonne. Nous commençons par présenter de nouvelles approches théorique et numérique de la quantification gloutonne. Nous établissons de nouveaux résultats d'optimalité du taux de convergence pour une classe plus large de distributions, et nous réalisons une étude numérique approfondie apportant de nombreuses améliorations dans le domaine de l'intégration numérique basé sur la quantification gloutonne. Parmi ces études, nous présentons des propriétés numériques intéressantes des suites de quantification gloutonne leur permettant de constituer un adversaire avantageux vis-à-vis les suites utilisées dans d'autres méthodes d'intégration numérique, comme les suites à discrépance faible dans la méthode quasi-Monte Carlo par exemple. De plus, nous montrons que, lorsqu'une suite de quantification gloutonne L^r-optimale est dilatée ou contractée de manière appropriée, elle reste à taux de convergence L^s-optimal. Ceci est parfois conditionné par une hypothèse de moment sur la loi de probabilité sous-jacente.

La deuxième partie de ce manuscrit est consacrée à l'approximation d'une équation différentielle stochastique rétrograde réfléchie par quantification vectorielle. Nous établissons d'abord des bornes supérieures de l'erreur dans L^p, $p \in (1, 2+d)$, induite par la quantification récursive d'une chaîne de Markov générale d'une part, et par une sorte de quantification récursive "hybride", méthode introduite dans cette thèse, d'autre part. Ensuite, nous établissons des bornes d'erreur dans L^p, $p \in (1, 2+d)$, pour le schéma de discrétisation spatiale basé sur la quantification et correspondant à l'équation différentielle stochastique rétrograde réfléchie. Cette méthode est utilisée pour évaluer les options financières, principalement les options américaines, et illustrée dans plusieurs exemples où nous comparons le comportement de la quantification récursive à celui de la quantification gloutonne en termes de précision et de coût en temps. Nous utilisons également cette technique de discrétisation pour l'évaluation du prix des options barrière.

Mots-clés: Quantification optimale- Quantification gloutonne - Quantification récursive - Discrépance - Equations différentielles stochastiques rétrogrades réfléchies - Intégration numérique - Taux de convergence optimal - Évaluation des prix d'options - Algorithme de Lloyd.

Abstract

This thesis contains two parts in which we treat greedy vector quantization with some financial applications.

In the first part, we focus on greedy vector quantization. We start by presenting new theoretical and numerical approaches of greedy quantization. We establish new rate optimality results for a larger class of distributions, and carry out an extensive numerical study bringing many improvements in the greedy quantization-based numerical integration field. Among these studies, we present interesting numerical properties of greedy quantization sequences allowing them to become an advantageous component compared to sequences used in other numerical integration methods, like the low discrepancy sequences in the quasi-Monte Carlo method for example. Furthermore, we show that, when an L^r-optimal greedy quantization sequence is dilated or contracted in an appropriate way, it remains L^s-rate optimal. This is sometimes conditioned by a certain moment assumption on the underlying probability distribution.

The second part of this manuscript is devoted to the approximation of a reflected Backward Stochastic Differential Equation by vector quantization. First, we establish upper bounds for the L^p-error, $p \in (1, 2+d)$, induced by recursive quantization of a general Markov chain one the one hand, and by a kind of "hybrid" recursive quantization, a method introduced in this thesis, on the other hand. Then, we establish L^p-error bounds, $p \in (1, 2+d)$, for the quantization-based space discretization scheme corresponding to the reflected Backward Stochastic Differential Equation. This is used for pricing financial options, mainly American options, and illustrated in several examples where we compare the behavior of recursive quantization versus greedy quantization in terms of precision and time cost. We use this discretization technique for the pricing of Barrier options as well.

Keywords: Optimal quantization - Greedy quantization - Recursive quantization - Discrepancy - Reflected Backward Stochastic Differential Equations - Numerical integration - Rate optimality - Option pricing - Lloyd's algorithm.

Contents

List of Figures

List of Tables

Chapter 1

Introduction-Français

Cette thèse est divisée en deux parties principales. La première partie contient les chapitres 3, 4 et 5 où nous présentons de nouveaux résultats et aspects théoriques de la quantification gloutonne, ainsi que quelques études numériques intéressantes. En résumé, nous adoptons une nouvelle approche pour étendre les résultats d'optimalité du taux de convergence de l'erreur de quantification gloutonne à une classe plus large de distributions et établir des résultats de L^s-optimalité du taux de convergence pour des suites de quantification gloutonne L^r-optimales dilatées ou contractées. Numériquement, nous réalisons quelques expériences et mettons en évidence certaines propriétés de la quantification gloutonne qui la rendent avantageuse face à d'autres méthodes d'approximation (principalement la méthode de quasi Monte Carlo). Dans la deuxième partie, composée des chapitres 6 et 7, nous établissons d'abord des bornes supérieures de l'erreur de quantification récursive dans L^p d'une chaîne de Markov d-dimensionnelle pour $p \in (1, 2 + d)$ et, ensuite, nous étendons les bornes d'erreur dans L^p induite par les schémas de discrétisation, basés sur la quantification récursive, des équations différentielles stochastiques rétrogrades réfléchies.

1.1 Quantification optimale: principe, définitons et principaux résultats

La quantification vectorielle optimale est une technique qui remonte aux années 1950 (voir [30]) lorsqu'elle a été conçue pour la première fois dans le domaine du traitement du signal afin de discrétiser les signaux continus pour leur transmission. Elle a ensuite été étendue à de nombreux domaines tels que la théorie de l'information, l'analyse de classification non supervisée, etc., et puis introduite comme outil mathématique dans les années 1990. Elle a d'abord été utilisée comme formule de quadrature dans le domaine de l'intégration numérique pour le calcul des espérances (voir [54]), puis, au début des années 2000, pour l'approximation des espérancess conditionnelles en vue d'applications financières, principalement l'évaluation des prix d'options américaines (voir [3, 4, 5]), des problèmes de filtrage non linéaire (voir [58]) et la simulation d'équations différentielles stochastiques (voir [3, 67]), etc.

Le problème mathématique de la quantification optimale consiste à trouver la meilleure approximation, dans un sens à préciser plus tard, d'une distribution de probabilité (éventuellement) continue par une distribution de probabilité discrète dont le support est de cardinal fini, ou, en d'autres termes, la meilleure approximation d'une variable aléatoire multidimensionnelle X par

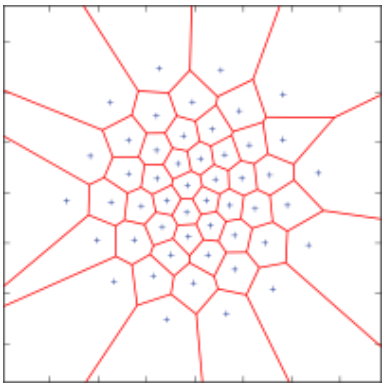

Figure 1.1: Exemple de diagramme de Voronoï dans $\mathbb{R}^2$ muni de la norme euclidienne.

une variable aléatoire Y prenant un nombre fini n de valeurs. Soit $d \geq 1$ et X une variable aléatoire d-dimensionnelle définie sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ telle que $X \in L^r(\mathbb{P}), r > 0$, c'est-à-dire $\mathbb{E}|X|^r < +\infty$ où $|.|$ désigne à priori toute norme sur $\mathbb{R}^d$. On note $P = \mathbb{P}_X$ la distribution de probabilité de X. Le but est d'approcher X par $q(X)$, où q est une fonction Borélienne définie sur $\mathbb{R}^d$ à valeurs dans une *grille* d-dimensionnelle $\Gamma = \{x_1, \ldots, x_n\}$ de taille n. Le meilleur choix pour q, Γ étant fixé, est clairement toute projection Borélienne du plus proche voisin $\pi_\Gamma : \mathbb{R}^d \to \Gamma$ définie par $\pi_\Gamma(\xi) = \sum_{i=1}^n x_i \mathbb{1}_{C_i(\Gamma)}(\xi)$, où

$$C_i(\Gamma) \subset \{\xi \in \mathbb{R}^d : |\xi - x_i| \leq \min_{j \neq i} |\xi - x_j|\}, \qquad i = 1, \ldots, n, \tag{1.1}$$

est une partition Borélienne de $\mathbb{R}^d$ appelée diagramme de Voronoï induit par Γ. Les ensembles boréliens $C_i(\Gamma)$ constituent les cellules de Voronoï de la partition induite par Γ. Un exemple de diagramme de Voronoï dans $\mathbb{R}^2$ muni de la norme euclidienne est présenté dans la figure 1.1.

Ainsi, la quantification de Voronoï de X est la composition de π_Γ et X:

$$\widehat{X}^\Gamma = \pi_\Gamma(X) := \sum_{i=1}^n x_i \mathbb{1}_{C_i(\Gamma)}(X).$$

Sa distribution est caractérisée par la grille $\Gamma = \{x_1, \ldots, x_n\}$ et les poids des cellules de Voronoï correspondantes donnés, pour chaque $i \in \{1, \ldots, n\}$, par

$$p_i^n = P(\widehat{X}^\Gamma = x_i) = P\big(X \in C_i(\Gamma)\big).$$

Nous noterons souvent, $\widehat{X}$ au lieu de $\widehat{X}^\Gamma$ pour alléger les notations. L'erreur de quantification L^r associée à une grille Γ est définie, pour chaque $r \in (0, +\infty)$, par

$$e_r(\Gamma, X) = \|X - \pi_\Gamma(X)\|_r = \|X - \widehat{X}^\Gamma\|_r = \|\mathrm{dist}(X, \Gamma)\|_r \tag{1.2}$$

où $\|Y\|_r = \big(\mathbb{E}|Y|^r\big)^{\frac{1}{r}}$ désigne la norme $L^r(\mathbb{P})$ d'un vecteur aléatoire Y (ou quasi-norme si $0 < r < 1$). Nous définissons également la fonction de L^r-distorsion G_n^r sur $\big(\mathbb{R}^d\big)^n$ par

$$G_n^r(x_1, \ldots, x_n) = e_r(\{x_1, \ldots, x_n\}, X)^r. \tag{1.3}$$

Cette fonction est différentiable si les $(x_i)_{1\le i\le n}$ sont deux-à-deux distincts ou, de manière équivalente, si $\Gamma = \{x_1, \dots, x_n\}$ est de taille n, et si les frontières du diagramme de Voronoï sont négligeables par rapport à la distribution P de X. Ceci dépend aussi de la différentiabilité de la norme sous-jacente elle-même. Son gradient est donné par

$$\nabla G_n^r(x_1, \dots, x_n) = r\left(\mathbb{E}\Big[\mathbb{1}_{X\in C_i(\Gamma)}\frac{(x_i - X)}{|x_i - X|}|x_i - X|^{r-2}\Big]\right)_{1\le i\le n}.$$

Le problème de quantification optimale consiste à trouver une grille Γ qui minimise l'erreur de quantification (1.2), c'est-à-dire qui résout le problème de minimisation suivant

$$e_{r,n}(X) \;=\; \inf_{\Gamma,\, \mathrm{card}(\Gamma)\le n} e_r(\Gamma, X). \tag{1.4}$$

Si $X \in L^r_{\mathbb{R}^d}(\mathbb{P})$, ce problème admet toujours au moins une solution Γ appelée quantifieur optimal, ou grille de quantification optimale, de taille n de X ou P, et l'erreur de quantification correspondante converge vers 0 lorsque la taille n tend vers l'infini. Pour une preuve, nous nous référons entre autres à [32, 56, 57]. Le taux de convergence de l'erreur de quantification L^r vers 0 est donné par deux résultats bien connus exposés dans le théorème suivant. Le premier est un résultat asymptotique et le second est universel et non asymptotique.

Theorem 1.1.1. (*a*) Théorème de Zador (voir [75]) *: Soit $r > 0$ et $X \in L^{r+\eta}_{\mathbb{R}^d}(\mathbb{P})$, $\eta > 0$, de distribution P tel que $dP(\xi) = \varphi(\xi)d\lambda_d(\xi) + d\nu(\xi)$ où λ_d est la mesure de Lebesgue sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Alors,*

$$\lim_{n\to+\infty} n^{\frac{1}{d}} e_{r,n}(X) = \tilde{J}_{r,d}\|\varphi\|^{\frac{1}{r}}_{L^{\frac{r}{r+d}}(\lambda_d)} \tag{1.5}$$

où $\tilde{J}_{r,d} = \inf_{n\ge1} n^{\frac{1}{d}} e_{r,n}(U([0,1]^d)) \in (0, +\infty)$.
(*b*) Lemme de Pierce étendu (voir [44])*: Soit $r, \eta > 0$. Il existe une constante $\kappa_{d,r,\eta} \in (0, +\infty)$ tel que, pour toute variable aléatoire $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^d$,*

$$\forall n \ge 1, \quad e_{r,n}(X) \le \kappa_{d,r,\eta}\sigma_{r+\eta}(X)n^{-\frac{1}{d}} \tag{1.6}$$

où, pour $p \in (0, +\infty)$, $\sigma_p(X) = \inf_{a\in\mathbb{R}^d} \|X - a\|_p < +\infty$.

Une propriété importante des quantifieurs quadratiques optimaux (dans L^2) est la stationnarité. Une grille de quantification Γ est dite stationnaire si les frontières des cellules de Voronoï sont P-négligeables et

$$\widehat{X}^\Gamma = \mathbb{E}\left(X \,|\, \widehat{X}^\Gamma\right). \tag{1.7}$$

En effet, tout quantifieur quadratique optimal (par rapport à la norme euclidienne) a des frontières P-négligeables (voir la proposition 4.2 dans [32]). Cette propriété est très importante dans la plupart des applications, notamment parce que la plupart des algorithmes conçus pour construire les quantifieurs optimaux sont basés sur cette propriété de stationnarité, même si tous les quantifieurs stationnaires ne sont pas optimaux. Son importance est également soulignée dans le domaine de l'intégration numérique basé sur la quantification, ce sujet est expliqué en détails dans ce qui suit.

Les quantifieurs optimaux sont utilisés dans le domaine de l'intégration numérique pour approcher les espérances de la forme $\mathbb{E}f(X)$ pour une variable aléatoire X et une fonction continue f. Puisque l'erreur de quantification $\|X - \widehat{X}^\Gamma\|_r$ converge vers 0, alors $\widehat{X}^\Gamma$ converge vers X dans L^r et donc en loi. On obtient donc une approximation de $\mathbb{E}f(X)$ par

$$\mathbb{E}f(\widehat{X}^\Gamma) = \sum_{i=1}^{n} p_i^n f(x_i)$$

où $p_i^n = \mathbb{P}(X \in C_i(\Gamma)), i = 1, \ldots, n$, sont les poids des cellules de Voronoï correspondants au quantifieur optimal $\Gamma = \{x_1, \ldots, x_n\}$ de taille n de X. Des bornes supérieures de l'erreur induite par ce type d'approximation ont été établies pour des quantifieurs optimaux stationnaires, en fonction de la régularité de la fonction f (voir [42, 56, 57]). Par exemple, si f est une fonction continue Lipschitzienne de coefficient de Lipschitz $[f]_{\text{Lip}}$ et Γ est une grille quelconque, alors

$$|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^\Gamma)| \leq [f]_{\text{Lip}} \|X - \widehat{X}^\Gamma\|_1 \leq [f]_{\text{Lip}}\, e_1(X, \Gamma) \leq [f]_{\text{Lip}}\, e_2(X, \Gamma).$$

Si, en outre, Γ est un quantifieur stationnaire pour X ou P, alors, si f est différentiable avec un gradient ∇f α-Hölderien, on a (voir [57] par exemple)

$$|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^\Gamma)| \leq \frac{1}{1+\alpha}[\nabla f]_\alpha \|X - \widehat{X}^\Gamma\|_{1+\alpha}^{1+\alpha}.$$

En particulier, si ∇f est continu Lipschitzien, alors

$$|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^\Gamma)| \leq \frac{1}{2}[\nabla f]_{\text{Lip}} \|X - \widehat{X}^\Gamma\|_2^2.$$

1.1.1 Construction des quantifieurs optimaux

Soit P une loi de probabilité définie sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ et soit X une variable aléatoire de loi P. La plupart des algorithmes conçus pour construire des quantifieurs optimaux (quadratiques) de X ou de sa distribution P sont basés sur la différentiabilité de la fonction de distorsion et sur la propriété de stationnarité (1.7). En fait, la proposition suivante constitue le point de départ des méthodes numériques pour calculer les quantifieurs optimaux dans le cas quadratique.

Proposition 1.1.2. *Soit $X \in L^2(P)$ une variable aléatoire telle que* $\text{card}(\text{supp}(\text{P})) \geq n$. *Toute grille Γ minimisant la fonction de distorsion quadratique G_n^2 associée à X est un quantifieur stationnaire de X.*

En outre, il a été prouvé, dans [39, 73], que si $d = 1$ et si la fonction de densité de probabilité de X est log-concave, alors il existe un quantifieur stationnaire unique de X et ce quantifieur est un minimum global de la fonction de distorsion.

Algorithme de Newton-Raphson Il s'agit d'une procédure déterministe utilisée lorsque la distribution de probabilité est connue explicitement. Supposons que la distribution P de X est absolument continue par rapport à la mesure de Lebesgue avec une densité continue φ. Le quantifieur L^r-optimal est obtenu comme suit : En notant $x = (x_1, \ldots, x_n)$ la grille à construire, on a

$$x^{[l+1]} = x^{[l]} - \left(\nabla^2 G_n^r(x^{[l]})\right)^{-1} \nabla G_n^r(x^{[l]})$$

partant de $x^{[0]}$ appartenant à l'enveloppe convexe du support de X, où $\nabla^2 G_n^r(x)$ est la matrice Hessienne de G_n^r. Ceci peut être amélioré en utilisant l'algorithme de Levenberg-Marquardt

$$x^{[l+1]} = x^{[l]} - \Big(\nabla^2 G_n^r(x^{[l]}) + \lambda_l I_d\Big)^{-1} \nabla G_n^r(x^{[l]})$$

pour un choix approprié des coefficients d'"amortissement" λ_l.

Competitive Learning Vector Quantization (CLVQ) Il s'agit d'un algorithme de descente de gradient stochastique utilisé pour le calcul de quantifieurs d-dimensionnels, $d \geq 1$, également connu sous le nom d'algorithme k-means. En dimension supérieure (toujours dans le cas quadratique), on profite de la représentation de G_n^2 sous forme d'espérance et on présente l'algorithme CLVQ défini par la récursion suivante

$$x^{[l+1]} = x^{[l]} - \gamma_{l+1}\Big(\mathbb{1}_{X\in C_i(x^{([l])})}(x_i^{[l]} - X)\Big)_{1\leq i\leq n}$$

partant de $x^{[0]}$ appartenant à l'enveloppe convexe du support de X, où $(\gamma_l)_{l\geq 1}$ est une suite de paramètres satisfaisant $\sum_{l\geq 1}\gamma_l = +\infty$ et $\sum_{l\geq 1}\gamma_l^2 < +\infty$.

Algorithme de Lloyd Il s'agit d'une recherche de point fixe basée directement sur la propriété de stationnarité. Dans le cas unidimensionnel, il s'agit d'une procédure déterministe utilisée lorsque la distribution de probabilité est connue explicitement et définie par

$$x_i^{[l+1]} = \frac{\mathbb{E}(X\mathbb{1}_{X\in C_i(x^{[l]})})}{P(X\in C_i(x^{[l]}))}.$$

à partir de $x^{[0]}$ appartenant à l'enveloppe convexe du support de X.

Algorithme randomisé de Lloyd En dimension supérieure, la procédure ci-dessus devient non-enviseagable, on passe donc à l'algorithme aléatoire de Lloyd. Les espérances et les probabilités sont calculées par une simulation Monte Carlo de taille M comme suit

$$x_i^{[l+1]} = \frac{\sum_{m=1}^M X^m \mathbb{1}_{X_m\in C_i(x^{[l]})}}{\mathrm{card}(X^m\,;\, X^m\in C_i(x^{[l]}))} \tag{1.8}$$

partant de $x^{[0]}$ appartenant à l'enveloppe convexe du support de X, où $(X^m)_{1\leq m\leq M}$ sont M copies i.i.d. de X.

Pour plus de détails sur les procédures ci-dessus, nous renvoyons à [56, 57]. Notez que des grilles de quantification trés précises de la loi $\mathcal{N}(0; I_q)$ pour des dimensions $d = 1$ à 10 et de tailles régulièrement échantillonnées de $N = 1$ à 1 000 peuvent être téléchargées sur le site de quantification www.quantize.maths-fi.com (à des fins non commerciales).

Pour quelques distributions de probabilité uni-dimensionnelles, il existe des formules fermées ou semi-fermées pour des grilles de quantification optimales. Par exemple, le quantifieur optimal Γ_n de taille n de la loi Uniforme $\mathcal{U}([0,1])$ est donné par

$$\Gamma_n = \Big\{\frac{2i-1}{2n},\ i = 1,\dots,n\Big\},$$

et les formules semi-fermées sont données dans [29, 32] pour les lois exponentielle, puissance, puissance inverse et Laplace.

Dans le cadre bidimensionnel, une approche déterministe pour optimiser les quantifieurs quadratiques est développée dans [52]. Elle repose sur l'approximation d'intégrales bidimensionnelles sur des polygones convexes par des formules de quadrature très efficaces.

Pour les dimensions supérieures, les procédures d'optimisation stochastique peuvent devenir trop coûteuses et trop exigeantes en termes de calcul. Lorsque la loi cible est un produit tensoriel de ses lois marginales indépendantes, on peut s'appuyer sur la quantification produit au lieu des procédures multidimensionnelles standards. Ceci consiste à obtenir des quantifieurs multidimensionnels grâce au produit tensoriel des suites unidimensionnelles, déjà calculées par l'un des algorithmes cités ci-dessus.

1.2 Quantification gloutonne

Lorsque la dimension d augmente, la recherche d'une solution au problème de quantification (1.4) devient plus compliquée et plus exigeante en terme de calcul. D'où la nécessité d' introduire une solution sous-optimale plus facile à calculer, dont le taux de convergence reste similaire à celui des quantifieurs optimaux. Cette solution est fournie par la quantification vectorielle gloutonne.

1.2.1 Principe et résultats

Soit X une variable aléatoire de loi P définie sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. La quantification vectorielle gloutonne a d'abord été introduite et étudiée dans [12] pour des distributions P à support compact (dans un cadre L^1) comme modèle de planification d'expériences à court terme par rapport à la planification d'expériences à long terme représentée par la quantification régulière dans L^1 à un niveau donné n. Elle a ensuite été réintroduite indépendemment et étudiée de manière approfondie dans [45] pour plusieurs classes de lois de probabilité à support non nécessairement borné. Dans les deux cas, elle consiste à déterminer, pour un vecteur (ou une distribution) aléatoire de moment d'ordre r fini, une suite $(a_n)_{n\geq 1}$ dans $\mathbb{R}^d$ qui soit récursivement L^r-optimale étape par étape. En d'autres termes, ayant déjà calculé les n premiers points de la suite $a^{(n)} = \{a_1, \ldots, a_n\}$, on ajoute le $(n+1)$-ième point de la suite comme étant une solution à

$$a_{n+1} \in \operatorname{argmin}_{\xi \in \mathbb{R}^d} e_r(a^{(n)} \cup \{\xi\}, X), \tag{1.9}$$

avec $a^{(0)} = \varnothing$. Notez que a_1 est une/la L^r-médiane de la loi P de X. Une solution à ce problème existe toujours et s'appelle une suite de quantification gloutonne L^r-optimale pour X ou sa loi P. Cependant, cette solution n'est pas nécessairement unique même si la L^r-médiane a_1 l'est. Ceci est dû à la dépendance de la quantification gloutonne de la symétrie de la loi P. Cette existence a été établie en toute généralité dans [45] où les auteurs ont également montré que l'erreur de quantification L^r correspondante est décroissante en fonction du nombre de points n de la suite et qu'elle converge vers 0 lorsque n tend vers l'infini. Le taux optimal de convergence $n^{-\frac{1}{d}}$ de l'erreur de quantification a également été montré dans [45]. Il repose sur l'intégrabilité de la fonction b-maximale associée à la suite de quantification gloutonne L^r-optimale $(a_n)_{n\geq 1}$

définie, pour $b \in (0, \frac{1}{2})$ et $\xi \in \mathbb{R}^d$, par

$$\Psi_b(\xi) = \sup_{n\in\mathbb{N}} \frac{\lambda_d\left(B(\xi, b\,\mathrm{dist}(\xi, a^{(n)}))\right)}{P\left(B(\xi, b\,\mathrm{dist}(\xi, a^{(n)}))\right)}. \tag{1.10}$$

Le théorème ci-dessous traite l'optimalité du taux de convergence de l'erreur de quantification gloutonne dans L^r.

Theorem 1.2.1. *Soit $X \in L^r(P)$, $r \in (0, +\infty)$ et $(a_n)_{n\geq 1}$ une suite de quantification gloutonne L^r-optimale de X. S'il existe $b \in (0, \frac{1}{2})$ tel que la fonction b-maximale $\Psi_b \in L^{\frac{r}{r+d}}(P)$, alors*

$$\limsup_n n^{\frac{1}{d}} e_r(a^{(n)}, X) < +\infty.$$

La fonction b-maximale Ψ_b est également utilisée pour montrer que les suites de quantification gloutonnes satisfont le problème de mismatch de distorsion, c'est-à-dire la propriété que le taux optimal des quantifieurs L^r détient pour les quantifieurs L^s pour $s > r$. Ce problème a déjà été étudié pour les quantifieurs optimaux dans [33] et ensuite dans [65]. Pour les suites de quantification gloutonnes, le théorème suivant, établi dans [45], résout le problème.

Theorem 1.2.2. *Soit $s \in (r, +\infty)$, $X \in L^r(P)$ et $(a_n)_{n\geq 1}$ une suite de quantification gloutonne L^r-optimale de X. Supposons que $\Psi_b \in L^{\frac{s}{r+d}}(P)$ pour $b \in (0, \frac{1}{2})$. Alors, $X \in L^s(P)$ et*

$$\limsup_n n^{\frac{1}{d}} e_s(a^{(n)}, X) < +\infty.$$

Construction des suites de quantification gloutonne

Les suites de quantification gloutonnes sont construites par des variantes d'algorithmes habituels de construction des quantifieurs optimaux, tels l'algorithme de Lloyd ou l'algorithme CLVQ, mais de manière récursive. Cela signifie qu'à chaque itération de l'algorithme, on ajoute un seul point aux points de la suite déjà calculés, puis on met en œuvre une procédure d'optimisation en gelant les autres points calculés précédemment. Nous donnons une brève idée sur cette procédure dans le cas quadratique lorsque $d = 1$ et $d \geq 2$.

Cadre unidimensionnel Lorsque $d = 1$ et la distribution de X est absolument continue avec une fonction de densité positive et continue φ, on peut mettre en œuvre des procédures déterministes basées sur la connaissance de la fonction de répartition F_X et de la fonction de moment de premier ordre K_X de la loi de X. L'idée est la suivante : à la n-ième itération, nous gelons les $n-1$ points $a^{(n-1)} = \{a_1, \ldots, a_{n-1}\}$ de la suite $(a_n)_{n\geq 1}$ qui ont déjà été calculés et nous les trions par ordre croissant

$$a_1^{(n-1)} < \ldots < a_{n-1}^{(n-1)}.$$

Ensuite, nous calculons les inerties locales inter-points données par

$$\sigma_i^2 := \int_{a_i^{(n-1)}}^{a_{i+\frac{1}{2}}^{(n-1)}} |a_i^{(n-1)} - \xi|^2 \mu(d\xi) + \int_{a_{i+\frac{1}{2}}^{(n-1)}}^{a_{i+1}^{(n-1)}} |a_{i+1}^{(n-1)} - \xi|^2 \mu(d\xi), \quad i = 0, \ldots, n-1,$$

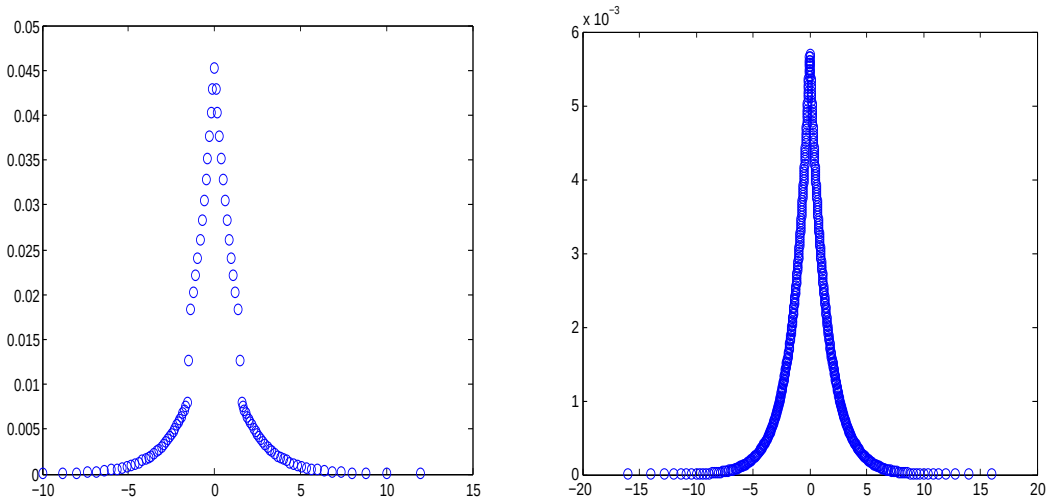

Figure 1.2: Graphe de $a_i \mapsto p_i^n$ où $a^{(n)}$ est une suite de quantification gloutonne L^2-optimale de Laplace$(0,1)$ et $(p_i^n)_{1\leq i\leq n}$ sont les poids des cellules de Voronoï correspondants pour $n = 100$ (gauche) et $n = 511$ (droite).

où $a_0^{(n-1)} = -\infty$, $a_n^{(n-1)} = +\infty$ et $a_{i+\frac{1}{2}}^{(n-1)}$ est le milieu du segment $[a_i^{(n-1)}, a_{i+1}^{(n-1)}]$:

$$a_{\frac{1}{2}}^{(n-1)} = -\infty, \quad a_{i+\frac{1}{2}}^{(n-1)} = \frac{a_i^{(n-1)} + a_{i+1}^{(n-1)}}{2}, \quad a_{n-\frac{1}{2}}^{(n-1)} = +\infty.$$

Nous ajoutons un point aléatoire $\bar{a}_0$ dans la zone interpoint d'inertie locale maximale $(a_{i_0}^{(n-1)}, a_{i_0+1}^{(n-1)})$ où i_0 est l'indice tel que

$$\sigma_{i_0}^2 = \max_{0\leq i\leq n-1} \sigma_i^2.$$

Ce point $\bar{a}_0$ est le point de départ de la procédure d'optimisation considérée, qui converge vers le n-ième point a_n de la suite. Plusieurs procédures sont détaillées dans la première partie du chapitre 4 telles que l'algorithme de Lloyd et l'algorithme CLVQ, et des suites de quantification gloutonnes de plusieurs lois de probabilité unidimensionnelles sont construites par l'algorithme de Lloyd. Par exemple, la figure 1.2 représente les poids des cellules de Voronoï des n premiers points $(n = 100, n = 511)$ d'une suite de quantification gloutonne quadratique de la loi de Laplace de paramètres 0 et 1.

Cadre bidimensionnel Nous étendons la variante déterministe des algorithmes gloutons au cas bidimensionnel dans le chapitre 4. Nous suivons la même procédure que pour les lois unidimensionnelles et nous nous appuyons sur une formule de quadrature très efficace pour calculer numériquement les intégrales nécessaires à la construction des suites de quantification gloutonnes. Dans la figure 1.3, nous observons une suite de quantification gloutonne quadratique déterministe de la loi Gaussienne standard $\mathcal{N}(0, I_2)$.

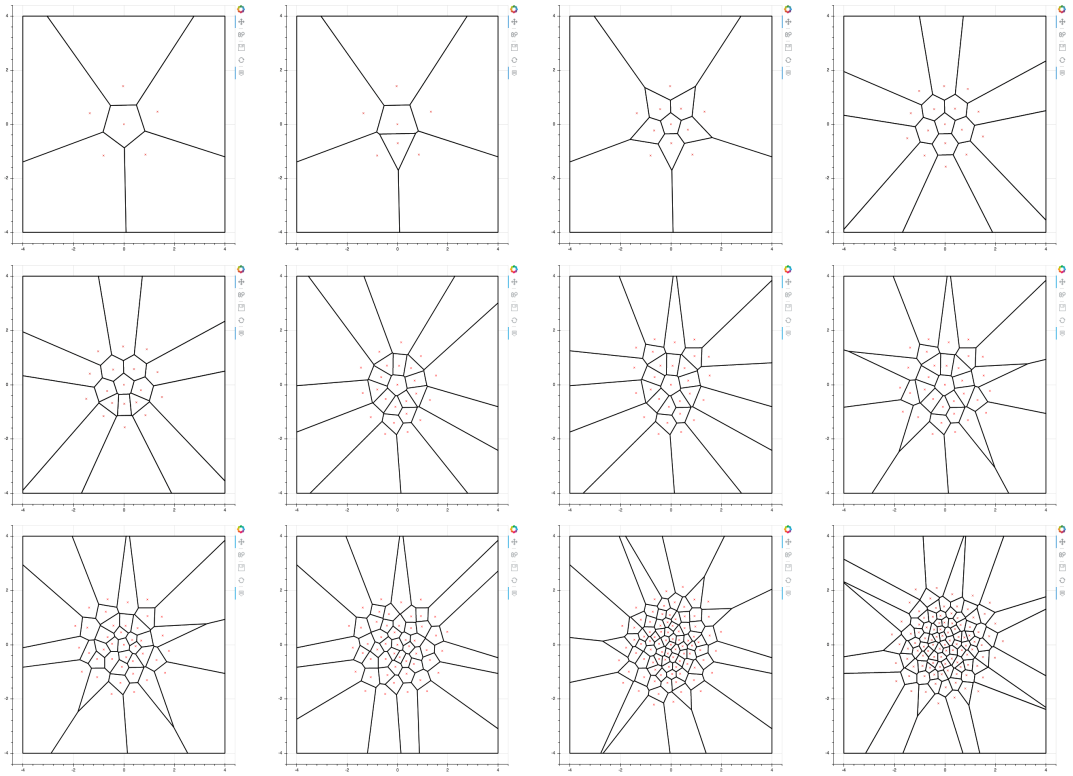

Figure 1.3: Suite de quantification gloutonne de $\mathcal{N}(0, I_2)$ obtenue par un algorithme de Lloyd déterministe de tailles $n = 6, 7, 11, 16, 18, 24, 28, 32, 39, 51, 86, 100$ (à partir du haut à gauche).

Cadre mutlidimensionnel En dimension $d > 2$, les procédures déterministes deviennent trop exigeantes, on passe donc à des procédures stochastiques où le calcul des intégrales est remplacé par des simulations de Monte Carlo couplées de recherches du plus proche voisin. Les versions gloutonnes de l'algorithme stochastique de Lloyd et de l'algorithme multidimensionnel CLVQ sont expliquées en détail dans le chapitre 4.
Ces procédures peuvent être très coûteuses en raison des nombreuses intégrales à calculer. Bien que nous expliquons, dans le chapitre 3, comment la quantification gloutonne permet une réduction du nombre de calculs à chaque étape, cela ne rend toujours pas les procédures d'optimisation stochastique faciles à mettre en œuvre. Une alternative est la quantification gloutonne produit où l'on s'appuie sur des suites de quantification gloutonne unidimensionnelles pour calculer des suites multidimensionnelles lorsque la loi de probabilité s'écrit comme un produit tensoriel de ses lois marginales. La suite multidimensionnelle est obtenue comme résultat du produit tensoriel de plusieurs suites unidimensionnelles. Ceci est expliqué en détail dans les chapitres 3 et 4.

1.2.2 Contributions et nouveaux résultats

La première contribution de cette thèse, dans le chapitre 3, est consacrée à l'extension de certains résultats théoriques de la quantification gloutonne d'une loi de probabilité P de moment d'ordre r fini à une classe plus large de distributions, principalement des résultats d'optimalité du taux de convergence et de mismatch de distorsion. Une étude numérique approfondie est également menée pour mettre en évidence les avantages des suites de quantification gloutonne, comparées principalement aux méthodes de Monte Carlo et de quasi-Monte Carlo. Dans le chapitre 5,

des résultats d'optimalité du taux de convergence dans L^s des suites dilatées de quantification gloutonnes L^r-optimales sont établis, inspirés par des résultats similaires pour les quantifieurs optimaux dilatés dans [71].

Optimalité du taux de convergence de l'erreur de quantification (chapitre 3)

Comme déjà mentionné dans la section 1.2.1, des résultats sur le taux de convergence de l'erreur de quantification gloutonne et le problème de mismatch de distorsion sont établis dans [45]. Ils sont basés sur l'intégrabilité de la fonction b-maximale Ψ_b définie par (1.10). Dans le chapitre 3, basé sur l'article soumis [24], nous étendons ces résultats à une classe de distributions beaucoup plus large.

Soit X une variable aléatoire de loi P définie sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Le principal outil de notre étude est de considérer des distributions de probabilité auxiliaires ν satisfaisant le contrôle suivant sur les boules par rapport à une L^r-médiane a_1 de P : Supposons qu'il existe $\varepsilon_0 \in (0,1]$ tel que, pour tout $\varepsilon \in (0, \varepsilon_0)$, il existe une fonction Borélienne $g_\varepsilon : \mathbb{R}^d \to [0, +\infty]$ satisfaisant, pour tout $x \in \mathrm{supp}(P)$ et tout $t \in [0, \varepsilon \|x - a_1\|]$,

$$\nu(B(x,t)) \geq g_\varepsilon(x) V_d \, t^d. \tag{1.11}$$

Cette classe de distributions auxiliaires sera l'outil principal pour diverses études théoriques des suites de quantification gloutonnes. En notant que la L^r-méédiane a_1 de P appartient à $a^{(n)}$ pour tout $n \geq 1$ (par construction de la suite de quantification gloutonne), on obtient une borne supérieure de la forme

$$\forall n \geq 2, \quad e_r(a^{(n)}, P) \leq \varphi_r(\varepsilon)^{-\frac{1}{d}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{1}{d}} \left(\int g_\varepsilon^{-\frac{r}{d}} dP\right)^{\frac{1}{r}} (n-1)^{-\frac{1}{d}} \tag{1.12}$$

où V_d est le volume de la boule unité et $\varphi_r(u) = \left(\frac{1}{3^r} - u^r\right) u^d$. La preuve de ce résultat repose sur une nouvelle micro-macro inégalité impliquant les distributions auxiliaires ν.

L'une des principales contributions présentées dans le chapitre 3 est l'extension des résultats universels non-asymptotiques de type Pierce (1.6) pour le taux de convergence de l'erreur de quantification L^r-gloutonne. Ceci est obtenu en spécifiant la mesure ν et la fonction g_ε, satisfaisant (1.11), dans la borne supérieure (1.12). Par exemple, on cite la borne supérieure suivante pour l'erreur de quantification dans L^r : Si $\int |x|^{r+\delta} dP(x) < +\infty$ pour $\delta > 0$, alors pour tout $n \geq 2$,

$$e_r(a^{(n)}, P) \leq \kappa_{d,\delta,r}^{\text{Greedy,Pierce}} \sigma_{r+\delta}(P) (n-1)^{-\frac{1}{d}},$$

où $\kappa_{d,\delta,r}^{\text{Greedy, Pierce}}$ est une constante finie définie dans le théorème 3.2.4, qui repose sur la mesure $\nu(dx) = \gamma_{r,\delta}(x) \lambda_d(dx)$ où

$$\gamma_{r,\delta}(x) = \frac{K_{\delta,r}}{(1 \vee |x - a_1|)^{d(1+\frac{\delta}{r})}} \qquad \text{et} \qquad K_{\delta,r} = \left(\int \frac{dx}{(1 \vee |x|)^{d(1+\frac{\delta}{r})}}\right)^{-1} < +\infty,$$

et la fonction

$$g_\varepsilon(x) = \frac{K_{\delta,r}}{(1 \vee [(1+\varepsilon)|x - a_1|])^{d(1+\frac{\delta}{r})}}, \quad \varepsilon \in (0, \tfrac{1}{3}).$$

Un résultat plus précis, mais moins explicite en termes de constantes, est ensuite énoncé (voir le théorème 3.2.4) en se basant sur des lois de probabilités vérifiant une propriété de log-intégrabilité de la forme $\int_{\mathbb{R}^d} |x|^r (\log^+|x|)^{\frac{r}{d}+\delta} dP(x) < +\infty$.

Enfin, un résultat hybride Zador-Pierce est démontré pour des densités radiales, c'est-à-dire une borne supérieure non asymptotique (de type Pierce (1.6)) dépendant de $\|h\|_{\frac{d}{d+r}}$ comme dans le théorème de Zador (1.5). En d'autres termes, nous établissons une borne supérieure de la forme

$$e_r(a^{(n)}, P) \leq C\|h\|_{\frac{d}{d+r}} n^{-\frac{1}{d}}$$

pour une constante réelle C où h est la densité de la composante absolument continue de P, censée être radiale. Une fonction $f : \mathbb{R}^d \to \mathbb{R}_+$ est dite radiale non croissante sur un ensemble A, telle que $\text{supp}(P) \subset A \subset \mathbb{R}^d$, par rapport à $a \in A$ (voir définition 3.2.6), s'il existe une norme $\|\cdot\|_0$ sur $\mathbb{R}^d$ et une constante réelle $M \in (0, 1]$ telle que

$$f(y) \geq Mf(x) \qquad \text{pour tout } x, y \in A \setminus \{a\} \quad \text{tels que} \quad \|y - a\|_0 \leq \|x - a\|_0.$$

Le résultat est obtenu en considérant

$$\nu = \frac{h^{\frac{d}{d+r}}}{\int h^{\frac{d}{d+r}} d\lambda_d}.\lambda_d$$

et en se basant sur une borne inférieure de $\nu(B(x, t))$, où $B(x, t)$ est la boule de centre $x \in \mathbb{R}^d$ et de rayon $t > 0$, établie dans le Lemme 3.2.10 du chapitre 3.

Le problème de mismatch de distorsion pour cette classe plus large de lois de probabilités est également résolu en considérant les mêmes distributions auxiliaires définies par (1.11). Les résultats sont donnés dans la Section 3.3 et on cite l'erreur suivante pour $s \in (r, d + r)$.

$$e_s(a^{(n)}, P) \leq \kappa_{d,r,\varepsilon}^{\text{Greedy}} \left(\int g_\varepsilon^{-\frac{s}{d+r-s}} dP \right)^{\frac{d+r-s}{s(d+r)}} \left(\int g_\varepsilon^{-\frac{r}{d}} dP \right)^{\frac{1}{d+r}} (n-2)^{-\frac{1}{d}}$$

pour $n \geq 3$ et une constante positive finie $\kappa_{d,r,\varepsilon}^{\text{Greedy}}$ définie plus tard dans le théorème 3.3.1.

Algorithmes et observations numériques

Dans la deuxième partie du chapitre 3 et dans le chapitre 4, plusieurs expériences numériques sont réalisées afin de mettre en évidence certaines propriétés intéressantes des suites de quantification gloutonnes unidimensionnelles. Entre autres, nous concluons numériquement que, même si les suites gloutonnes dans L^r ne sont pas optimales à chaque niveau n, elles peuvent néanmoins être sous-optimales dans le sens où il existe des sous-suites de $a^{(n)}$ qui sont elles-mêmes L^r-optimales. Cela a été déduit en observant les graphiques représentant les poids des cellules de Voronoï de ces suites. Nous spécifions ces sous-suites pour les lois $\mathcal{N}(0, 1)$ et $\mathcal{U}([0, 1])$ et concluons par une conjecture concernant les densités unimodales symétriques par rapport à leur médiane.

D'un autre point de vue, lorsque l'on travaille sur le cube unité, il est naturel de comparer des suites de quantification gloutonnes à des suites à discrépance faible utilisées dans les méthodes

quasi-Monte Carlo (QMC). En fait, avec l'intégration numérique basée sur la quantification, on approche des espérances de la forme $\mathbb{E}f(X)$, pour une fonction continue Lipschitzienne f et une variable aléatoire X, avec un taux de convergence de $\mathcal{O}(n^{-\frac{1}{d}})$. Alors que l'approximation par la méthode quasi-Monte Carlo donne un taux de convergence de $\mathcal{O}\Big(\frac{\log n}{n^{\frac{1}{d}}}\Big)$, ceci est dû au théorème de Proïnov (voir [70] ou théorème 3.4.1 au chapitre 3). Le prix à payer pour l'absence du facteur $(\log n)$ avec la quantification gloutonne est le fait que les poids des cellules de Voronoï correspondants à la suite gloutonne $a^{(n)}$ ne sont pas uniformes (c'est-à-dire égal à $\frac{1}{n}$) ce qui induit une plus grande complexité lors de la mise en œuvre "naïve" des formules de quadrature résultantes. Nous montrons, dans la deuxième partie du chapitre 3, que la récursivité de la quantification gloutonne permet de réduire le nombre de calculs afin que la quantification gloutonne et QMC deviennent comparables en termes de complexité. De plus, ce caractère permet de conserver l'atout d'une suite qui est une formule récursive pour les cubatures, faisant ainsi de la quantification gloutonne une composante avantageuse face aux méthodes Quasi-Monte Carlo.

Pour être plus précis, lors de la procédure de construction de la suite gloutonne, on remarque qu'à chaque itération, on ajoute un seul point à la suite alors que les autres restent gelés. Ainsi, les cellules de Voronoï, qui sont loin du nouveau point ajouté, restent intactes et inchangées. Cela signifie que leurs poids, ainsi que l'inertie locale inter-points correspondante, n'ont pas besoin d'être calculés à chaque itération. Cette remarque permet d'éviter un grand nombre de calculs inutiles à chaque itération de l'algorithme. Outre la réduction importante du coût de calcul, ce caractère récursif de la quantification gloutonne nous amène à déduire une formule itérative récursive pour la cubature dans les cadres unidimensionnels et multidimensionnels. Lorsque $d = 1$, nous approchons $\mathbb{E}f(X)$ par $I_n(f)$ donné par

$$I_n(f) \quad = \quad I_{n-1}(f) - p_-^n\Big(f(a_{i_0-1}^{(n)}) - f(a_{i_0}^{(n)})\Big) - p_+^n\Big(f(a_{i_0+1}^{(n)}) - f(a_{i_0}^{(n)})\Big),$$

où $a_{i_0}^{(n)}$ est le point ajouté à la suite gloutonne à l'itération n, $a_{i_0-1}^{(n)}$ et $a_{i_0+1}^{(n)}$ sont les points inférieur et supérieur à $a_{i_0}^{(n)}$ et

$$p_-^n = P\Big([a_{i_0-\frac{1}{2}}^{(n)}, a_{\text{mil}}^{(n)}]\Big) \qquad \text{et} \qquad p_+^n = P\Big([a_{\text{mil}}^{(n)}, a_{i_0+\frac{1}{2}}^{(n)}]\Big)$$

où $a_{i_0\pm\frac{1}{2}}^{(n)} = \frac{a_{i_0}^{(n)} + a_{i_0\pm1}^{(n)}}{2}$ et $a_{\text{mil}}^{(n)} = \dfrac{a_{i_0+1}^{(n)} + a_{i_0-1}^{(n)}}{2}$, avec $a_0 = -\infty$ et $a_n = +\infty$.
Cette formule est généralisée au cadre multidimensionnel (voir (3.20)) lorsque l'on considère des suites gloutonnes "produit", comme expliqué dans le chapitre 3.

Par ailleurs, on note qu'il existe une relation entre la discrépance d'une suite Ξ et l'erreur de quantification induite par cette suite par rapport à la loi Uniforme. Basée sur le théorème de Proïnov (voir [70] et théorème 3.4.1 dans le chapitre 3), elle est donnée par

$$e_1(\Xi, \mathcal{U}([0,1]^d)) \leq D_n^*(\Xi)^{\frac{1}{d}}$$

où $D_n^*(\Xi)$ est la discrépance à l'origine de la suite $\Xi = (\xi_i)_{1\leq i\leq n}$ d'ordre n définie par

$$D_n^*(\Xi) = \sup_{u\in[0,1]^d} \Big|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\xi_i\in[0,u]^d} - \lambda_d([0,u]^d)\Big|.$$

Cela nous conduit à réaliser une étude en vue d'une comparaison entre les suites de quantification gloutonnes et les suites à discrépance faible. Deux approches principales sont considérées:

- Calculer la discrépance des suites gloutonnes et la comparer à celles des suites à discrépance faible,
- Traiter les suites à discrépance faible comme des suites de quantification (sous-optimales), c'est-à-dire leur attribuer un diagramme de Voronoï et des poids non uniformes, afin de comparer leurs performances avec des suites de quantification gloutonnes.

Plusieurs simulations numériques sont effectuées et certaines conclusions sont tirées et détaillées à la fin du chapitre 3 et dans le chapitre 4. Disons en bref que, lorsque $d = 1$, les suites de quantification gloutonnes peuvent être utilisées comme suites à discrépance faible, et, elles sont plus performantes que les suites à discrépance faible traitées comme des suites de quantification. Cependant, lorsque $d \geq 2$, nous n'avons pas des résultats aussi optimistes pour les suites de quantification gloutonnes standard en termes de discrépance faible.

Taux de convergence L^s-optimal des suites de quantification gloutonne L^r-optimales dilatées/contractées

Dans le chapitre 5, nous étudions l'optimalité du taux de convergence dans L^s des suites dilatées/contractées de quantification gloutonne dans L^r. Cette étude s'inspire de résultats similaires obtenus pour les quantifieurs L^r-optimaux dans [71], où les quantifieurs L^r-optimaux, une fois dilatés ou contractés de manière appropriée, s'avèrent avoir un taux de convergence L^s-optimal, c'est-à-dire de $\mathcal{O}(n^{-\frac{1}{d}})$, pour $s > r$. Cela a des conséquences importantes en pratique puisque, généralement, on n'a accès qu'à des quantifieurs quadratiques optimaux (comme pour la loi $\mathcal{N}(0, I_d)$, $d = 1, \ldots, 10$, sur le site web de quantification www.quantize.maths-fi.com ou pour d'autres lois (1D) pour lesquelles des formules semi-fermées sont disponibles (voir [29, 32] par exemple)). On peut citer, d'une part, le domaine de l'intégration numérique où les bornes d'erreur des formules de cubature basées sur la quantification impliquent souvent l'erreur L^s de quantification induite par les quantifieurs L^r-optimaux, $s > r$, qui doit être traitée. D'autre part, les quantifieurs L^r-optimaux dilatés sont de bons candidats pour l'initialisation des algorithmes de conception des suites de quantification L^s-optimales (voir [71] pour plus de détails sur ce sujet).

Le but du chapitre 5 est d'établir des résultats similaires pour les suites de quantification gloutonnes L^r-optimales. Pour cela, nous nous appuyons sur des distributions auxiliaires, vérifiant un critère similaire à celui donné par (1.11). Également, nous généralisons les résultats séminaux de [71] en s'appuyant sur notre nouvelle approche basée sur des fonctions auxiliaires. Soyons plus précis.

Soit X une variable aléatoire de loi P définie sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ et soit $a^{(n)}$ une suite de quantification gloutonne L^r-optimale de taille n, $r \geq 1$. La suite dilatée ou contractée correspondante est notée $a^{(n)}_{\theta,\mu}$ et définie, pour tout $\theta > 0$ et $\mu \in \mathbb{R}^d$, par $a^{(n)}_{\theta,\mu} = \{\mu + \theta(a_i - \mu),\ ; a_i \in a^{(n)}\}$. De même, $f_{\theta,\mu}$ désigne la fonction $f_{\theta,\mu}(x) = f(\mu + \theta(x - \mu))$. Et, si $X \sim P = f.\lambda_d$, alors $P_{\theta,\mu}$ désigne la loi de la variable aléatoire $\frac{X-\mu}{\theta} + \mu$ et $dP_{\theta,\mu} = \theta^d f_{\theta,\mu}.d\lambda_d$. Nous nous appuyons sur une inégalité micro-macro impliquant des distributions auxiliaires satisfaisant le contrôle (1.11) sur les boules pour obtenir deux principaux résultats non asymptotiques d'optimalité du taux

de convergence dans L^s en fonction de s.

• Soit $s \in (r, d+r)$ et P une loi ayant des moments polynomiaux finis de tout ordre. Supposons

$$\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{\frac{(d+r)(r+\delta-\eta)}{(d+r-s)(r+\delta-\eta)-ds}} f d\lambda_d < +\infty. \tag{1.13}$$

Alors, pour toute fonction Borélienne g_ε, $\varepsilon \in (0, \frac{1}{3})$, vérifiant (1.11) et tout $n \geq 3$,

$$e_s(a^{(n)}_{\theta,\mu}, P) \leq \theta^{1+\frac{d}{s}} \kappa^{\text{Greedy,Pierce}}_{\theta,\mu} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{\frac{(d+r)(r+\delta-\eta)}{(d+r-s)(r+\delta-\eta)-ds}} f d\lambda_d \right)^{\frac{1}{|q|q'(d+r)}} \sigma_{r+\delta}(P)(n-2)^{-\frac{1}{d}}. \tag{1.14}$$

où $q = \frac{-s}{d+r-s}$, $q' = \frac{r+\delta-\eta}{r+\delta-\eta-d|q|}$, $p' = \frac{q'}{q'-1}$, $e_{r+\delta}(a^{(1)}, P) = \sigma_{r+\delta}(P) < +\infty$ et

$$\kappa^{\text{Greedy,Pierce}}_{\theta,\mu} = 2^{\frac{1}{d}+\frac{r+\delta}{r+d}(1+\frac{1}{|q|p'})} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} \min_{\varepsilon\in(0,\frac{1}{3})} \left[(1+\varepsilon)\varphi_r(\varepsilon)^{-\frac{1}{d}}\right] \left(\int (1 \vee |x|)^{\frac{r+\delta}{r+\delta-\eta}} dx \right)^{\frac{1}{d}}.$$

• Soit $s < r$. Supposons

$$\int f^{-\frac{s}{r-s}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d < +\infty.$$

Alors, pour toute distribution ν et toute fonction g_ε satisfaisant (1.11) et tout $n \geq 3$,

$$e_s(a^{(n)}_{\theta,\mu}, P) \leq \widetilde{\kappa}^{\text{Greedy,Pierce}}_{\theta,\mu} \theta^{1+\frac{d}{s}} \left(\int_{\{f>0\}} f^{-\frac{s}{r-s}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d \right)^{\frac{r-s}{sr}} \sigma_{r+\delta}(P)(n-2)^{-\frac{1}{d}} \tag{1.15}$$

où $e_{r+\delta}(a^{(1)}, P) = \sigma_{r+\delta}(P) < +\infty$ et

$$\widetilde{\kappa}^{\text{Greedy,Pierce}}_{\theta,\mu} = 2^{1+\frac{1}{d}+\frac{\delta}{r}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} \min_{\varepsilon\in(0,\frac{1}{3})} \left[(1+\varepsilon)\varphi_r(\varepsilon)^{-\frac{1}{d}}\right] \left(\int (1 \vee |x|)^{-d(1+\frac{\delta}{r})} dx \right)^{-\frac{1}{d}}.$$

Les résultats ci-dessus sont des avatars du Lemme de Pierce (1.6). Une étude particulière pour des densités radiales donne des résultats similaires sous une hypothèse de moment sur P.

Après avoir montré qu'une suite dilatée ou contractée de quantification gloutonne L^r-optimale $a^{(n)}_{\theta,\mu}$ a un taux de convergence optimal dans L^s sous l'une des conditions mentionnées ci-dessus, en fonction des valeurs de s, nous déterminons l'ensemble des paramètres (θ, μ) qui satisfait ces conditions. En général, la valeur optimale pour μ^* est la L^r-médiane de la loi P. Quant à θ, le problème dépend entièrement de la loi P. Nous menons cette étude pour plusieurs lois de probabilité et déterminons, pour chacune d'entre elles, les valeurs de θ pour lesquelles la suite dilatée a un taux de convergence optimal dans L^s. De plus, dans certains cas, nous montrons que la suite $\alpha^{(n)}_{\theta^*,\mu}$ satisfait le "théorème de la mesure empirique" pour une valeur particulière θ^* de θ qui sera déterminée. Cette valeur particulière θ^* permet à la borne inférieure (5.6) induite par $\alpha^{(n)}_{\theta^*,\mu}$ d'atteindre la constante du théorème de Zador.

Pour la loi Normale d-dimensionnelle $\mathcal{N}(m, \Sigma_d)$, cette valeur est $\theta^* = \sqrt{\frac{s+d}{r+d}}$, pour les lois hyper exponentielles de paramètres α et λ, on obtient $\theta^* = \left(\frac{s+d}{r+d}\right)^{\frac{1}{\alpha}}$ et, pour la loi hyper gamma de paramètres α, β et λ, la suitee dilatée/contractée satisfait le théorème de la mesure empirique pour $\theta^* = \left(\frac{s+d}{r+}\right)^{\frac{1}{\alpha}}$, uniquement lorsque $\beta = \frac{d+r}{d(d+s)}$.

1.3 Quantification récursive et application aux E.D.S. rétrogrades réfléchies

1.3.1 Principe et résultats existants

La quantification markovienne et la quantification récursive ont été initialement introduites dans [59] et [63] pour produire des schémas de discrétisation spatiale des chaînes de Markov, typiquement des schémas de discrétisation temporelle de processus stochastiques comme les processus de diffusion. La quantification récursive est une version de la quantification markovienne qui permet en dimension 1, mais aussi en dimensions moyennes, une "optimisation déterministe" rapide des grilles de quantification impliquées dans ces schémas numériques. Elle a d'abord été étudiée en profondeur dans [63] pour la discrétisation d'un schéma d'Euler à valeurs dans $\mathbb{R}^d$ d'un processus de diffusion où les auteurs ont proposé un algorithme rapide pour construire, de manière déterministe, l'arbre de quantification dans un cadre unidimensionnel. Dans [51], la quantification récursive a été étendue à des schémas d'ordre supérieur, toujours dans le cadre unidimensionnel. Pour les problèmes en dimension supérieure, la quantification récursive produit a été introduite et utilisée dans [15, 28] entre autres.

Considérons un processus de diffusion Brownien $(X_t)_{t\geq 0}$ à valeurs dans $\mathbb{R}^d$ et solution de

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s\,, \qquad X_0 = x_0 \in \mathbb{R}^d, \tag{1.16}$$

où $b : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ est le coefficient de dérive, $\sigma : [0,T] \times \mathbb{R}^d \to \mathcal{M}(d,q)$ est le coefficient de la matrice de diffusion et $(W_t)_{t\geq 0}$ est un mouvement Brownien q-dimensionnel défini sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ équipé de sa filtration naturelle augmentée $(\mathcal{F}_t)_{t\geq 0}$ où $\mathcal{F}_t = \sigma(W_s, s \leq t, \mathcal{N}_{\mathbb{P}})$, $\mathcal{N}_{\mathbb{P}}$ désigne la classe de tous les ensembles $\mathbb{P}$-négligeables de $\mathcal{A}$. Le schéma d'Euler associé au processus $(X_t)_{t\in[0,T]}$, de maillage uniforme $t_k = k\Delta$, $k \in \{0, \ldots, n\}$ et de pas de temps $\Delta = \frac{T}{n}$, est défini récursivement par

$$\bar{X}^n_{t_{k+1}} = \bar{X}^n_{t_k} + \Delta b(t_k, \bar{X}^n_{t_k}) + \sigma(t_k, \bar{X}^n_{t_k})(W_{t_{k+1}} - W_{t_k}), \qquad \bar{X}^n_{t_0} = X_0 = x_0 \in \mathbb{R}^d. \tag{1.17}$$

La quantification récursive consiste à construire une chaîne de Markov à valeurs dans une grille (ou quantifieur) Γ_k de taille N_k du schéma d'Euler discret $\bar{X}_{t_k}$ au temps t_k. Notre objectif est donc d'optimiser les grilles Γ_k de manière récursive, de sorte que cette optimisation est effectuée "pas à pas" à partir du temps $t_0 = 0$ jusqu'au temps $t_n = T$. Premièrement, nous indiquons par

$$F_k(x, \varepsilon_{k+1}) = x + \Delta b(t_k, x) + \sqrt{\Delta}\sigma(t_k, x)\varepsilon_{k+1}$$

l'opérateur d'Euler de pas de temps Δ, où $(\varepsilon_k)_{0\leq k\leq n}$ est une suite de variables aléatoires i.i.d. de loi $\mathcal{N}(0, I_q)$, en d'autres termes, $\varepsilon_k = \sqrt{\frac{n}{T}}(W_{t_{k+1}} - W_{t_k})$. Notez que ce processus est de loi Normale

$$F_k(x, \varepsilon_{k+1}) \sim \mathcal{N}(m_k, \Sigma_k)$$

où $m_k = x + \Delta b(t_k, x)$ et $\Sigma_k = \sqrt{\Delta}\sigma(t_k, x)$. La quantification récursive $(\widehat{X}^{\Gamma_k}_{t_k})_{0\leq k\leq n}$ de $(\bar{X}_{t_k})_{0\leq k\leq n}$ est effectuée par la récursion suivante : A partir de $\widehat{X}_{t_0} = \bar{X}_{t_0} = x_0$,

$$\begin{cases} \widetilde{X}_{t_k} &= F_{k-1}(\widehat{X}^{\Gamma_{k-1}}_{t_{k-1}}, \varepsilon_k), \\ \widehat{X}^{\Gamma_k}_{t_k} &= \mathrm{Proj}_{\Gamma_k}(\widetilde{X}_{t_k}), \qquad \forall k = 1, \ldots, n \end{cases} \tag{1.18}$$

où Γ_k est un quantifieur optimal de $\widetilde{X}_{t_k}$ de taille N_k pour tout $k \in \{1, \dots, n\}$.

Des bornes supérieures de l'erreur de quantification induite par l'approximation de $\bar{X}_{t_k}$ par $\widehat{X}_{t_k}^{\Gamma_k}$ sont établies dans [63] dans le cadre quadratique où les auteurs ont montré que, sous certaines hypothèses de continuité Lipschitzienne (en x, uniformément en $t \in [0,1]$) sur b et σ, on a, pour chaque $k \in \{0, \dots, n\}$,

$$\|\bar{X}_{t_k} - \widehat{X}_{t_k}^{\Gamma_k}\|_2 \leq K \sum_{l=1}^{k} c_l N_l^{-\frac{1}{d}}$$

pour des constantes positives finies K et c_l. Par souci de simplicité, on note $\widehat{X}_{t_k}$ au lieu de $\widehat{X}_{t_k}^{\Gamma_k}$.

La construction de quantifieurs récursifs $\widehat{X}_{t_k}$ de $\bar{X}_{t_k}$ est principalement réduite au calcul des grilles de quantification optimales Γ_k de $\widetilde{X}_{t_k}$ de taille N_k. Dans le cadre quadratique, ceci est effectué par des algorithmes déterministes standards, tels l'algorithme de Lloyd ou le CLVQ. Dans cette thèse, nous utiliserons principalement l'algorithme de Lloyd pour calculer les grilles $(\Gamma_k)_{1 \leq k \leq n}$ de manière récursive. En effet, à l'instant t_{k+1}, la grille $\Gamma_{k+1} = \{x_1^{k+1}, \dots, x_{N_{k+1}}^{k+1}\}$ est construite en fonction de la grille $\Gamma_k = \{x_1^k, \dots, x_{N_k}^k\}$ déjà obtenue à l'instant t_k.

Le principal avantage de cette approche est la préservation de la propriété de Markov. La loi de la chaîne de Markov $(\widehat{X}_{t_k})_{0 \leq k \leq n}$ est entièrement caractérisée par la loi initiale et les matrices de transition $P_k = (p_{ij}^k)_{i,j}$, pour tout $k \in \{1, \dots, n\}$, qui constituent un outil très important dans plusieurs applications. La probabilité de transition de $(\widehat{X}_{t_k})_{0 \leq k \leq n}$ de x_i^k à x_j^{k+1} entre les temps t_k et t_{k+1} est donnée par

$$p_{ij}^k = \mathbb{P}\left(\widetilde{X}_{t_{k+1}} \in C_j(\Gamma_{k+1}) \,|\, \widetilde{X}_{t_k} \in C_i(\Gamma_k)\right) = \mathbb{P}\left(F_k(x_i^k, \varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\right)$$

où $\big(C_i(\Gamma_k)\big)_{1 \leq i \leq N_k}$ est le diagramme de Voronoï associé au quantifieur Γ_k à l'instant t_k. Les poids des cellules de Voronoï $(p_j^{k+1})_{1 \leq j \leq N_{k+1}}$ sont obtenus par l'équation de Kolmogorov classique (temps discret). Pour tout $j \in \{1, \dots, N_{k+1}\}$, on a

$$p_j^{k+1} = \mathbb{P}(\widetilde{X}_{t_{k+1}} \in C_j(\Gamma_{k+1})) = \sum_{i=1}^{N_k} p_i^k \mathbb{P}\Big(F_k(x_i^k, \varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\Big).$$

Dans le cadre unidimensionnel, le calcul des poids de transition p_{ij}^k est basé sur la fonction de répartition de la loi Gaussienne. Lorsque la dimension d augmente, on s'appuie sur les simulations de Monte Carlo pour ces calculs.

Dans le chapitre 7, nous donnons des détails sur le calcul des suites de quantification récursive de modèles spécifiques dans le cas unidimensionnel, comme le modèle de Black-Scholes et le modèle CEV, discrétisés selon un schéma d'Euler ou un schéma de Milstein. Dans la figure 1.4, on illustre les fonctions $x_i^k \mapsto p_i^k, k = 1, \dots, n$, où $(x_i^k)_{1 \leq i \leq N_k}$ est la grille de quantification récursive d'un processus de diffusion suivant un modèle de Black-Scholes et discrétisé suivant un schéma d'Euler, c'est-à-dire

$$\bar{X}_{t_{k+1}} = \bar{X}_{t_k} + r\Delta \bar{X}_{t_k} + \sigma\sqrt{\Delta}\bar{X}_{t_k}\,\varepsilon_{k+1} := F_k(\bar{X}_{t_k}, \varepsilon_{k+1})$$

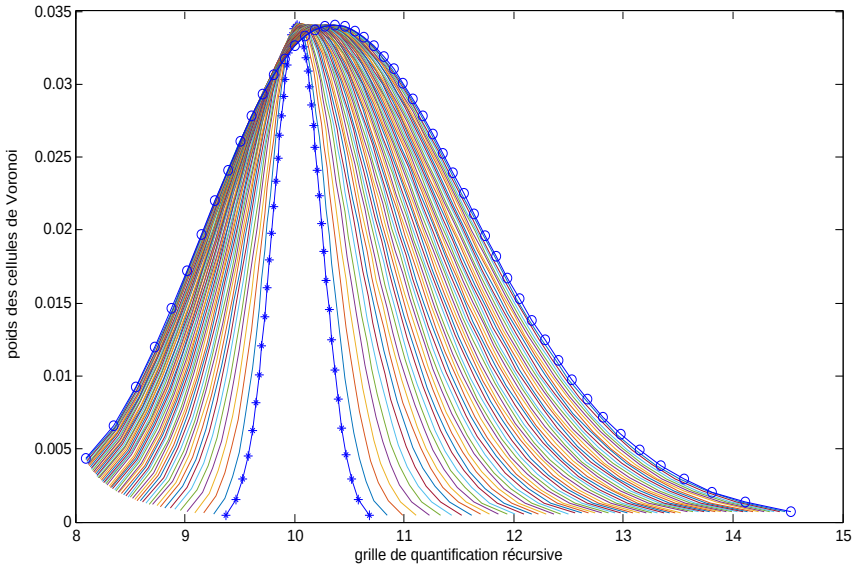

Figure 1.4: Graphes de $x_i^k \mapsto p_i^k$ où $(x_i^k)_{1\leq i\leq N_k}$ est la grille de quantification récursive, pour tout $k \in \{1,\dots,n\}$, dans un modèle de Black-Scholes (* correspond à $k=2$ et $\circ$ correspond à $k=n=30$).

On considère $n = 30$ pas de temps et contruit des grilles de taille $N_k = 50$, pour tout $k \in \{1,\dots,n\}$. On considère

$$T = 1, \quad X_0 = 100, \quad r = 0.006, \quad \sigma = 0.2.$$

1.3.2 Contributions de cette thèse

Dans le chapitre 6, nous établissons des bornes supérieures de l'erreur L^p de quantification récursive d'un modèle de Markov général de la forme $X_{k+1} = F_k(X_k, \varepsilon_{k+1})$, $(\varepsilon_k)_{1\leq k\leq n}$ étant une suite de variables aléatoires i.i.d. de loi Normale. Nous étendons les résultats obtenus dans un cadre L^2 dans [63] et estimons des erreurs dans L^p pour $p \in (1, 2+d)$. Nous considérons que les grilles Γ_k, dans (1.18), sont des quantifieurs quadratiques optimaux de $\widetilde{X}_{t_k}$. Ceci est important car la propriété de stationnarité (1.7) satisfaite par les quantificateurs quadratiques optimaux sera nécessaire pour notre étude.

Puisque nous estimons des bornes supérieures de l'erreur L^p de quantification récursive en utilisant des quantifieurs L^2-optimaux de $\widetilde{X}_{t_k}$, nous nous trouvons dans une position où nous devons traiter l'erreur de quantification L^p d'un quantifieur L^2-optimal. Pour cela, nous nous appuyons sur les résultats du problème de mismatch de distorsion, aussi connu sous le nom de problème (L^r - L^s), rappelé dans le théorème 6.2.2. De plus, nous montrons et utilisons un lemme technique qui permet de contrôler l'espérance de la forme $\mathbb{E}|a + A\sqrt{h}Z|^r$ pour $r \geq 2$, $a \in \mathbb{R}^d$, $h > 0$, $A \in \mathcal{M}(d,q,\mathbb{R})$ et $Z \in L^r_{\mathbb{R}^q}(\mathbb{P})$ une variable aléatoire à valeurs dans $\mathbb{R}^q$ tel que

$\mathbb{E}[Z] = 0$, plus précisemment

$$\mathbb{E}|a + A\sqrt{h}Z|^r \leq |a|^r\Big(1 + 2^{(r-3)_+}(r-1)(r-2)h\Big) + 2^{(r-3)_+}(r-1)h\|A\|^r\mathbb{E}|Z|^r\left(1 + \frac{r}{2}h^{\frac{r}{2}-1}\right).$$

Cette inégalité sera très utile pour établir plusieurs résultats théoriques permettant d'aboutir à des bornes d'erreur.

Avec tous les outils nécessaires, nous montrons que l'erreur de quantification récursive dans L^p du schéma d'Euler est bornée par

$$\forall k \in \{0, \dots, n\} \qquad \|\bar{X}_{t_k} - \widehat{X}_{t_k}\|_p \leq K\sum_{l=1}^{k} C_l\|\widehat{X}_{t_l} - \widetilde{X}_{t_l}\|_p \leq K'\sum_{l=1}^{k} C_l N_l^{-\frac{1}{d}} \tag{1.19}$$

où N_l est la taille du quantifieur Γ_l de $\widetilde{X}_{t_l}$ et K, K' et C_l sont des constantes finies positives à préciser ultérieurement dans le théorème 6.2.1 dépendant de p, d, b, σ et ε_k.

Lorsque la dimension d augmente, une technique de substitution intéressante est la quantification récursive produit qui, cependant, devient très exigeante pour les dimensions très élevées. Dans le chapitre 6, nous présentons et étudions une alternative, la *quantification récursive hybride* qui consiste en la quantification du bruit Gaussien dans (1.18) de sorte que la quantification hybride récursive de $\bar{X}_{t_k}$ est donnée par le schéma récursif suivant

$$\begin{cases} \widetilde{X}_{t_k} &= F_{k-1}(\widehat{X}_{t_{k-1}}, \widehat{\varepsilon}_k), \\ \widehat{X}_{t_k} &= \mathrm{Proj}_{\Gamma_k}(\widetilde{X}_{t_k}), \qquad \forall k = 1, \dots, n. \end{cases}$$

où $(\widehat{\varepsilon}_k)_k$ est maintenant une suite de quantifieurs optimaux de la loi Normale $\mathcal{N}(0, I_q)$, qui sont déjà calculés et stockés off-line. En se basant sur les mêmes outils utilisés pour établir des bornes supérieures pour la quantification récursive standard, nous établissons des borness d'erreur dans L^p pour la quantification récursive hybride pour $p \in (1, 2 + d)$, comme suit

$$\|\bar{X}_{t_k} - \widehat{X}_{t_k}\|_p \leq K\sum_{l=1}^{k} C_X (N_l^X)^{-\frac{1}{d}} + K\sum_{l=1}^{k} C_\varepsilon (N_l^\varepsilon)^{-\frac{1}{d}}$$

où N_l^X est la taille du quantifieur optimal de $\widetilde{X}_{t_l}$, N_l^ε est la taille des quantifieurs optimaux du vecteur aléatoire Gaussien et K, C_X, C_ε sont des constantes positives finies. En pratique, puisque les ε_k sont des variables aléatoires i.i.d., nous construisons des quantifieurs correspondants $\hat{\varepsilon}_k$ de la même taille $N_k^\varepsilon = N^\varepsilon$ pour tout $k \in \{1, \dots, n\}$.

1.3.3 Application à la discrétisation des Equations Différentielles Stochastiques Rétrogrades Réfléchies

La quantification récursive est une technique de discrétisation spatiale utilisée dans les applications financières. On peut citer le pricing d'options dans un modèle de volatilité stochastique (voir [15]) et le pricing d'un panier d'options (voir [28]). Dans le chapitre 6, nous nous appuyons sur la quantification récursive pour la discrétisation spatiale de la solution d'une Equation Différentielle Stochastique Rétrograde Réfléchie (RBSDE). Des approximations de telles équations

ont déjà été établies par plusieurs méthodes. Par exemple, on peut citer les méthodes de régression avec des simulations de Monte Carlo (voir [9]), les itérations de Picard combinées avec une décomposition dans le chaos de Wiener (voir [17]) et la quantification optimale (voir [3, 4, 37]).

On considère la RBSDE de maturité T

$$Y_t = g(X_T) + \int_t^T f(s, X_s, Y_s, Z_s)ds + K_T - K_t - \int_t^T Z_s.dW_s\,, \qquad t \in [0,T], \tag{1.20}$$

$$Y_t \geq h(t, X_t) \quad \text{and} \quad \int_0^T (Y_s - h(s, X_s))dK_s = 0.$$

où le processus $(X_t)_{t\in[0,T]}$ est une diffusion donnée par (1.16) et f, g et h sont des fonctions continues Lipschitziennes. La solution de cette équation est un triplet (Y_t, Z_t, K_t) et une telle solution existe et est unique comme établi dans [25] sous des hypothèses de Lipschitz appropriées. Toutefois, cette solution n'admet pas une forme fermée en général. Il faut donc l'approcher par des schémas de discrétisation spatio-temporelle. Le schéma de discrétisation temporelle $(\bar{Y}_t^n, \bar{\zeta}_t^n)$ associé à (Y_t, Z_t) est basé sur le schéma d'Euler du processus $(X_t)_{t\in[0,T]}$. Plusieurs choix sont possibles (voir [3, 9, 48]). Notre choix dans ce travail est d'insérer l'espérance conditionnelle dans le driver f comme suit

$$\begin{aligned}
&\bar{Y}_T^n = g(\bar{X}_T^n)\\
&\widetilde{Y}_{t_k}^n = \mathbb{E}(\bar{Y}_{t_{k+1}}^n|\mathcal{F}_{t_k}) + \Delta f\big(t_k, \bar{X}_{t_k}^n, \mathbb{E}(\bar{Y}_{t_{k+1}}^n|\mathcal{F}_{t_k}), \bar{\zeta}_{t_k}^n\big)\,, \quad k = 0, \dots, n-1,\\
&\bar{\zeta}_{t_k}^n = \frac{1}{\Delta}\mathbb{E}(\bar{Y}_{t_{k+1}}^n(W_{t_{k+1}} - W_{t_k})\,|\,\mathcal{F}_{t_k})\,, \quad k = 0, \dots, n-1,\\
&\bar{Y}_{t_k}^n = \widetilde{Y}_{t_k}^n \vee h(t_k, \bar{X}_{t_k}^n)\,, \quad k = 0, \dots, n-1.
\end{aligned}$$

De tels schémas ont été envisagés pour des BSDE (sans réflexion) dans [65] ou pour les BSDE à double réflexion dans [37], alors que dans la plupart de la littérature, l'espérance est généralement appliquée en dehors de la fonction f. Dans certains articles motivés par les options américaines, f ne dépend pas du processus Z_t.

Ce schéma ne peut pas être simulé à cause des espérances conditionnelles, nous sommes donc amenés, comme nos prédécesseurs, à effectuer une discrétisation spatiale supplémentaire basée ici sur la quantification récursive du processus $\bar{X}_{t_k}$. Le schéma résultant est le suivant

$$\begin{aligned}
&\widehat{Y}_T^n = g(\widehat{X}_T)\\
&\widehat{\zeta}_{t_k}^n = \frac{1}{\Delta}\mathbb{E}\left(\widehat{Y}_{t_{k+1}}^n(W_{t_{k+1}} - W_{t_k})\,|\,\mathcal{F}_{t_k}\right), \quad k = 0, \dots, n-1,\\
&\widehat{Y}_{t_k}^n = \max\left(h_k(\widehat{X}_{t_k})\,, \mathbb{E}\left(\widehat{Y}_{t_{k+1}}^n\,|\,\mathcal{F}_{t_k}\right) + \Delta f\big(t_k, \widehat{X}_{t_k}, \mathbb{E}(\widehat{Y}_{t_{k+1}}^n\,|\,\mathcal{F}_{t_k}), \widehat{\zeta}_{t_k}^n\big)\right), \quad k = 0, \dots, n-1,
\end{aligned}$$

On établit des bornes supérieures pour les erreurs induites par les discrétisations temporelle et spatiale mentionnées ci-dessus.

Discrétisation temporelle Pour l'erreur de discrétisation en temps, nous établissons des bornes supérieures de l'erreur dans L^2. A cette fin, nous introduisons un processus continu en

temps qui étend $\bar{Y}_{t_k}$, en se basant sur le théorème de représentation de martingale. Cela conduit à définir un processus càdlàg $\widetilde{Y}_t$ sur $[t_k, t_{k+1})$ et un processus làdcàg $\bar{Y}_t$ sur $(t_k, t_{k+1}]$, par

$$\widetilde{Y}_t = \bar{Y}_t = \bar{Y}_{t_{k+1}} - (t_{k+1} - t) f_k(\bar{X}_{t_k}, \mathbb{E}(\bar{Y}_{t_{k+1}} \,|\, \mathcal{F}_{t_k}), \bar{\zeta}_{t_k}) - \int_t^{t_{k+1}} \bar{Z}_s dW_s, \tag{1.21}$$

conduisant à la représentation suivante

$$\widetilde{Y}_t = \bar{Y}_T + \int_t^T f(\underline{s}, \bar{X}_{\underline{s}}, \mathbb{E}(\bar{Y}_{\bar{s}} \,|\, \mathcal{F}_{\underline{s}}), \bar{\zeta}_{\underline{s}})\, ds - \int_t^{t_{k+1}} \bar{Z}_s dW_s + \bar{K}_T - \bar{K}_t$$

où $\underline{s} = t_k$ et $\bar{s} = t_{k+1}$ si $s \in (t_k, t_{k+1})$, $\bar{Z}_t$ est un processus tel que $\mathbb{E} \sup_{[0,T]} |\bar{Z}_s|^2 < +\infty$ et $\bar{K}_{t_k}$ est un processus càdlàg croissant, nul au temps 0, défini par

$$\bar{K}_{t_k} = \sum_{j=0}^{k} \left(h_j(\bar{X}_{t_j}) - \widetilde{Y}_{t_k} \right)_+$$

et tel que $\bar{K}_t = \bar{K}_{t_k}$ pour tout $t \in (t_k, t_{k+1})$. Cela conduit à la borne supérieure suivante pour l'erreur de discrétisation temporelle, pour tout $k \in \{1, \dots, n\}$,

$$\mathbb{E}|Y_{t_k} - \bar{Y}_{t_k}|^2 \leq C_{b,\sigma,f,h,T} \left(\Delta + \int_0^T \mathbb{E}|Z_s - Z_{\underline{s}}|^2 ds \right)$$

où $C_{b,\sigma,f,h,T}$ est une constante réelle positive. Cela montre classiquement que le taux de convergence du schéma de discrétisation temporelle est régi par la régularité du processus $(Z_t)_{t\in[0,T]}$ (qui peut être analysé par les méthodes PDE lorsque b et σ sont suffisamment régulières).

Discrétisation spatiale En ce qui concerne la discrétisation spatiale, on établit des bornes d'erreur dans L^p pour $p \in (1, 2+d)$ et $k \in \{1, \dots, n\}$, comme suit

$$\|\bar{Y}_{t_k} - \widehat{Y}_{t_k}\|_p \leq K \| \max_{k \leq l \leq n} |\bar{X}_{t_l} - \widehat{X}_{t_l}| \|_p$$

où K est une constante finie positive définie ultérieurement dans le chapitre 6. Les normes $\|\bar{X}_{t_l} - \widehat{X}_{t_l}\|_p$ sont des erreurs de quantification récursive déjà contrôlées par (1.19).

De point de vue algorithmique, one montre par récurrence rétrograde qu'il existe des fonctions $\widehat{y}_k : \Gamma_k \mapsto \mathbb{R}, k \in \{0, \dots, n\}$, telles que $\widehat{Y}_k = \widehat{y}_k(\widehat{X}_k)$, pour tout $k \in \{0, \dots, n\}$, définies récursivement par le principe de programmation dynamique rétrograde (BDPP) suivant

$$\begin{cases} \widehat{y}_n &= h_n \\ \widehat{y}_k &= \max \left(h_k,\, \widehat{P}_k \widehat{y}_{k+1} + \Delta f_k(., \widehat{P}_k \widehat{y}_{k+1}, \widehat{Q}_k \widehat{y}_{k+1}) \right), \quad k = 0, \dots, n-1, \end{cases}$$

où

$$\widehat{P}_k \widehat{y}_{k+1}(\widehat{X}_k) = \mathbb{E}\left(\widehat{y}_{k+1}(\widehat{X}_{k+1}) \,|\, \mathcal{F}_{t_k} \right) \qquad \text{et} \qquad \widehat{Q}_k \widehat{y}_{k+1}(\widehat{X}_k) = \frac{1}{\sqrt{\Delta}} \mathbb{E}\left(\widehat{y}_{k+1}(\widehat{X}_{k+1}) \varepsilon_{k+1} \,|\, \mathcal{F}_{t_k} \right).$$

De même, il existe des fonctions $\widehat{z}_k$ telles que $\widehat{\zeta}_k = \widehat{z}_k(\widehat{X}_k)$, définies par

$$\widehat{z}_k = \widehat{Q}_k \widehat{y}_{k+1}.$$

En s'appuyant sur ces BDPP et sur la quantification récursive $\widehat{X}_{t_k}^{\Gamma_k}$ de $\bar{X}_{t_k}$, $\Gamma_k = \{x_1^k, \ldots, x_{N_k}^k\}$, on approche la solution Y_0 de la RBSDE à l'instant 0 par la valeur initale $\widehat{y}_0$ du schéma

$$\begin{cases} \widehat{y}_n(x_i^n) &= h_n(x_i^n)\,, \qquad i = 1, \ldots, N_n, \\ \widehat{y}_k(x_i^k) &= \max\Big(h_k(x_i^k),\, \widehat{\alpha}_k(x_i^k) + \Delta f_k(x_i^k, \widehat{\alpha}_k(x_i^k), \widehat{\beta}_k(x_i^k))\Big)\,, \qquad i = 1, \ldots, N_k, \end{cases}$$

où

$$\widehat{\alpha}_k(x_i^k) = \sum_{j=1}^{N_{k+1}} \widehat{y}_{k+1}(x_j^{k+1}) p_{ij}^k \qquad \text{et} \qquad \widehat{\beta}_k(x_i^k) = \frac{1}{\Delta} \sum_{j=1}^{N_{k+1}} \widehat{y}_{k+1}(x_j^{k+1}) \pi_{ij}^k$$

avec

$$\pi_{ij}^k = \frac{\sqrt{\Delta}}{p_i^k}\, \mathbb{E}\Big(\varepsilon_{k+1} \mathbb{1}_{\{\widehat{X}_{k+1} = x_j^{k+1},\, \widehat{X}_k = x_i^k\}}\Big) = \sqrt{\Delta}\, \mathbb{E}\Big(\varepsilon_{k+1} \mathbb{1}_{\{F_k(x_i^k, \varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\}}\Big).$$

Nous illustrons cette approximation par plusieurs exemples numériques unidimensionnels et multidimensionnels à la fin du chapitre 6 et dans le chapitre 7. Dans le cadre unidimensionnel, nous considérons le prix d'une option d'achat américaine sur un marché avec un écart acheteur-vendeur sur les taux d'intérêt et le prix d'une option de vente américaine sous la probabilité historique, les deux exemples sont considérés à la fois dans un modèle Black-Scholes et un modèle CEV. En ce qui concerne le cadre multidimensionnel, nous évaluons le prix d'une option de change américaine bidimensionnelle dans un modèle de Black-Scholes et considérons un exemple multidimensionnel dû à J.F. Chassagneux.
En outre, nous considérons l'évaluation du prix des options de vente américaines pour $d = 1$ et $d = 2$. Nous montrons que les estimations des bornes d'erreur L^p induites par la discrétisation spatiale correspondante peuvent être obtenues directement. En fait, puisque $(\bar{X}_{t_k})_{0 \le k \le n}$ et $(\widehat{X}_{t_k})_{0 \le k \le n}$ sont des chaînes de Markov, $\bar{Y}_{t_k}$ et $\widehat{Y}_{t_k}$ s'écrivent sous forme d'enveloppes de Snell comme suit: pour tout $k \in \{1, \ldots, n\}$,

$$\bar{Y}_{t_k} = \mathbb{P}\text{-esssup}\{\mathbb{E}\,(h_\tau(\bar{X}_\tau)\,|\,\mathcal{F}_\tau),\, \tau \in \{t_k, \ldots, T\}\, \mathcal{F}_\tau\text{-temps d'arrêt}\}$$

et

$$\widehat{Y}_{t_k} = \mathbb{P}\text{-esssup}\{\mathbb{E}\,(h_\tau(\widehat{X}_\tau)\,|\,\mathcal{F}_\tau),\, \tau \in \{t_k, \ldots, T\}\, \mathcal{F}_\tau\text{-temps d'arrêt}\}$$

où $h(x) = \max(K - x, 0)$. Par conséquent, on obtient, pour tout $k \in \{1, \ldots, n\}$,

$$\|\bar{Y}_{t_k} - \widehat{Y}_{t_k}\|_p \le [h]_{\text{Lip}} \big\| \max_{l \ge k} |\bar{X}_{t_l} - \widehat{X}_{t_l}| \big\|_p.$$

Dans tous les exemples, nous comparons les résultats obtenus par quantification récursive à ceux obtenus par d'autres types de quantification. Si $d = 1$, nous comparons la quantification récursive à la quantification optimale, gloutonne et récursive gloutonne. Et, si $d > 1$, one adopte la quantification récursive hybride, au lieu de la quantification récursive standard, et on compare les résultats à ceux obtenus par quantification optimale et gloutonne. Toutes les méthodes mentionnées sont détaillées dans les chapitres 6 et 7.

Chapter 2

Introduction

This thesis is divided into two main parts. The first part contains Chapters 3, 4 and 5 where we present new theoretical results and aspects of greedy quantization, as well as some numerical studies. Briefly, we adopt a new approach to extend rate optimality and distortion mismatch results for greedy quantization to a wider class of distributions and establish L^s-rate optimality results for L^r-dilated or contracted greedy quantization sequences. Numerically, we carry out some experiments and emphasize some properties of greedy quantization that make it advantageous in face of other approximation methods (mainly quasi Monte Carlo). In the second part consisting of Chapters 6 and 7, we establish, first, L^p-error bounds for recursive quantization of a general d-dimensional Markov model for $p \in (1, 2+d)$ and, then, extend error bounds for the recursive quantization-based discretization schemes of reflected Backward Stochastic Differential Equations to the L^p-framework .

2.1 Optimal quantization: Principle, definitions and main results

Optimal vector quantization is a technique going back to the 1950's (see [30]) when it was first devised in the signal processing field to discretize continuous signals for their transmission. It was then extended to many domains such as Information theory, cluster analysis, etc., until it was introduced as a mathematical tool in the 1990's. It was first used as a quadrature formula in the numerical integration field for the computation of expectations (see [54]), and then, in the early 2000's, for the approximation of conditional expectations in view of financial applications, mainly pricing of American options (see [3, 4, 5]), of non-linear filtering problems (see [58]) and simulation of Stochastic Differential Equations (see [3, 67]), etc.

The mathematical problem of optimal quantization consists in finding the best approximation, in a sense to be specified, of a (possibly) continuous probability distribution by a discrete probability distribution whose support is of finite cardinal, or, in other words, the best approximation of a multidimensional random variable X by a random variable Y taking a finite number n of values. Let $d \geq 1$ and X be a d-dimensional random variable defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $X \in L^r(\mathbb{P}), r > 0$ i.e. $\mathbb{E}|X|^r < +\infty$ where $|.|$ denotes a priori any norm on $\mathbb{R}^d$. We denote $P = \mathbb{P}_X$ the probability distribution of X. The goal is to approximate X by $q(X)$, where q is a Borel function defined on $\mathbb{R}^d$ and having values in a d-dimensional *grid*

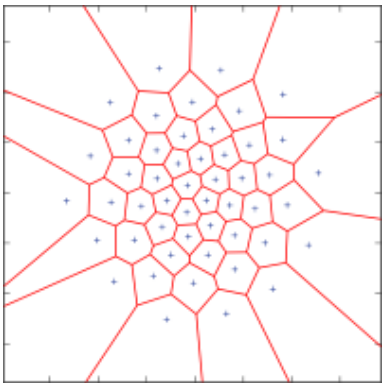

Figure 2.1: Example of a Voronoï diagram in $\mathbb{R}^2$ w.r.t. the Euclidean norm.

$\Gamma = \{x_1, \dots, x_n\}$ of size n. The best choice for q, Γ being fixed, is clearly any Borel nearest neighbor projection $\pi_\Gamma : \mathbb{R}^d \to \Gamma$ defined by $\pi_\Gamma(\xi) = \sum_{i=1}^n x_i \mathbb{1}_{C_i(\Gamma)}(\xi)$, where

$$C_i(\Gamma) \subset \{\xi \in \mathbb{R}^d : |\xi - x_i| \leq \min_{j \neq i} |\xi - x_j|\}, \qquad i = 1, \dots, n, \tag{2.1}$$

is a Borel partition of $\mathbb{R}^d$ called the Voronoï diagram induced by Γ. The Borel sets $C_i(\Gamma)$ are called the Voronoï cells of the partition induced by Γ. An example of a Voronoï diagram in $\mathbb{R}^2$ equipped with the Euclidean norm is presented in Figure 2.1.

Then, the Voronoï quantization of X is the composition of π_Γ and X:

$$\widehat{X}^\Gamma = \pi_\Gamma(X) := \sum_{i=1}^n x_i \mathbb{1}_{C_i(\Gamma)}(X).$$

Its distribution is characterized by the grid $\Gamma = \{x_1, \dots, x_n\}$ and the weights of the corresponding Voronoï cells given, for every $i \in \{1, \dots, n\}$, by

$$p_i^n = P(\widehat{X}^\Gamma = x_i) = P\big(X \in C_i(\Gamma)\big).$$

We will often denote, $\widehat{X}$ instead of $\widehat{X}^\Gamma$ to alleviate notations. The L^r-quantization error associated to a grid Γ is defined, for every $r \in (0, +\infty)$, by

$$e_r(\Gamma, X) = \|X - \pi_\Gamma(X)\|_r = \|X - \widehat{X}^\Gamma\|_r = \|\mathrm{dist}(X, \Gamma)\|_r \tag{2.2}$$

where $\|Y\|_r = \big(\mathbb{E}|Y|^r\big)^{\frac{1}{r}}$ denotes the $L^r(\mathbb{P})$-norm of a random vector Y (or quasi-norm if $0 < r < 1$). We also define the L^r-distortion function G_n^r on $(\mathbb{R}^d)^n$ by

$$G_n^r(x_1, \dots, x_n) = e_r\big(\{x_1, \dots, x_n\}, X\big)^r. \tag{2.3}$$

This function is differentiable if the $(x_i)_{1 \leq i \leq n}$ are pairwise distinct or, equivalently, $\Gamma = \{x_1, \dots, x_n\}$ has full size n, and the boundaries of the Voronoï diagram are negligible w.r.t. the distribution P of X, it depends also on the differentiability of the underlying norm itself. Its gradient is given by

$$\nabla G_n^r(x_1, \dots, x_n) = r \left(\mathbb{E}\Big[\mathbb{1}_{X \in C_i(\Gamma)} \frac{(x_i - X)}{|x_i - X|} |x_i - X|^{r-2}\Big] \right)_{1 \leq i \leq n}.$$

Optimal quantization problem consists in finding a grid Γ that minimizes the quantization error (2.2), i.e. solves the following minimization problem

$$e_{r,n}(X) = \inf_{\Gamma,\, \mathrm{card}(\Gamma)\leq n} e_r(\Gamma, X). \tag{2.4}$$

If $X \in L^r_{\mathbb{R}^d}(\mathbb{P})$, this problem always admits at least one solution Γ called optimal quantizer of size n of X or P, and the corresponding quantization error converges to 0 when the size n goes to $+\infty$. For a proof, we refer to [32, 56, 57] among others. The rate of convergence of the L^r-quantization error to 0 is given by two well known results exposed in the following theorem. The first one is a sharp asymptotic result and the second one is universal and non-asymptotic.

Theorem 2.1.1. *(a)* Zador's Theorem (see [75]) *: Let $r > 0$ and let $X \in L^{r+\eta}_{\mathbb{R}^d}(\mathbb{P})$ for some $\eta > 0$, with distribution P such that $dP(\xi) = \varphi(\xi)d\lambda_d(\xi) + d\nu(\xi)$ where λ_d denotes the Lebesgue measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then,*

$$\lim_{n\to+\infty} n^{\frac{1}{d}} e_{r,n}(X) = \tilde{J}_{r,d} \|\varphi\|^{\frac{1}{r}}_{L^{\frac{r}{r+d}}(\lambda_d)} \tag{2.5}$$

where $\tilde{J}_{r,d} = \inf_{n\geq 1} n^{\frac{1}{d}} e_{r,n}(U([0,1]^d)) \in (0, +\infty)$.
(b) Extended Pierce's Lemma (see [44])*: Let $r, \eta > 0$. There exists a constant $\kappa_{d,r,\eta} \in (0, +\infty)$ such that, for any random vector $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^d$,*

$$\forall n \geq 1, \quad e_{r,n}(X) \leq \kappa_{d,r,\eta} \sigma_{r+\eta}(X) n^{-\frac{1}{d}} \tag{2.6}$$

where, for every $p \in (0, +\infty)$, $\sigma_p(X) = \inf_{a\in\mathbb{R}^d} \|X - a\|_p$ is the L^p-standard deviation of X.

An important property shared by quadratic L^2-optimal quantizers is *stationarity*. A quantization grid Γ is said to be stationary iff the boundaries of the Voronoï partitions are P-negligible and

$$\widehat{X}^\Gamma = \mathbb{E}\,(X \,|\, \widehat{X}^\Gamma). \tag{2.7}$$

In fact, any quadratic optimal quantizer (w.r.t. the Euclidean norm) has P-negligible boundaries (see Proposition 4.2 in [32]). This property is very important in most applications, especially because most algorithms devised to compute optimal quantizers are based on this stationarity property, even if not all stationary quantizers are optimal. Its importance is also emphasized in the quantization-based numerical integration, this topic is explained in details in the following.

Optimal quantizers are used in numerical integration to approximate expectations of the form $\mathbb{E}f(X)$ for a random variable X and a continuous function f. Since the L^r-quantization error $\|X - \widehat{X}^\Gamma\|_r$ converges to 0, then $\widehat{X}^\Gamma$ converges to X in L^r and hence in distribution. So, one approximates $\mathbb{E}f(X)$ by

$$\mathbb{E}f(\widehat{X}^\Gamma) = \sum_{i=1}^n p_i^n f(x_i)$$

where $p_i^n = \mathbb{P}(X \in C_i(\Gamma)), i = 1, \ldots, n$, are the weights of the Voronoï cells corresponding to the optimal quantizer $\Gamma = \{x_1, \ldots, x_n\}$ of size n of X. Upper bounds for the error induced by this type of approximation have been established for stationary optimal quantizers, depending

on the regularity of the function f (see [42, 56, 57]). For example, if f is a Lipschitz continuous function with Lipschitz coefficient $[f]_{\text{Lip}}$ and Γ is any grid, then

$$\left|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{\Gamma})\right| \leq [f]_{\text{Lip}} \|X - \widehat{X}^{\Gamma}\|_1 \leq [f]_{\text{Lip}}\, e_1(X,\Gamma) \leq [f]_{\text{Lip}}\, e_2(X,\Gamma).$$

If, furthermore, Γ is a stationary quantizer for X or P, then, if f is differentiable with an α-Hölder gradient ∇f, one has (see [57] for example)

$$\left|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{\Gamma})\right| \leq \frac{1}{1+\alpha}[\nabla f]_{\alpha} \|X - \widehat{X}^{\Gamma}\|_{1+\alpha}^{1+\alpha}.$$

In particular, if ∇f is Lipschitz continuous, then

$$\left|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{\Gamma})\right| \leq \frac{1}{2}[\nabla f]_{\text{Lip}} \|X - \widehat{X}^{\Gamma}\|_2^2.$$

2.1.1 Construction of optimal quantizers

Let P be a probability defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and let X be a random variable with distribution P. Most devised algorithms for designing (quadratic) optimal quantizers of X or its distribution P are based on the differentiability of the distortion function and on the stationarity property (2.7). In fact, the following Proposition makes up the starting point of the numerical methods to compute optimal quantizers in the quadratic case.

Proposition 2.1.2. *Let $X \in L^2(P)$ be a random variable such that* $\text{card}(\text{supp}(\text{P})) \geq n$. *Any grid Γ minimizing the quadratic distortion function G_n^2 associated to X is a stationary quantizer of X.*

Furthermore, it has been proved, in [39, 73], that if $d = 1$ and the probability density function of X is log-concave, then there exists a unique stationary quantizer of X and this quantizer is a global minimum of the distortion function.

Newton-Raphson zero search algorithm This is a deterministic procedure used if the probability distribution is known explicitly. Assume that the distribution P of X is absolutely continuous with respect to the Lebesgue measure with continuous density φ. The L^r-optimal quantizer is obtained as follows: Denoting $x = (x_1, \ldots, x_n)$ the grid to build, one has

$$x^{[l+1]} = x^{[l]} - \left(\nabla^2 G_n^r(x^{[l]})\right)^{-1} \nabla G_n^r(x^{[l]})$$

starting at $x^{[0]}$ belonging to the convex hull of the support of X, where $\nabla^2 G_n^r(x)$ is the Hessian matrix of G_n^r. This can be improved by using the Levenberg-Marquardt algorithm

$$x^{[l+1]} = x^{[l]} - \left(\nabla^2 G_n^r(x^{[l]}) + \lambda_l I_d\right)^{-1} \nabla G_n^r(x^{[l]})$$

for an appropriate choice of the "damping" coefficients λ_l.

Competitive Learning Vector Quantization (CLVQ) This is a stochastic gradient descent algorithm used for the computation of d-dimensional quantizers for $d \geq 1$, also known as k-means algorithm. In higher dimensions in the quadratic case, one takes advantage of the representation of G_n^2 as an expectation and switches to the CLVQ algorithm defined by the following recursion

$$x^{[l+1]} = x^{[l]} - \gamma_{l+1}\Big(\mathbb{1}_{X \in C_i(x^{([l]})}(x_i^{[l]} - X)\Big)_{1 \leq i \leq n}$$

starting at $x^{[0]}$ belonging to the convex hull of the support of X, where $(\gamma_l)_{l \geq 1}$ is a sequence of step parameters satisfying $\sum_{l \geq 1} \gamma_l = +\infty$ and $\sum_{l \geq 1} \gamma_l^2 < +\infty$.

Lloyd's algorithm This is a fixed-point search based directly on the stationarity property. In the one-dimensional case, it is a deterministic procedure used if the probability distribution is known explicitly and defined by

$$x_i^{[l+1]} = \frac{\mathbb{E}(X \mathbb{1}_{X \in C_i(x^{[l]})})}{P(X \in C_i(x^{[l]}))}.$$

starting at $x^{[0]}$ belonging to the convex hull of the support of X.

Randomized Lloyd's algorithm In higher dimensions, the procedure above becomes intractable so one switches to the randomized Lloyd's algorithm. The expectations and probabilities are computed by a Monte Carlo simulation of size M as follows

$$x_i^{[l+1]} = \frac{\sum_{m=1}^{M} X^m \mathbb{1}_{X_m \in C_i(x^{[l]})}}{\mathrm{card}(X^m \,;\, X^m \in C_i(x^{[l]}))} \tag{2.8}$$

starting at $x^{[0]}$ belonging to the convex hull of the support of X, where $(X^m)_{1 \leq m \leq M}$ are M i.i.d. copies of X.

For further details on the above procedures, we refer to [56, 57]. Note that highly accurate quantization grids of $\mathcal{N}(0; I_q)$ distributions for dimensions $d = 1$ up to 10 and regularly sampled sizes from $N = 1$ to 1 000 can be downloaded from the quantization website www.quantize.maths-fi.com (for non-commercial purposes).

For some few scalar probability distributions, there exists closed or semi-closed forms for optimal quantization grids. For example, the optimal quantizer Γ_n of size n of the Uniform distribution $\mathcal{U}([0, 1])$ is given by

$$\Gamma_n = \Big\{\frac{2i-1}{2n},\ i = 1, \ldots, n\Big\},$$

and semi-closed forms were given in [29, 32] for the exponential, power, inverse power and Laplace distributions.

In the two-dimensional framework, a deterministic approach to optimize quadratic quantizers is developed in [52]. It relies on the approximation of two-dimensional integrals over convex

polygons by very effective numerical quadrature formulas.

For higher dimensions, stochastic optimization procedures may become too expensive and computationally too demanding. When the target law is a tensor product of its independent marginal laws, one can rely on product quantization instead of standard multi-dimensional procedures. It consists in obtaining multi-dimensional quantizers as a result of the tensor product of one-dimensional sequences, already computed by one of the algorithms cited above.

2.2 Greedy quantization

When the dimension d increases, the search of a solution to the quantization problem (2.4) becomes more complicated and computationally too demanding. Therefore, one needs to introduce a sub-optimal solution which is easier to compute, as long as the rate of convergence remains similar to that of optimal quantizers. This solution is provided by greedy vector quantization.

2.2.1 Principle and existing results

Let X be a random variable with probability P defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Greedy vector quantization has first been introduced and investigated in [12] for compactly supported distributions P (in a L^1 sense) as a model of short term experiment planning versus long term experiment planning represented by regular L^1-quantization at a given level n. It has been then reintroduced independently and studied extensively in [45] for various classes of distributions with possibly unbounded support. In both cases, it consists in determining, for a random vector (or distribution) with finite r-th moment, a sequence $(a_n)_{n\geq 1}$ in $\mathbb{R}^d$ which is recursively L^r-optimal step by step. In other words, having already computed the first n points of the sequence $a^{(n)} = \{a_1, \ldots, a_n\}$, one adds the $(n+1)$-the point of the sequence as a solution to

$$a_{n+1} \in \operatorname{argmin}_{\xi \in \mathbb{R}^d} e_r(a^{(n)} \cup \{\xi\}, X), \tag{2.9}$$

with $a^{(0)} = \varnothing$. Note that a_1 is an/the L^r-median of the distribution P of X. A solution to this problem always exists and is called an L^r-optimal greedy quantization sequence for X or its distribution P. However, this solution may not be unique even if the L^r-median a_1 is. This is due to the dependency of greedy quantization on the symmetry of the distribution P. This existence has been proved in full generality in [45] where the authors also showed that the corresponding L^r-quantization error is decreasing w.r.t. the number n of points of the sequence and it converges to 0 as n goes to infinity. The optimal $n^{-\frac{1}{d}}$-rate of convergence has also been proved in [45]. It relies on the integrability of the b-maximal function associated to the L^r-optimal greedy quantization sequence $(a_n)_{n\geq 1}$ defined, for every $b \in (0, \frac{1}{2})$ and every $\xi \in \mathbb{R}^d$, by

$$\Psi_b(\xi) = \sup_{n\in\mathbb{N}} \frac{\lambda_d\left(B(\xi, b\,\mathrm{dist}(\xi, a^{(n)}))\right)}{P\left(B(\xi, b\,\mathrm{dist}(\xi, a^{(n)}))\right)}. \tag{2.10}$$

The theorem below deals with the L^r-rate optimality of greedy quantization sequences.

Theorem 2.2.1. *Let $X \in L^r(P)$, $r \in (0, +\infty)$ and let $(a_n)_{n\geq 1}$ be an L^r-greedy quantization sequence of X. If there exists $b \in (0, \frac{1}{2})$ such that the b-maximal function $\Psi_b \in L^{\frac{r}{r+d}}(P)$, then*

$$\limsup_n n^{\frac{1}{d}} e_r(a^{(n)}, X) < +\infty.$$

The b-maximal function Ψ_b is also used to show that greedy quantization sequences satisfy the distortion mismatch problem, i.e. the property that the optimal rate of L^r-quantizers holds for L^s-quantizers for $s > r$. This problem was already investigated for optimal quantizers in [33] and then in [65]. For greedy quantization sequences, the following theorem, established in [45], solves the problem.

Theorem 2.2.2. *Let $s \in (r, +\infty)$, $X \in L^r(P)$ and $(a_n)_{n\geq 1}$ an L^r-optimal greedy quantization sequence of X. Assume that $\Psi_b \in L^{\frac{s}{r+d}}(P)$ for some $b \in (0, \frac{1}{2})$. Then, $X \in L^s(P)$ and*

$$\limsup_n n^{\frac{1}{d}} e_s(a^{(n)}, X) < +\infty.$$

How to obtain greedy quantization sequences

Greedy quantization sequences are computed by implementing variants of usual algorithms of computing optimal quantizers, such as Lloyd's algorithm or CLVQ algorithm, but in a recursive way. This means that at each iteration of the algorithm, one adds only one point to the previously computed points of the sequence, then one implements an optimization procedure keeping in mind that all the previously computed points are frozen. We give a brief idea on how to build such greedy sequences in the quadratic case when $d = 1$ and when $d \geq 2$.

One-dimensional setting When $d = 1$ and the distribution of X is absolutely continuous with a continuous positive density φ, one can implement deterministic procedures based on the knowledge of the cumulative distribution function F_X and the first moment function K_X of the distribution of X. The implementation is as follows: at the n-th iteration, we freeze the $n-1$ points of $a^{(n-1)} = \{a_1, \ldots, a_{n-1}\}$ of the sequence $(a_n)_{n\geq 1}$ which have been already computed and we sort them in an increasing order

$$a_1^{(n-1)} < \ldots < a_{n-1}^{(n-1)}.$$

Then, we compute the inter-point local inertia given by

$$\sigma_i^2 := \int_{a_i^{(n-1)}}^{a_{i+\frac{1}{2}}^{(n-1)}} |a_i^{(n-1)} - \xi|^2 \mu(d\xi) + \int_{a_{i+\frac{1}{2}}^{(n-1)}}^{a_{i+1}^{(n-1)}} |a_{i+1}^{(n-1)} - \xi|^2 \mu(d\xi), \quad i = 0, \ldots, n-1,$$

where $a_0^{(n-1)} = -\infty$, $a_n^{(n-1)} = +\infty$ and $a_{i+\frac{1}{2}}^{(n-1)}$ is the mid-point of $[a_i^{(n-1)}, a_{i+1}^{(n-1)}]$:

$$a_{\frac{1}{2}}^{(n-1)} = -\infty, \quad a_{i+\frac{1}{2}}^{(n-1)} = \frac{a_i^{(n-1)} + a_{i+1}^{(n-1)}}{2}, \quad a_{n-\frac{1}{2}}^{(n-1)} = +\infty.$$

We add a random point $\bar{a}_0$ in the inter-point zone with the maximal local inertia $(a_{i_0}^{(n-1)}, a_{i_0+1}^{(n-1)})$ where i_0 is the index such that

$$\sigma_{i_0}^2 = \max_{0\leq i\leq n-1} \sigma_i^2.$$

This point $\bar{a}_0$ is the starting point of the optimization procedure considered, which converges to the n-th point a_n of the sequence. Several procedures are detailed in the first part of Chapter 4 such as Lloyd's algorithm and CLVQ algorithm, and greedy quantization sequences of several

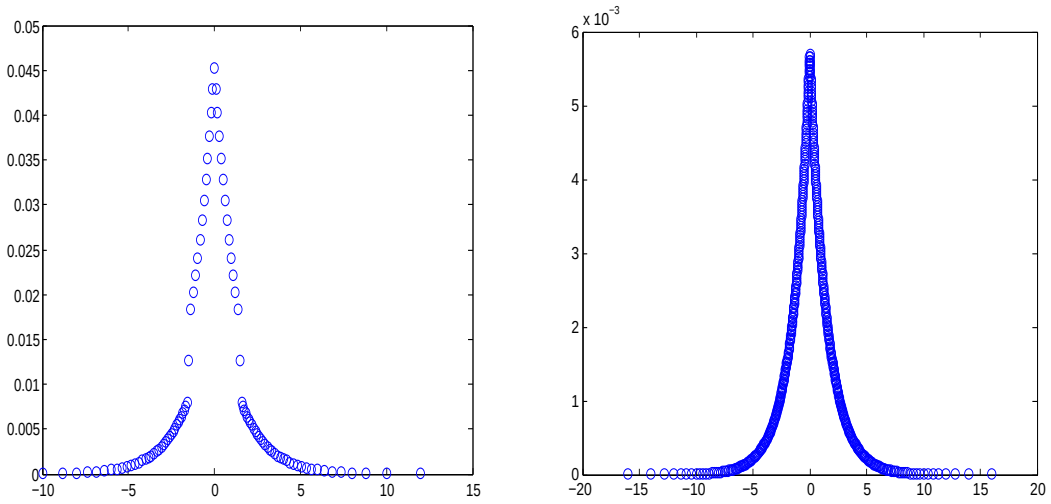

Figure 2.2: Graph of $a_i \mapsto p_i^n$ where $a^{(n)}$ is an L^2-greedy quantization sequence of Laplace$(0,1)$ and $(p_i^n)_{1\leq i\leq n}$ are the weights of the Voronoï cells for $n = 100$ (left) and $n = 511$ (right).

scalar probability distributions are computed by Lloyd's algorithm. For an example, Figure 2.2 depicts the graph representing the weights of the Voronoï cells of the first n terms ($n = 100, n = 511$) of an L^2-greedy quantization sequence of the Laplace distribution with parameters 0 and 1.

Two-dimensional setting We extend the deterministic variant of greedy algorithms to the two-dimensional case in Chapter 4. We follow the same procedure as for the scalar distributions and rely on highly effective quadrature formula to numerically compute the integrals necessary for the construction of greedy quantization sequences. In Figure 2.3, we observe a deterministic L^2-greedy quantization sequence of the standard Gaussian distribution $\mathcal{N}(0, I_2)$.

Multi-dimensional case In higher dimensions, deterministic procedures become too demanding so one switches to stochastic procedures where the computation of the integrals is replaced by large Monte Carlo simulations coupled with a nearest neighbor search. Randomized greedy Lloyd's algorithm and multi-dimensional CLVQ algorithm are explained in detail in Chapter 4. However, these procedures can be very demanding due to the several integrals that need to be computed. Although we explain, in Chapter 3, how greedy quantization allows a reduction of the number of computations at each step, this still does not make the stochastic optimization procedures easy to implement. An alternative is greedy product quantization where one relies on one-dimensional greedy quantization sequences to compute multi-dimensional sequences when the probability distribution can be written as a tensor product of its marginal laws. The multi-dimensional sequence is obtained as a result of the tensor product of multiple one-dimensional sequences. This is explained deeply in Chapters 3 and 4.

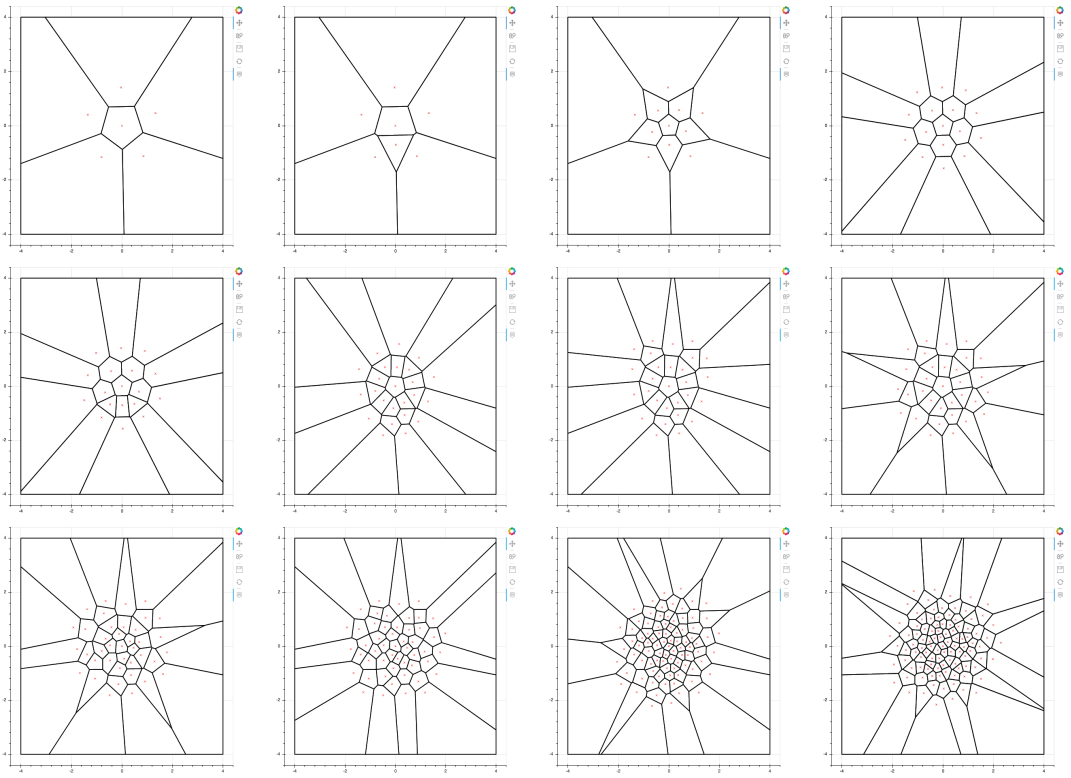

Figure 2.3: Greedy quantization sequence of $\mathcal{N}(0, I_2)$ obtained by a deterministic Lloyd's algorithm of sizes $n = 6, 7, 11, 16, 18, 24, 28, 32, 39, 51, 86, 100$ (starting from the upper left corner).

2.2.2 Contributions and new results

The first contribution of this thesis, in Chapter 3, is devoted to extending some theoretical results of L^r-greedy quantization of a distribution P with finite r-th moment to a wider class of distributions, mainly rate optimality and distortion mismatch results. An extensive numerical study is also carried out to highlight the advantages of greedy quantization sequences, compared mostly to the Monte Carlo and quasi-Monte Carlo methods. In Chapter 5, L^s-rate optimality results of L^r-dilated greedy quantization sequences are established, inspired by similar results for dilated optimal quantizers in [71].

Rate optimality and distortion mismatch (Chapter 3)

As already mentioned in Section 2.2.1, results on the rate of convergence of the greedy quantization error and the distortion mismatch problem have been established in [45]. They were based on the integrability of the b-maximal function Ψ_b defined by (2.10). In Chapter 3, based on the submitted paper [24], we extend these results to a much larger class of distributions.

Let X be a random variable with probability P defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The key in our study is to consider auxiliary probability distributions ν satisfying the following control on balls with respect to an L^r-median a_1 of P: we assume the existence of $\varepsilon_0 \in (0, 1]$ such that for every $\varepsilon \in (0, \varepsilon_0)$, there exists a Borel function $g_\varepsilon : \mathbb{R}^d \to [0, +\infty)$ satisfying, for every $x \in \mathrm{supp}(P)$

and every $t \in [0, \varepsilon\|x - a_1\|]$,

$$\nu(B(x,t)) \geq g_\varepsilon(x) V_d \, t^d. \tag{2.11}$$

This class of auxiliary distributions will be the key tool for various theoretical studies of greedy quantization sequences. Noting that the L^r-median a_1 of P belongs to $a^{(n)}$ for every $n \geq 1$ by construction of the greedy quantization sequence, we obtain an upper bound of the form

$$\forall n \geq 2, \quad e_r(a^{(n)}, P) \leq \varphi_r(\varepsilon)^{-\frac{1}{d}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{1}{d}} \left(\int g_\varepsilon^{-\frac{r}{d}} dP\right)^{\frac{1}{r}} (n-1)^{-\frac{1}{d}} \tag{2.12}$$

where V_d is the volume of the hyper-unit cube and $\varphi_r(u) = \left(\frac{1}{3^r} - u^r\right) u^d$. The proof of this result relies on a new micro-macro inequality involving the auxiliary distributions ν.

One of the main contributions presented in Chapter 3 is the extension of the non-asymptotic universal Pierce type results (2.6) for the rate of convergence of the L^r-greedy quantization error to 0. This is achieved by specifying the measure ν and the function g_ε, satisfying (2.11), in the general upper bound (2.12). For example, we can cite the following upper bound for the L^r-quantization error: If $\int |x|^{r+\delta} dP(x) < +\infty$ for some $\delta > 0$, then for every $n \geq 2$,

$$e_r(a^{(n)}, P) \leq \kappa_{d,\delta,r}^{\text{Greedy,Pierce}} \sigma_{r+\delta}(P) (n-1)^{-\frac{1}{d}},$$

for a finite constant $\kappa_{d,\delta,r}^{\text{Greedy, Pierce}}$ to be specified in Theorem 3.2.4, which relies on the measure $\nu(dx) = \gamma_{r,\delta}(x)\lambda_d(dx)$ where

$$\gamma_{r,\delta}(x) = \frac{K_{\delta,r}}{(1 \vee |x - a_1|)^{d(1+\frac{\delta}{r})}} \qquad \text{and} \qquad K_{\delta,r} = \left(\int \frac{dx}{(1 \vee |x|)^{d(1+\frac{\delta}{r})}}\right)^{-1} < +\infty,$$

and the function

$$g_\varepsilon(x) = \frac{K_{\delta,r}}{(1 \vee [(1+\varepsilon)|x - a_1|])^{d(1+\frac{\delta}{r})}}, \quad \varepsilon \in (0, \tfrac{1}{3}).$$

A sharper result, but less explicit in terms of constants, is then stated (see Theorem 3.2.4) based on distributions satisfying a "log"-integrability property of the form $\int_{\mathbb{R}^d} |x|^r (\log^+ |x|)^{\frac{r}{d}+\delta} dP(x) < +\infty$.

Finally, a hybrid Zador-Pierce result is proved for almost radial non-increasing densities, i.e. an upper bound that is non-asymptotic (Pierce-type (2.6)) with a controlling bound relying on $\|h\|_{\frac{d}{d+r}}$ as in Zador's Theorem (2.5). In other words, we establish an upper bound of the form

$$e_r(a^{(n)}, P) \leq C \|h\|_{\frac{d}{d+r}} n^{-\frac{1}{d}}$$

for some real constant C where h is the density of the absolutely continuous component of P, supposed to be radial non-increasing. A function $f : \mathbb{R}^d \to \mathbb{R}_+$ is said to be almost radial non-increasing on a set A, such that $\text{supp}(P) \subset A \subset \mathbb{R}^d$, w.r.t. some $a \in A$ (see definition 3.2.6), if there exists a norm $\|\cdot\|_0$ on $\mathbb{R}^d$ and a real constant $M \in (0, 1]$ such that

$$f(y) \geq M f(x) \qquad \text{for all } x, y \in A \setminus \{a\} \quad \text{for which} \quad \|y - a\|_0 \leq \|x - a\|_0.$$

For this purpose, we consider

$$\nu = \frac{h^{\frac{d}{d+r}}}{\int h^{\frac{d}{d+r}} d\lambda_d}.\lambda_d$$

and rely on a lower bound of $\nu(B(x,t))$, where $B(x,t)$ is the ball with center $x \in \mathbb{R}^d$ and radius $t > 0$, established in Lemma 3.2.10 of Chapter 3.

The distortion mismatch problem for this larger class of probability distributions is solved also by considering the same auxiliary distributions defined in (2.11). The results are given in Section 3.3 and we cite the following error bound for $s \in (r, d+r)$

$$e_s(a^{(n)}, P) \leq \kappa^{\text{Greedy}}_{d,r,\varepsilon} \left(\int g_\varepsilon^{-\frac{s}{d+r-s}} dP \right)^{\frac{d+r-s}{s(d+r)}} \left(\int g_\varepsilon^{-\frac{r}{d}} dP \right)^{\frac{1}{d+r}} (n-2)^{-\frac{1}{d}}$$

for every $n \geq 3$ and a finite positive constant $\kappa^{\text{Greedy}}_{d,r,\varepsilon}$ defined later in Theorem 3.3.1.

Algorithmics and numerical observations

In the second part of Chapter 3 and in Chapter 4, several numerical experiments are carried out in order to emphasize some interesting properties of one-dimensional greedy quantization sequences. Among others, we conclude numerically that, even though L^r-greedy sequences cannot be optimal at each level n, they can still be sub-optimal in the sense that there exist sub-sequences of $a^{(n)}$ which are L^r-optimal themselves. This was deduced by observing the graphs representing the weights of the Voronoï cells of these sequences. We specify these sub-sequences for the $\mathcal{N}(0,1)$ and the $\mathcal{U}([0,1])$ distributions and conclude with a conjecture concerning unimodal densities symmetric w.r.t. their L^r-median.

From another point of view, when working on the unit cube, it is natural to compare greedy quantization sequences to sequences with low discrepancy commonly used in quasi-Monte Carlo (QMC) methods. In fact, with quantization-based numerical integration, one approximates expectations of the form $\mathbb{E}f(X)$, for a Lipschitz continuous function f and a random variable X, with an $\mathcal{O}(n^{-\frac{1}{d}})$ rate of convergence to 0. While the approximation by the quasi-Monte Carlo method yields an $\mathcal{O}\left(\frac{\log n}{n^{\frac{1}{d}}}\right)$ rate of convergence, this is due to Proïnov's Theorem (see [70] or Theorem 3.4.1 in Chapter 3). The price to pay for the absence of the $(\log n)$-factor with greedy quantization is the fact that the weights of the Voronoï cells corresponding to the greedy sequence $a^{(n)}$ are not uniform (i.e. equal to $\frac{1}{n}$) which induces a higher complexity when naïvely implementing the resulting quadrature formulas. We show, in the second part of Chapter 3, how the recursivity of greedy quantization allows to reduce the number of computations so that greedy quantization sequence and QMC become equivalent in terms of complexity. Moreover, this character allows us to keep the asset of a sequence which is a recursive formula for cubatures, hence making of greedy quantization an advantageous component in the face of Quasi-Monte Carlo methods, since the error bounds are lower by a $\log(n)$ factor.

To be more precise, during the procedure of building the greedy sequence, we notice that, at each iteration, one adds a single point to the sequence while the rest remain frozen. So, the Voronoï cells, which are far from the new added point, remain untouched and unchanged. This

means that their weights, as well as the corresponding inter-point local inertia, do not need to be computed at each iteration. This remark allows to avoid a huge number of unnecessary computations at each iteration of the algorithm. Besides the dramatic reduction in the computational cost, this recursive character of greedy quantization leads us to deduce an iterative recursive formula for cubature in the one and multi-dimensional frameworks. When $d = 1$, we approximate $\mathbb{E}f(X)$ by $I_n(f)$ given by

$$I_n(f) \quad = \quad I_{n-1}(f) - p_-^n \Big(f(a_{i_0-1}^{(n)}) - f(a_{i_0}^{(n)}) \Big) - p_+^n \Big(f(a_{i_0+1}^{(n)}) - f(a_{i_0}^{(n)}) \Big),$$

where $a_{i_0}^{(n)}$ is the point added to the greedy sequence at the n-th iteration, $a_{i_0-1}^{(n)}$ and $a_{i_0+1}^{(n)}$ are the points lower and greater than $a_{i_0}^{(n)}$ and

$$p_-^n = P\Big([a_{i_0-\frac{1}{2}}^{(n)}, a_{\text{mil}}^{(n)}] \Big) \qquad \text{and} \qquad p_+^n = P\Big([a_{\text{mil}}^{(n)}, a_{i_0+\frac{1}{2}}^{(n)}] \Big)$$

where $a_{i_0\pm\frac{1}{2}}^{(n)} = \frac{a_{i_0}^{(n)} + a_{i_0\pm1}^{(n)}}{2}$ and $a_{\text{mil}}^{(n)} = \dfrac{a_{i_0+1}^{(n)} + a_{i_0-1}^{(n)}}{2}$, with $a_0 = -\infty$ and $a_n = +\infty$.
This formula can be generalized to the multidimensional framework (see (3.20)) when considering "product" greedy sequences, as explained in Chapter 3.

Moreover, note that there exists a relation between the discrepancy of a sequence Ξ and the quantization error induced by this sequence with respect to the Uniform distribution. Based on Proïnov's Theorem (see [70] and Theorem 3.4.1 in Chapter 3), it is given by

$$e_1(\Xi, \mathcal{U}([0,1]^d)) \leq D_n^*(\Xi)^{\frac{1}{d}}$$

where $D_n^*(\Xi)$ is the star-discrepancy of the sequence $\Xi = (\xi_i)_{1\leq i\leq n}$ at order n defined by

$$D_n^*(\Xi) = \sup_{u\in[0,1]^d} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\xi_i \in [0,u]^d} - \lambda_d([0,u]^d) \Big|.$$

This led us to carry out a study in view of comparison between greedy quantization sequences and sequences with low discrepancy. Two major directions are followed:

- Computing the discrepancy of greedy sequences and comparing it to that of low dicrepancy sequences,
- Treating low discrepancy sequences as (sub-optimal) quantization sequences, i.e. assigning to them a Voronoï diagram and non-uniform weights, in order to compare their performance with greedy quantization sequences.

Various numerical simulations are made and some conclusions are drown and detailed in the end of Chapter 3 and in Chapter 4. Let us say in short that, when $d = 1$, greedy quantization sequences can be used as low discrepancy sequences, and, they are better performing than low discrepancy sequences treated as quantization sequences. However, when $d \geq 2$, we don't have such optimistic results for standard greedy quantization sequences in terms of low discrepancy.

L^s-rate optimality of dilated/contracted L^r-greedy quantization sequences

In Chapter 5, we investigate the L^s-rate optimality of dilated/contracted L^r-greedy quantization sequences. This study is inspired by similar results obtained for L^r-optimal quantizers in [71], where L^r-optimal quantizers, once dilated or contracted in an appropriate way, turn out to remain L^s-rate optimal, i.e. having an $\mathcal{O}(n^{-\frac{1}{d}})$ rate of decay, for $s > r$. This may have important consequences for practical application since, usually, one has only access to quadratic optimal quantizers (like for the $\mathcal{N}(0, I_d)$ distribution, $d = 1, \ldots, 10$, on the quantization website www.quantize.maths-fi.com or for other (1D) distributions for which semi-closed forms are available (see [29, 32] for example)). One can cite, on one hand, the numerical integration field where the error bounds of quantization-based cubature formulas often involve the L^s-quantization error induced by L^r-optimal quantizers, $s > r$, which needs to be handled. On the other hand, the dilated L^r-optimal quantizers turn out to be good candidates for the initialization of the algorithms of designing L^s-quantization sequences (see [71] for further details on this topic).

The purpose of Chapter 5 is to establish similar results for L^r greedy quantization sequences. To do so, we rely on auxiliary distributions, satisfying a similar criteria to (2.11). On our way, we also generalize the seminal results from [71] taking advantage of our approach based on auxiliary functions. Let us be more precise.

Let X be a random variable with probability P defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and let $a^{(n)}$ be a corresponding L^r-optimal greedy quantization sequence of size n, $r \geq 1$. The L^r-dilated or contracted greedy quantization sequence is denoted by $a^{(n)}_{\theta,\mu}$ defined, for every $\theta > 0$ and $\mu \in \mathbb{R}^d$, by $a^{(n)}_{\theta,\mu} = \{\mu+\theta(a_i-\mu),\ a_i \in a^{(n)}\}$. Likewise, $f_{\theta,\mu}$ denotes the function $f_{\theta,\mu}(x) = f(\mu+\theta(x-\mu))$. And, if $X \sim P = f.\lambda_d$, then $P_{\theta,\mu}$ denotes the probability distribution of the random variable $\frac{X-\mu}{\theta} + \mu$ and $dP_{\theta,\mu} = \theta^d f_{\theta,\mu}.d\lambda_d$. We rely on a micro-macro inequality involving some auxiliary distributions and consider auxiliary distributions satisfying the control on balls (2.11) to obtain two main non-asymptotic L^s-rate optimality results depending on the value of s.

• Let $s \in (r, d + r)$ and P be with finite polynomial order at any order. Assume

$$\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{\frac{(d+r)(r+\delta-\eta)}{(d+r-s)(r+\delta-\eta)-ds}} f d\lambda_d < +\infty. \tag{2.13}$$

Then, for every Borel function g_ε, $\varepsilon \in (0, \frac{1}{3})$, satisfying (2.11) and every $n \geq 3$,

$$e_s(a^{(n)}_{\theta,\mu}, P) \leq \theta^{1+\frac{d}{s}} \kappa^{\text{Greedy,Pierce}}_{\theta,\mu} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{\frac{(d+r)(r+\delta-\eta)}{(d+r-s)(r+\delta-\eta)-ds}} f d\lambda_d\right)^{\frac{1}{|q|q'(d+r)}} \sigma_{r+\delta}(P)(n-2)^{-\frac{1}{d}}. \tag{2.14}$$

where $q = \frac{-s}{d+r-s}$, $q' = \frac{r+\delta-\eta}{r+\delta-\eta-d|q|}$, $p' = \frac{q'}{q'-1}$, $e_{r+\delta}(a^{(1)}, P) = \sigma_{r+\delta}(P) < +\infty$ denotes the $L^{r+\delta}$-standard deviation of P and

$$\kappa^{\text{Greedy,Pierce}}_{\theta,\mu} = 2^{\frac{1}{d}+\frac{r+\delta}{r+d}(1+\frac{1}{|q|p'})} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} \min_{\varepsilon\in(0,\frac{1}{3})} \left[(1+\varepsilon)\varphi_r(\varepsilon)^{-\frac{1}{d}}\right] \left(\int (1 \vee |x|)^{\frac{r+\delta}{r+\delta-\eta}} dx\right)^{\frac{1}{d}}.$$

• Let $s < r$. Assume

$$\int f^{-\frac{s}{r-s}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d < +\infty.$$

Then, for every distribution ν, every function g_ε satisfying (2.11) and every $n \geq 3$,

$$e_s(a^{(n)}_{\theta,\mu}, P) \leq \widetilde{\kappa}^{\text{Greedy,Pierce}}_{\theta,\mu} \theta^{1+\frac{d}{s}} \left(\int_{\{f>0\}} f^{-\frac{s}{r-s}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d \right)^{\frac{r-s}{sr}} \sigma_{r+\delta}(P)(n-2)^{-\frac{1}{d}} \tag{2.15}$$

where $e_{r+\delta}(a^{(1)}, P) = \sigma_{r+\delta}(P) < +\infty$ and

$$\widetilde{\kappa}^{\text{Greedy,Pierce}}_{\theta,\mu} = 2^{1+\frac{1}{d}+\frac{\delta}{r}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} \min_{\varepsilon \in (0,\frac{1}{3})} \left[(1+\varepsilon)\varphi_r(\varepsilon)^{-\frac{1}{d}} \right] \left(\int (1 \vee |x|)^{-d(1+\frac{\delta}{r})} dx \right)^{-\frac{1}{d}}.$$

The above results are avatars of Pierce's Lemma (2.6). A particular study for almost radial non-increasing densities yields similar results under a particular moment assumption on P.

After showing that an L^r-dilated or contracted greedy quantization sequence $a^{(n)}_{\theta,\mu}$ is L^s-rate optimal under one of the conditions mentioned above, depending on the values of s, we determine the set of parameters (θ, μ) that satisfies these conditions. In general, the optimal value for μ^* is the L^r-median of the distribution P. As for θ, this problem depends entirely on the distribution P. We lead this study for several particular density distributions and determine, for each one, the values of θ for which the dilated sequence is L^s-rate optimal. Moreover, in some cases, we show that the sequence $\alpha^{(n)}_{\theta^*,\mu}$ satisfies the so-called "empirical measure Theorem" for a particular value θ^* of θ that will be determined. This particular value θ^* allows the lower bound (5.6) induced by $\alpha^{(n)}_{\theta^*,\mu}$ to attain the sharp constant in Zador's Theorem.

For the multivariate Normal distribution $\mathcal{N}(m, I_d)$, this value is $\theta^* = \sqrt{\frac{s+d}{r+d}}$, for the hyper exponential distributions with parameters α and λ, we obtain $\theta^* = \left(\frac{s+d}{r+d}\right)^{\frac{1}{\alpha}}$ and, for the hyper gamma distribution with parameters α, β and λ, the dilated/contracted sequence satisfies the empirical measure Theorem for $\theta^* = \left(\frac{s+d}{r+}\right)^{\frac{1}{\alpha}}$, only when $\beta = \frac{d+r}{d(d+s)}$.

2.3 Recursive quantization and application to reflected BSDEs

2.3.1 Principle and existing results

Markovian quantization and recursive quantization have been originally introduced in [59] and [63] to produce spatial discretization schemes of Markov chains, typically time discretization schemes of stochastic processes like diffusion processes. Recursive quantization is a version of Markovian quantization which allows in dimension 1, but also in medium dimensions, a fast "embedded deterministic optimization" of the quantization grids involved in these numerical schemes. It has been first studied deeply in [63] for the discretization of an $\mathbb{R}^d$-valued Euler scheme of a diffusion process where the authors proposed a fast algorithm for building, in a deterministic way, the quantization tree in a one-dimensional framework. In [51], recursive quantization was extended to higher order schemes, always in the one-dimensional framework. For problems in higher dimensions, product recursive quantization has been introduced and used in [15, 28] among others.

Let us consider a Brownian diffusion process $(X_t)_{t\geq 0}$ taking values in $\mathbb{R}^d$ and solution to

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s\,, \qquad X_0 = x_0 \in \mathbb{R}^d, \tag{2.16}$$

where $b : [0,T]\times\mathbb{R}^d \to \mathbb{R}^d$ is the drift coefficient, $\sigma : [0,T]\times\mathbb{R}^d \to \mathcal{M}(d,q)$ is the matrix diffusion coefficient and $(W_t)_{t\geq 0}$ is a q-dimensional Brownian motion defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ equipped with its augmented natural filtration $(\mathcal{F}_t)_{t\geq 0}$ where $\mathcal{F}_t = \sigma(W_s, s \leq t, \mathcal{N}_{\mathbb{P}})$, $\mathcal{N}_{\mathbb{P}}$ denotes the class of all $\mathbb{P}$-negligible sets of $\mathcal{A}$. The Euler scheme associated to the process $(X_t)_{t\in[0,T]}$, with the uniform mesh $t_k = k\Delta$, $k \in \{0,\dots,n\}$ and timestep $\Delta = \frac{T}{n}$, is recursively defined by

$$\bar{X}^n_{t_{k+1}} = \bar{X}^n_{t_k} + \Delta b(t_k, \bar{X}^n_{t_k}) + \sigma(t_k, \bar{X}^n_{t_k})(W_{t_{k+1}} - W_{t_k}), \qquad \bar{X}^n_{t_0} = X_0 = x_0 \in \mathbb{R}^d. \tag{2.17}$$

Recursive quantization (as a Markovian quantization) consists in building a Markov chain having values into a *grid (or quantizer)* Γ_k of size N_k of the discrete Euler scheme $\bar{X}_{t_k}$ at time t_k. So, our goal is to optimize the grids Γ_k in a recursive way as a kind of "embedded" procedure. By "embedded" we mean that this optimization is performed "step by step" starting from time $t_0 = 0$ to time $t_n = T$. First, we denote by

$$F_k(x, \varepsilon_{k+1}) = x + \Delta b(t_k, x) + \sqrt{\Delta}\sigma(t_k, x)\varepsilon_{k+1}$$

the Euler operator with step Δ, where $(\varepsilon_k)_{0\leq k\leq n}$ is an i.i.d. sequence of random variables with distribution $\mathcal{N}(0, I_q)$, in other words, $\varepsilon_k = \sqrt{\frac{n}{T}}(W_{t_{k+1}} - W_{t_k})$. Note that this operator is with Normal distribution

$$F_k(x, \varepsilon_{k+1}) \sim \mathcal{N}(m_k, \Sigma_k)$$

where $m_k = x + \Delta b(t_k, x)$ and $\Sigma_k = \sqrt{\Delta}\sigma(t_k, x)$. The recursive quantization $(\widehat{X}^{\Gamma_k}_{t_k})_{0\leq k\leq n}$ of $(\bar{X}_{t_k})_{0\leq k\leq n}$ is performed via the following recursion: Starting at $\widehat{X}_{t_0} = \bar{X}_{t_0} = x_0$,

$$\begin{cases} \widetilde{X}_{t_k} &= F_{k-1}(\widehat{X}^{\Gamma_{k-1}}_{t_{k-1}}, \varepsilon_k), \\ \widehat{X}^{\Gamma_k}_{t_k} &= \mathrm{Proj}_{\Gamma_k}(\widetilde{X}_{t_k}), \qquad\qquad \forall k = 1,\dots,n \end{cases} \tag{2.18}$$

where Γ_k is an optimal quantizer of $\widetilde{X}_{t_k}$ of size N_k for every $k \in \{1,\dots,n\}$.

Upper bounds for the quantization error induced by the approximation of $\bar{X}_{t_k}$ by $\widehat{X}^{\Gamma_k}_{t_k}$ have been established in [63] in the quadratic framework where the authors showed that, under some Lipschitz assumptions (in x, uniformly in $t \in [0,1]$) on b and σ, one has, for every $k \in \{0,\dots,n\}$,

$$\|\bar{X}_{t_k} - \widehat{X}^{\Gamma_k}_{t_k}\|_2 \leq K\sum_{l=1}^{k} c_l N_l^{-\frac{1}{d}}$$

for finite positive constants K and c_l. For simplicity, we denote $\widehat{X}_{t_k}$ instead of $\widehat{X}^{\Gamma_k}_{t_k}$.

The construction of recursive quantizers $\widehat{X}_{t_k}$ of $\bar{X}_{t_k}$ is mainly reduced to the computation of optimal quantization grids Γ_k of $\widetilde{X}_{t_k}$ of size N_k. In the quadratic framework, it is performed by standard deterministic algorithms, such as Lloyd's algorithm or CLVQ. In this thesis, we

will mainly use the Lloyd algorithm to compute the grids $(\Gamma_k)_{1\leq k\leq n}$ in a recursive way. In fact, at time t_{k+1}, the grid $\Gamma_{k+1} = \{x_1^{k+1}, \dots, x_{N_{k+1}}^{k+1}\}$ is computed as a function of the grid $\Gamma_k = \{x_1^k, \dots, x_{N_k}^k\}$ already computed at the previous time t_k.

The main advantage of this approach is the preservation of the Markov property. The distribution of the Markov chain $(\widehat{X}_{t_k})_{0\leq k\leq n}$ is entirely characterized by the initial distribution and the transition matrices $P_k = (p_{ij}^k)_{i,j}$, for every $k \in \{1, \dots, n\}$, which constitute a very important tool in various applications. The transition probability of $(\widehat{X}_{t_k})_{0\leq k\leq n}$ from x_i^k to x_j^{k+1} between times t_k and t_{k+1} is given by

$$p_{ij}^k = \mathbb{P}\left(\widetilde{X}_{t_{k+1}} \in C_j(\Gamma_{k+1}) \,|\, \widetilde{X}_{t_k} \in C_i(\Gamma_k)\right) = \mathbb{P}\left(F_k(x_i^k, \varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\right)$$

where $\big(C_i(\Gamma_k)\big)_{1\leq i\leq N_k}$ is a Voronoï partition associated to the quantizer Γ_k at time t_k. The weights of the Voronoï cells $(p_j^{k+1})_{1\leq j\leq N_{k+1}}$ are obtained via the classical (discrete time) forward Kolmogorov equation. For every $j \in \{1, \dots, N_{k+1}\}$, one has

$$p_j^{k+1} = \mathbb{P}(\widetilde{X}_{t_{k+1}} \in C_j(\Gamma_{k+1})) = \sum_{i=1}^{N_k} p_i^k \mathbb{P}\Big(F_k(x_i^k, \varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\Big).$$

In the one-dimensional setting, the transition weights p_{ij}^k are computed based on the cumulative distribution function of the Gaussian distribution. When the dimension d grows, one relies on Monte Carlo simulations for these computations.

In Chapter 7, we give details on how to compute the recursive quantization of specific models in the one-dimensional case, like the Black-Scholes model and the CEV model, discretized following either an Euler scheme or a Milstein scheme. In Figure 2.4, we present the functions $x_i^k \mapsto p_i^k, k = 1, \dots, n$, where $(x_i^k)_{1\leq i\leq N_k}$ is the recursive quantization grid of a diffusion process following a Black Scholes model and discretized following an Euler scheme, i.e.

$$\bar{X}_{t_{k+1}} = \bar{X}_{t_k} + r\Delta\bar{X}_{t_k} + \sigma\sqrt{\Delta}\bar{X}_{t_k}\,\varepsilon_{k+1} := F_k(\bar{X}_{t_k}, \varepsilon_{k+1})$$

We consider $n = 30$ time steps and design grids of size $N_k = 50$, for every $k \in \{1, \dots, n\}$. We consider

$$T = 1, \quad X_0 = 100, \quad r = 0.006, \quad \sigma = 0.2.$$

2.3.2 Contributions of this thesis

In Chapter 6, we establish L^p-error bounds for recursive quantization for a general Markov model of the form $X_{k+1} = F_k(X_k, \varepsilon_{k+1})$, $(\varepsilon_k)_{1\leq k\leq n}$ being a sequence of i.i.d. random variables. We extend the results obtained in a L^2-framework in [63] and estimate L^p-error bounds for $p \in (1, 2 + d)$. We consider that the grids Γ_k, in (2.18), are quadratic optimal quantizers of $\widetilde{X}_{t_k}$. This is important because the stationarity property (2.7) satisfied by quadratic optimal quantizers will be necessary in our study.

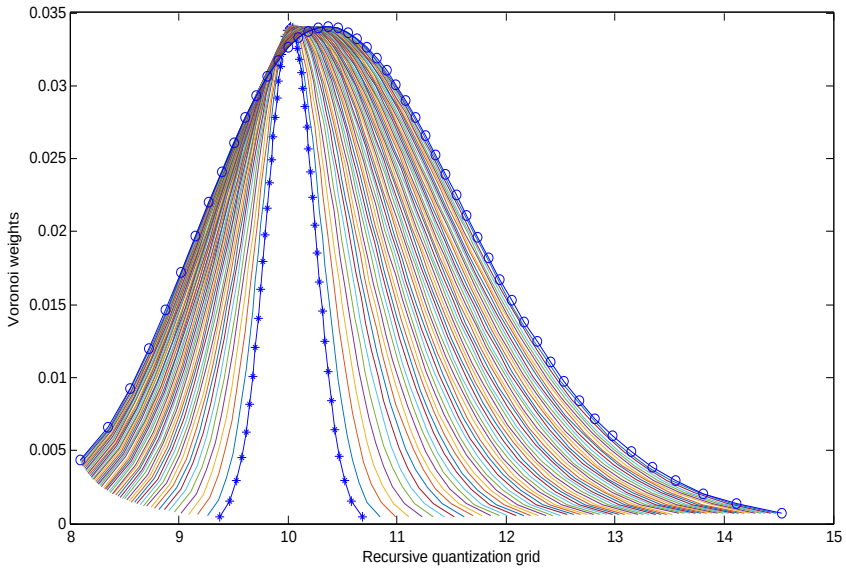

Figure 2.4: Representation of $x_i^k \mapsto p_i^k$ where $(x_i^k)_{1\leq i\leq N_k}$ is the recursive quantization grid, for every $k \in \{1,\dots,n\}$, in a Black-Scholes model (* corresponds to $k=2$ and $\circ$ corresponds to $k=n=30$).

Since we are estimating L^p-upper bounds for recursive quantization using L^2-optimal quantizers of $\widetilde{X}_{t_k}$, we find ourselves in a position where we need to handle L^p-quantization error of an L^2-optimal quantizer. For this, we rely on results on the distortion mismatch problem, also known as the $(L^r$ - $L^s)$ problem, recalled in Theorem 6.2.2. Moreover, we prove and use a technical lemma which makes possible to control the expectation of the form $\mathbb{E}|a+A\sqrt{h}Z|^r$ for some $r\geq 2$, $a\in\mathbb{R}^d$, $h>0$, $A\in\mathcal{M}(d,q,\mathbb{R})$ and $Z\in L^r_{\mathbb{R}^q}(\mathbb{P})$ an $\mathbb{R}^q$-valued random vector such that $\mathbb{E}[Z]=0$, namely

$$\mathbb{E}|a+A\sqrt{h}Z|^r \leq |a|^r\Big(1+2^{(r-3)_+}(r-1)(r-2)h\Big)+2^{(r-3)_+}(r-1)h\|A\|^r\mathbb{E}|Z|^r\left(1+\frac{r}{2}h^{\frac{r}{2}-1}\right).$$

This inequality will be of great use in proving several theoretical results to establish error bounds.

Having all the necessary tools, we show that the L^p-recursive quantization error of the Euler scheme is bounded by

$$\forall k\in\{0,\dots,n\} \qquad \|\bar{X}_{t_k}-\widehat{X}_{t_k}\|_p \leq K\sum_{l=1}^{k} C_l\|\widehat{X}_{t_l}-\widetilde{X}_{t_l}\|_p \leq K'\sum_{l=1}^{k} C_l N_l^{-\frac{1}{d}} \tag{2.19}$$

where N_l is the size of the quantizer Γ_l of $\widetilde{X}_{t_l}$ and K, K' and C_l are positive finite constants to be precised later in Theorem 6.2.1 depending on p, d, b, σ and ε_k.

When the dimension d increases, an interesting substituting technique is product recursive quantization. However, it still becomes very demanding for very high dimensions. We present and investigate an alternative, in the first part of Chapter 6, which is the *hybrid recursive*

quantization. It consists in the quantization of the white Gaussian noise in (2.18) so that the hybrid recursive quantization of $\bar{X}_{t_k}$ is given by the following recursive scheme

$$\begin{cases} \widetilde{X}_{t_k} &= F_{k-1}(\widehat{X}_{t_{k-1}}, \widehat{\varepsilon}_k), \\ \widehat{X}_{t_k} &= \mathrm{Proj}_{\Gamma_k}(\widetilde{X}_{t_k}), \qquad \forall k = 1, \dots, n. \end{cases}$$

where $(\widehat{\varepsilon}_k)_k$ is now a sequence of optimal quantizers of the Normal distribution $\mathcal{N}(0, I_q)$, which are already computed and kept off line. Based on the same tools used to establish upper bounds for the standard recursive quantization, we establish L^p-error bounds for the hybrid recursive quantization for $p \in (1, 2+d)$, namely

$$\|\bar{X}_{t_k} - \widehat{X}_{t_k}\|_p \leq K \sum_{l=1}^{k} C_X (N_l^X)^{-\frac{1}{d}} + K \sum_{l=1}^{k} C_\varepsilon (N_l^\varepsilon)^{-\frac{1}{d}}$$

where N_l^X is the size of the optimal quantizer of $\widetilde{X}_{t_l}$, N_l^ε is the size of the optimal quantizers of the Gaussian random vector and K, C_X, C_ε some finite positive constants. In practice, since the ε_k are i.i.d., we build corresponding quantizers $\hat{\varepsilon}_k$ of the same size $N_k^\varepsilon = N^\varepsilon$ for every $k \in \{1, \dots, n\}$.

2.3.3 Application to the discretization of Reflected Backward Stochastic Differential equations

Recursive quantization is a space discretization technique used in financial applications. We can cite the pricing in a Stochastic volatility model (see [15]) and the pricing of a Basket of options (see [28]). In Chapter 6, we rely on recursive quantization for the space discretization of the solution of a reflected Backward Stochastic Differential Equation (RBSDE). Approximations of these equations have already been established by several methods. For example, we can cite the regression methods with Monte Carlo simulations (see [9]), Picard iterates combined with a decomposition in Wiener chaos (see [17]) and optimal quantization (see [3, 4, 37]).

We consider the RBSDE with maturity T

$$Y_t = g(X_T) + \int_t^T f(s, X_s, Y_s, Z_s)ds + K_T - K_t - \int_t^T Z_s.dW_s\,, \qquad t \in [0, T], \tag{2.20}$$

$$Y_t \geq h(t, X_t) \quad \text{and} \quad \int_0^T (Y_s - h(s, X_s))dK_s = 0.$$

where the forward process $(X_t)_{t\in[0,T]}$ is a diffusion given by (2.16) and f, g and h are Lipschitz continuous functions. The solution of this equation is a triplet (Y_t, Z_t, K_t) and such a solution exists and is unique as established in [25] under some appropriate Lipschitz assumptions. However, this solution does not admit a closed form in general. So, one needs to approximate it by time-space discretization schemes. The time discretization scheme $(\bar{Y}_t^n, \bar{\zeta}_t^n)$ associated to (Y_t, Z_t) is based on the Euler scheme associated to the forward process $(X_t)_{t\in[0,T]}$. Several choices are possible (see [3, 9, 48]). Our choice in this work is to plug the conditional expectation inside the

driver f as follows

$$\begin{aligned}
\bar{Y}_T^n &= g(\bar{X}_T^n)\\
\widetilde{Y}_{t_k}^n &= \mathbb{E}(\bar{Y}_{t_{k+1}}^n|\mathcal{F}_{t_k}) + \Delta f\big(t_k, \bar{X}_{t_k}^n, \mathbb{E}(\bar{Y}_{t_{k+1}}^n|\mathcal{F}_{t_k}), \bar{\zeta}_{t_k}^n\big)\,, \quad k = 0, \dots, n-1,\\
\bar{\zeta}_{t_k}^n &= \frac{1}{\Delta}\mathbb{E}\big(\bar{Y}_{t_{k+1}}^n (W_{t_{k+1}} - W_{t_k})\,|\,\mathcal{F}_{t_k}\big)\,, \quad k = 0, \dots, n-1,\\
\bar{Y}_{t_k}^n &= \widetilde{Y}_{t_k}^n \vee h(t_k, \bar{X}_{t_k}^n)\,, \quad k = 0, \dots, n-1.
\end{aligned}$$

Such schemes were considered for BSDE (without reflection) in [65] or for doubly reflected BSDE in [37], whereas in most papers in the literature, the expectation is usually applied outside the driver f. In some seminal papers motivated by American options, the driver f does not depend on the process Z_t.

This scheme cannot be simulated due to the presence of conditional expectations, so we are led, like our predecessors, to perform an additional space discretization based here on a recursive quantization of the forward process $\bar{X}_{t_k}$. The fully discretized resulting scheme reads

$$\begin{aligned}
\widehat{Y}_T^n &= g(\widehat{X}_T)\\
\widehat{\zeta}_{t_k}^n &= \frac{1}{\Delta}\mathbb{E}\left(\widehat{Y}_{t_{k+1}}^n (W_{t_{k+1}} - W_{t_k})\,|\,\mathcal{F}_{t_k}\right), \quad k = 0, \dots, n-1,\\
\widehat{Y}_{t_k}^n &= \max\left(h_k(\widehat{X}_{t_k})\,, \mathbb{E}\left(\widehat{Y}_{t_{k+1}}^n\,|\,\mathcal{F}_{t_k}\right) + \Delta f\big(t_k, \widehat{X}_{t_k}, \mathbb{E}(\widehat{Y}_{t_{k+1}}^n\,|\,\mathcal{F}_{t_k}), \widehat{\zeta}_{t_k}^n\big)\right), \quad k = 0, \dots, n-1,
\end{aligned}$$

We establish upper bounds for the error induced by both time and space discretizations mentioned above.

Time discretization For the time discretization error, we establish L^2-upper bounds. To this end, we introduce a time continuous process which extends $\bar{Y}_{t_k}$, based on the martingale representation Theorem. This leads to defining a càdlàg process $\widetilde{Y}_t$ on $[t_k, t_{k+1})$ and a làdcàg process $\bar{Y}_t$ on $(t_k, t_{k+1}]$, by

$$\widetilde{Y}_t = \bar{Y}_t = \bar{Y}_{t_{k+1}} - (t_{k+1} - t) f_k(\bar{X}_{t_k}, \mathbb{E}(\bar{Y}_{t_{k+1}}\,|\,\mathcal{F}_{t_k}), \bar{\zeta}_{t_k}) - \int_t^{t_{k+1}} \bar{Z}_s dW_s, \tag{2.21}$$

leading to the following representation

$$\widetilde{Y}_t = \bar{Y}_T + \int_t^T f(\underline{s}, \bar{X}_{\underline{s}}, \mathbb{E}(\bar{Y}_{\bar{s}}\,|\,\mathcal{F}_{\underline{s}}), \bar{\zeta}_{\underline{s}})\, ds - \int_t^{t_{k+1}} \bar{Z}_s dW_s + \bar{K}_T - \bar{K}_t$$

where $\underline{s} = t_k$ and $\bar{s} = t_{k+1}$ if $s \in (t_k, t_{k+1})$, $\bar{Z}_t$ is a process such that $\mathbb{E}\sup_{[0,T]} |\bar{Z}_s|^2 < +\infty$ and $\bar{K}_{t_k}$ is an increasing càdlàg process, null at time 0, defined by

$$\bar{K}_{t_k} = \sum_{j=0}^{k} \left(h_j(\bar{X}_{t_j}) - \widetilde{Y}_{t_k}\right)_+$$

and such that $\bar{K}_t = \bar{K}_{t_k}$ for every $t \in (t_k, t_{k+1})$. This leads to the following upper bound for the time discretization error, for every $k \in \{1, \dots, n\}$,

$$\mathbb{E}|Y_{t_k} - \bar{Y}_{t_k}|^2 \le C_{b,\sigma,f,h,T}\left(\Delta + \int_0^T \mathbb{E}|Z_s - Z_{\underline{s}}|^2 ds\right)$$

where $C_{b,\sigma,f,h,T}$ is a positive real constant. This shows classically that the convergence rate of the time discretization scheme is ruled by the pathwise regularity of the process $(Z_t)_{t\in[0,T]}$ (which can be analyzed by PDE methods when b and σ are smooth enough).

Space discretization As concerns space discretization, we establish L^p-error bounds for $p \in (1, 2+d)$. It is given, for every $k \in \{1,\dots,n\}$, by

$$\|\bar{Y}_{t_k} - \widehat{Y}_{t_k}\|_p \le K \| \max_{k\le l\le n} |\bar{X}_{t_l} - \widehat{X}_{t_l}| \|_p$$

for a positive finite constant K defined later in Chapter 6. The quantities $\|\bar{X}_{t_l} - \widehat{X}_{t_l}\|_p$ are recursive quantization errors already upper-bounded in (2.19).

From an algorithmic point of view, one shows by a backward induction that there exists functions $\widehat{y}_k : \Gamma_k \mapsto \mathbb{R}, k \in \{0,\dots,n\}$, such that $\widehat{Y}_k = \widehat{y}_k(\widehat{X}_k)$, for every $k \in \{0,\dots,n\}$, recursively defined by the following Backward Dynamic Programming Principle (BDPP)

$$\begin{cases} \widehat{y}_n &= h_n \\ \widehat{y}_k &= \max\left(h_k,\, \widehat{P}_k\widehat{y}_{k+1} + \Delta f_k(.,\widehat{P}_k\widehat{y}_{k+1}, \widehat{Q}_k\widehat{y}_{k+1})\right), \quad k = 0,\dots,n-1, \end{cases}$$

where

$$\widehat{P}_k\widehat{y}_{k+1}(\widehat{X}_k) = \mathbb{E}\left(\widehat{y}_{k+1}(\widehat{X}_{k+1}) \,|\, \mathcal{F}_{t_k}\right) \qquad \text{and} \qquad \widehat{Q}_k\widehat{y}_{k+1}(\widehat{X}_k) = \frac{1}{\sqrt{\Delta}}\mathbb{E}\left(\widehat{y}_{k+1}(\widehat{X}_{k+1})\varepsilon_{k+1} \,|\, \mathcal{F}_{t_k}\right).$$

Likewise, there exists functions $\widehat{z}_k$ such that $\widehat{\zeta}_k = \widehat{z}_k(\widehat{X}_k)$, defined by

$$\widehat{z}_k = \widehat{Q}_k\widehat{y}_{k+1}.$$

Relying on this BDPP and on the recursive quantization $\widehat{X}_{t_k}^{\Gamma_k}$ of $\bar{X}_{t_k}$, $\Gamma_k = \{x_1^k,\dots,x_{N_k}^k\}$, we approximate the solution Y_0 of the RBSDE at time 0 by the initial value $\widehat{y}_0$ of the scheme

$$\begin{cases} \widehat{y}_n(x_i^n) &= h_n(x_i^n), \qquad i = 1,\dots,N_n, \\ \widehat{y}_k(x_i^k) &= \max\left(h_k(x_i^k),\, \widehat{\alpha}_k(x_i^k) + \Delta f_k(x_i^k, \widehat{\alpha}_k(x_i^k), \widehat{\beta}_k(x_i^k))\right), \qquad i = 1,\dots,N_k, \end{cases}$$

where

$$\widehat{\alpha}_k(x_i^k) = \sum_{j=1}^{N_{k+1}} \widehat{y}_{k+1}(x_j^{k+1})p_{ij}^k \qquad \text{and} \qquad \widehat{\beta}_k(x_i^k) = \frac{1}{\Delta}\sum_{j=1}^{N_{k+1}} \widehat{y}_{k+1}(x_j^{k+1})\pi_{ij}^k$$

with

$$\pi_{ij}^k = \frac{\sqrt{\Delta}}{p_i^k}\,\mathbb{E}\Big(\varepsilon_{k+1}\mathbb{1}_{\{\widehat{X}_{k+1}=x_j^{k+1},\, \widehat{X}_k = x_i^k\}}\Big) = \sqrt{\Delta}\,\mathbb{E}\Big(\varepsilon_{k+1}\mathbb{1}_{\{F_k(x_i^k,\varepsilon_{k+1})\in C_j(\Gamma_{k+1})\}}\Big).$$

We illustrate this with several one-dimensional and multi-dimensional numerical examples at the end of Chapter 6 and in Chapter 7. In the one-dimensional setting, we consider the pricing of an American call option in a market with bid-ask spread on interest rates and of an American put option under the historical probability, both examples are considered in both a Black-Scholes model and a CEV model. As for the multi-dimensional setting, we price a two-dimensional American exchange option in a Black-Scholes model and consider a multi-dimensional example

due to J.F. Chassagneux.
Moreover, we consider the pricing of American put options for $d = 1$ and $d = 2$. We show that estimates of the L^p-error bounds induced by the corresponding space discretization can be obtained directly. In fact, since both $(\bar{X}_{t_k})_{0\le k\le n}$ and $(\widehat{X}_{t_k})_{0\le k\le n}$ are Markov chains, $\bar{Y}_{t_k}$ and $\widehat{Y}_{t_k}$ can be written as the Snell envelopes: for every $k \in \{1, \dots, n\}$,

$$\bar{Y}_{t_k} = \mathbb{P}\text{-esssup}\{\mathbb{E}\,(h_\tau(\bar{X}_\tau)\,|\,\mathcal{F}_\tau),\ \tau \in \{t_k, \dots, T\}\,\mathcal{F}_\tau\text{-stopping time}\}$$

and

$$\widehat{Y}_{t_k} = \mathbb{P}\text{-esssup}\{\mathbb{E}\,(h_\tau(\widehat{X}_\tau)\,|\,\mathcal{F}_\tau),\ \tau \in \{t_k, \dots, T\}\,\mathcal{F}_\tau\text{-stopping time}\}$$

where $h(x) = \max(K - x, 0)$. Consequently, one has, for every $k \in \{1, \dots, n\}$,

$$\|\bar{Y}_{t_k} - \widehat{Y}_{t_k}\|_p \le [h]_{\mathrm{Lip}} \big\| \max_{l\ge k} |\bar{X}_{t_l} - \widehat{X}_{t_l}| \big\|_p.$$

In all the examples, we compare the results obtained by recursive quantization with others types of quantization. If $d = 1$, we compare recursive quantization to optimal, greedy and greedy recursive quantization. And, if $d > 1$, we rely on hybrid recursive quantization, instead of standard recursive quantization, and compare the results to those obtained by optimal and greedy product quantization. All the mentioned methods are detailed in Chapters 6 and 7.

Chapter 3

New approach to greedy vector quantization

This chapter corresponds to the paper "New approach to greedy vector quantization" submitted to *Bernoulli* journal and accessible on arXiv (http://arxiv.org/abs/2003.14145). It is a joint work with Harald Luschgy and Gilles Pagès.

Abstract We extend some rate of convergence results of greedy quantization sequences already investigated in 2015. We show, for a more general class of distributions satisfying a certain control, that the quantization error of these sequences have an optimal rate of convergence and that the distortion mismatch property is satisfied. We will give some non-asymptotic Pierce type estimates. The recursive character of greedy vector quantization allows some improvements to the algorithm of computation of these sequences and the implementation of a recursive formula to quantization-based numerical integration. Furthermore, we establish further properties of sub-optimality of greedy quantization sequences.

3.1 Introduction

Let $d \geq 1$, $r \in (0, +\infty)$ and $L^r_{\mathbb{R}^d}(\mathbb{P})$ (or simply $L^r(\mathbb{P})$) the set of d-dimensional random variables X defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathbb{E}\|X\|^r < +\infty$ where $\|.\|$ denotes any norm on $\mathbb{R}^d$. We denote $P = \mathbb{P}_X$ the probability distribution of X. Optimal vector quantization is a technique derived from signal processing, initially devised to optimally discretize a continuous (stationary) signal for its transmission. Originally developed in the $1950s$ (see [30]), it was introduced as a cubature formula for numerical integration in the early $1990s$ (see [54]) and for approximation of conditional expectations in the early $2000s$ for financial applications (see [4, 5]). Its goal is to find the best approximation of a continuous probability distribution by a discrete one, or in other words, the best approximation of a multidimensional random vector X by a random variable Y taking at most a finite number n of values.
Let $\Gamma = \{x_1, \ldots, x_n\}$ be a d-dimensional grid of size n. The idea is to approximate X by $q(X)$, where q is a Borel function defined on $\mathbb{R}^d$ and having values in Γ. If we consider, for q, the

nearest neighbor projection $\pi_\Gamma : \mathbb{R}^d \to \Gamma$ defined by $\pi_\Gamma(\xi) = \sum_{i=1}^{n} x_i \mathbb{1}_{W_i(\Gamma)}(\xi)$, where

$$W_i(\Gamma) \subset \{\xi \in \mathbb{R}^d : \|\xi - x_i\| \leq \min_{j \neq i} \|\xi - x_j\|\}, \qquad i = 1, \ldots, n, \tag{3.1}$$

is the Voronoï partition induced by Γ, then the Voronoï quantization of X is defined by

$$\widehat{X}^\Gamma = \pi_\Gamma(X) := \sum_{i=1}^{n} x_i \mathbb{1}_{W_i(\Gamma)}(X).$$

We will denote, most of the times, $\widehat{X}$ instead of $\widehat{X}^\Gamma$ when there is no need for specifications. The L^r-quantization error associated to the grid Γ is defined, for every $r \in (0, +\infty)$, by

$$e_r(\Gamma, X) = \|X - \pi_\Gamma(X)\|_r = \|X - \widehat{X}^\Gamma\|_r = \Big\| \min_{1 \leq i \leq n} |X - x_i| \Big\|_r \tag{3.2}$$

where $\|.\|_r$ denotes the $L^r(\mathbb{P})$-norm (or quasi-norm if $0 < r < 1$). Consequently, the optimal quantization problem comes down to finding the grid Γ that minimizes this error. It has been shown (see [32, 56, 57]) that this problem admits a solution and that the quantization error converges to 0 when the size n goes to $+\infty$. The rate of convergence is given by two well known results exposed in the following theorem.

Theorem 3.1.1. (a) Zador's Theorem (see [75]) *: Let $X \in L^{r+\eta}_{\mathbb{R}^d}(\mathbb{P})$, $\eta > 0$, with distribution P such that $dP(\xi) = \varphi(\xi) d\lambda_d(\xi) + d\nu(\xi)$. Then,*

$$\lim_{n \to +\infty} n^{\frac{1}{d}} e_{r,n}(X) = \tilde{J}_{r,d} \|\varphi\|^{\frac{1}{r}}_{L^{\frac{r}{r+d}}(\lambda_d)}$$

where $\tilde{J}_{r,d} = \inf_{n \geq 1} n^{\frac{1}{d}} e_{r,n}(U([0,1]^d)) \in (0, +\infty)$.
(b) Extended Pierce's Lemma (see [44])*: Let $r, \eta > 0$. There exists a constant $\kappa_{d,r,\eta} \in (0, +\infty)$ such that,*

$$\forall n \geq 1, \quad e_{r,n}(X) \leq \kappa_{d,r,\eta} \sigma_{r+\eta}(X) n^{-\frac{1}{d}}$$

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

However, the numerical implementation of multidimensional optimal quantizers requires the computation of grids of size $N \times d$ which becomes too expensive when N or d increase. Hence, there is a need to provide a sub-optimal solution to the quantization problem which is easier to handle and whose convergence rate remains similar (or comparable) to that induced by optimal quantizers. A so-called greedy version of optimal vector quantization has been developed in [45]. It consists this time in building a *sequence* of points $(a_n)_{n \geq 1}$ in $\mathbb{R}^d$ which is recursively optimal step by step, in the sense that it minimizes the L^r-quantization error at each iteration. This means that, having the first n points $a^{(n)} = \{a_1, \ldots, a_n\}$ for $n \geq 1$, we add, at the $(n+1)$-th step, the point a_{n+1} solution to

$$a_{n+1} \in \operatorname{argmin}_{\xi \in \mathbb{R}^d} e_r(a^{(n)} \cup \{\xi\}, X), \tag{3.3}$$

noting that $a^{(0)} = \varnothing$, so that a_1 *is simply an/the L^r-median of the distribution P of X.* The sequence $(a_n)_{n \geq 1}$ is called an L^r-optimal greedy quantization sequence for X or its distribution

P. The idea to design such an optimal sequence, which will hopefully produce quantizers with a rate-optimal behavior as n goes to infnity, is very natural and may be compared to sequences with low discrepancy in Quasi-Monte Carlo methods when working on the unit cube $[0,1]^d$. In fact, such sequences have already been developed and investigated in an L^1-setting for compactly supported distributions P as a model of short term experiment planning versus long term experiment planning (see [12]). In [45], the authors investigated independently a greedy version of vector quantization for L^r-random vectors taking values in $\mathbb{R}^d$, for numerical integration purposes. They showed that the problem (6.64) admits at least one solution $(a_n)_{n\geq 1}$ when X is an $\mathbb{R}^d$-valued random vector (the existence of such sequences can be proved in Banach spaces but, in this chapter, we only focus on $\mathbb{R}^d$). This sequence may not be unique since greedy quantization depends on the symmetry of the distribution (consider for example the $\mathcal{N}(0,1)$ distribution). However, note that, if the norm $\|.\|$ is strictly convex and $r>1$, then the L^r-median is unique. They also showed that the L^r-quantization error converges to 0 when n goes to infinity and, if $\mathrm{supp}(P)$ contains at least n elements, then the sequence $a^{(n)}$ lies in the convex hull of $\mathrm{supp}(P)$, $e_r(a^{(k)},X)$ is decreasing w.r.t. $k\in\{1,\dots,n\}$ and $P\left(\{\xi\in\mathbb{R}^d : \|\xi-a_n\|<\min_{1\leq i\leq n}\|\xi-a_i\|\}\right)>0$. Moreover, the authors showed that these sequences have an optimal rate of convergence to zero, compared to optimal quantizers, and that they satisfy the distortion mismatch problem, i.e. the property that the optimal rate of L^r-quantizers holds for L^s-quantizers for $s>r$. The proofs were based on the integrability of the b-maximal functions associated to an L^r-optimal greedy quantization sequence $(a_n)_{n\geq 1}$ given by

$$\forall \xi\in\mathbb{R}^d,\quad \Psi_b(\xi)=\sup_{n\in\mathbb{N}}\frac{\lambda_d\left(B(\xi,b\,\mathrm{dist}(\xi,a^{(n)}))\right)}{P\left(B(\xi,b\,\mathrm{dist}(\xi,a^{(n)}))\right)}. \tag{3.4}$$

In this chapter, we extend those rate of convergence and distortion mismatch results to a much larger class of functions. Instead of maximal functions, we will rely on a new micro-macro inequality involving an auxiliary probability distribution ν on $\mathbb{R}^d$. When ν satisfies an appropriate control on balls, defined later in section 3.2, we show that the rate of convergence of the L^r-quantization error of greedy sequences is $\mathcal{O}(n^{-\frac{1}{d}})$, just like the optimal quantizers. Furthermore, considering appropriate auxiliary distributions ν satisfying this control allows us to obtain Pierce type, and hybrid Zador-Pierce type, L^r-rate optimality results of the error quantization, instead of only Zador type results as given in [45].

A very important field of applications is to use these greedy sequences instead of n-optimal quantizers in quantization-based numerical integration schemes. In fact, the size of the grids used in these procedures is large in a way that the RAM storing of the quantization tree may exceed the storage capacity of the computing device. So, using greedy quantization sequences will dramatically reduce this drawback, especially since we will show that they behave similarly to optimal quantizers in terms of convergence rate. One demanding application that we can cite is the approximation of the solutions of Reflected Backward Stochastic Differential Equations (RBSDEs), including the pricing of American options, in [23], where greedy quantization proves itself to be quite performing compared to other types of quantization and more generally, to other usual numerical methods. The computation of greedy quantizers is performed by algorithms, detailed in [46], allowing also the computation of the weights $(p_i^n)_{1\leq i\leq n}$ of the Voronoï cells of the sequence $a^{(n)}$. These quantities are mandatory for the greedy quantization-based numerical

integration to approximate an integral I of a function f on $\mathbb{R}^d$ by the cubature formula

$$I(f) \approx \sum_{i=1}^{n} p_i^n f(a_i^{(n)}).$$

Compared to other methods of numerical approximation, such as quasi-Monte Carlo methods (QMC), the quantization-based methods present an advantage in terms of convergence rate, since QMC, for example, is known to induce an $\mathcal{O}\Big(\frac{\log(n+1)}{n^{\frac{1}{d}}}\Big)$ convergence rate when integrating Lipschitz functions (see [70]) while quantization -based numerical integration produces an $\mathcal{O}(n^{-\frac{1}{d}})$ rate (see [57]). However, it seems to have a drawback which is the computation of the non-uniform weights $(p_i^n)_{1\leq i\leq n}$, unlike the uniform weights in QMC (equal to $\frac{1}{n}$). In this chapter, we expose how the recursive character of greedy quantization provides several improvements to the algorithm, making it more advantageous. Moreover, this character induces the implementation of a recursive formula for numerical integration, that can replace the usual cubature formula, reducing the time and cost of the computations. This recursive formula will be introduced first in the one-dimensional case, and then extended to the multi-dimensional case for product greedy quantization sequences, computed from one-dimensional sequences, used to reduce the cost of implementations while always preserving the recursive character.

The chapter is organized as follows. We first show that greedy quantization sequences are rate optimal in section 3.2 where we extend the results presented in [45]. The distortion mismatch problem will be solved and extended in section 3.3. In section 3.4, we present the improvements applied to the algorithm of designing the greedy sequences, as well as the new approach for greedy quantization-based numerical integration. Numerical examples will illustrate the advantages brought by this new approach in section 3.5. Finally, section 3.6 is devoted to some numerical conclusions about further properties of greedy quantization sequences such as the sub-optimality, the convergence of empirical measures, the stationarity (or quasi-stationarity) and the discrepancy, to see to what extent greedy sequences can be close to optimality.

3.2 Rate optimality: Universal non-asymptotic bounds

In [45], the authors presented the rate optimality of L^r-greedy quantizers in the sense of Zador's Theorem based on the integrability of the b-maximal function $\Psi_b(\xi)$ defined by (3.4). Here, we present Pierce type non-asymptotic estimates relying on micro-macro inequalities applied to a certain class of auxiliary probability distributions ν. Different specifications of ν lead to various versions of Pierce's Lemma.
In all this section, we denote $V_d = \lambda_d(B(0,1))$ w.r.t. the norm $\|\cdot\|$. We recall, first, a micro-macro inequality that will be be used to prove the first result.

Proposition 3.2.1. *Assume $\int \|x\|^r dP(x) < +\infty$. Then, for every probability distribution ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, every $c \in (0, \frac{1}{2})$ and every $n \geq 1$*

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq \frac{(1-c)^r - c^r}{(c+1)^r} \int \nu\left(B\Big(x, \tfrac{c}{c+1} d(x, a^{(n)})\Big)\right) d(x, a^{(n)})^r dP(x).$$

Proof. Step 1: Micro-macro inequality
Let $\Gamma \subset \mathbb{R}^d$ be a finite quantizer of a random variable X with distribution P and $\Gamma_1 = \Gamma \cup \{y\}$,

$y \in \mathbb{R}^d$. For every $c \in (0, \frac{1}{2})$, we have $B(y, cd(y,\Gamma)) \subset W_y(\Gamma_1)$, where $W_y(\Gamma_1)$ is the Voronoï cell associated centroid y form a Voronoi partition induced by Γ_1, as defined by (6.11). Hence, for every $x \in B(y, cd(y,\Gamma))$, $d(x,\Gamma) \geq d(y,\Gamma) - \|x-y\| \geq (1-c)d(y,\Gamma)$. Consequently,

$$\begin{aligned} e_r(\Gamma, P)^r - e_r(\Gamma \cup \{y\}, P)^r &= \int_{\mathbb{R}^d} \left(d(x,\Gamma)^r - d(x,\Gamma_1)^r\right) dP(x) \\ &\geq \int_{W_y(\Gamma_1)} \left(d(x,\Gamma)^r - \|x-y\|^r\right) dP(x) \\ &\geq \int_{B(y,cd(y,\Gamma))} ((1-c)^r - c^r) d(y,\Gamma)^r dP(x). \end{aligned}$$

Finally, we obtain the micro-macro inequality

$$e_r(\Gamma, P)^r - e_r(\Gamma \cup \{y\}, P)^r \geq ((1-c)^r - c^r) P\left(B\left(y, cd\left(y,\Gamma\right)\right)\right) d\left(y,\Gamma\right)^r. \tag{3.5}$$

Step 2: We apply the micro-macro inequality (3.5) to the greedy quantization sequence $a^{(n)}$ and notice that $e_r(a^{(n+1)}, P) \leq e_r(a^{(n)} \cup \{y\}, P)$ for every $y \in \mathbb{R}^d$. This yields, for every $c \in \left(0, \frac{1}{2}\right)$ and every $y \in \mathbb{R}^d$,

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq ((1-c)^r - c^r)\, P\left(B\left(y, cd(y, a^{(n)})\right)\right) d(y, a^{(n)})^r.$$

We integrate this inequality with respect to ν to obtain

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq ((1-c)^r - c^r) \int P\left(B\left(y, cd(y, a^{(n)})\right)\right) d(y, a^{(n)})^r d\nu(y).$$

Now, we consider the closed sets $F_1 = \{(x,y) \in (\mathbb{R}^d)^2 : \|x-y\| \leq cd(y, a^{(n)})\}$ and $F_2 = \{(x,y) \in (\mathbb{R}^d)^2 : \|x-y\| \leq \frac{c}{c+1} d(x, a^{(n)})\}$, and notice that

$$F_2 \subset F_1 \cap \{(x,y) \in (\mathbb{R}^d)^2 : d(y, a^{(n)}) \geq \frac{1}{c+1} d(x, a^{(n)})\},$$

In fact, for $(x,y) \in F_2$,

$$d(y, a^{(n)}) \geq d(x, a^{(n)}) - \|x-y\| \geq d(x, a^{(n)}) - \frac{c}{c+1} d(x, a^{(n)}) \geq \frac{1}{c+1} d(x, a^{(n)})$$

and $\|x-y\| \leq \frac{c}{c+1} d(x, a^{(n)}) \leq cd(y, a^{(n)})$. Then,

$$\begin{aligned} \int P(B(y, cd(y, a^{(n)}))) d(y, a^{(n)})^r d\nu(y) &= \int\int \mathbb{1}_{F_1}(x,y) d(y, a^{(n)})^r d\nu(y) dP(x) \\ &\geq \frac{1}{(c+1)^r} \int\int \mathbb{1}_{F_2}(x,y) d(x, a^{(n)})^r d\nu(y) dP(x) \\ &= \frac{1}{(c+1)^r} \int \nu\Big(B\Big(x, \tfrac{c}{c+1} d(x, a^{(n)})\Big)\Big) d(x, a^{(n)})^r dP. \end{aligned}$$

In order to prove the rate optimality of the greedy quantization sequences and obtain a non-asymptotic Pierce type result, we will consider auxiliary probability distributions ν satisfying the following control on balls with respect to an L^r-median a_1 of P: for every $\varepsilon \in (0, \varepsilon_0)$, for

some $\varepsilon_0 \in (0,1]$, there exists a Borel function $g_\varepsilon : \mathbb{R}^d \to [0,+\infty)$ such that, for every $x \in \mathrm{supp}(P)$ and every $t \in [0, \varepsilon\|x - a_1\|]$,

$$\nu(B(x,t)) \geq g_\varepsilon(x) V_d t^d. \tag{3.6}$$

Of course, this condition is of interest only if the set $\{g_\varepsilon > 0\}$ is sufficiently large. Note that $a_1 \in a^{(n)}$ for every $n \geq 1$ by construction of the greedy quantization sequence. We begin by a technical lemma which will be used in the proof of the next proposition.

Lemme 3.2.2. *Let $C, \rho \in (0,+\infty)$ be some real constants and $(x_n)_{n\geq 1}$ be a non-negative sequence satisfying, for every $n \geq 1$, $x_{n+1} \leq x_n - C x_n^{1+\rho}$. Then for every $n \geq 1$,*

$$(n-1)^{\frac{1}{\rho}} x_n \leq \left(\frac{1}{C\rho}\right)^{\frac{1}{\rho}}.$$

Proof. We rely on the following Bernoulli inequalities, for every $x \geq -1$,

$$(1+x)^\rho \geq 1 + \rho x, \quad \text{if } \rho \geq 1, \qquad \text{and} \qquad (1+x)^\rho \leq 1 + \rho x, \quad \text{if } 0 < \rho < 1.$$

These inequalities can be obtained by studying the function f defined for every $x \in (-1,+\infty)$ by $f(x) = (1+x)^\rho - (1+\rho x)$. Assuming that $(x_n)_{n\geq 1}$ is non-increasing and that $x_n > 0$ for every $n \geq 1$, it follows from the assumption made on $(x_n)_{n\geq 1}$ that

$$\frac{1}{x_{n+1}^\rho} \geq \frac{1}{x_n^\rho (1 - C\, x_n^\rho)^\rho} \geq \frac{1}{x_n^\rho}(1 + C\, x_n^\rho)^\rho.$$

If $\rho \geq 1$, the Bernoulli inequalities imply $\frac{1}{x_{n+1}^\rho} \geq \frac{1}{x_n^\rho}(1 + C\,\rho\, x_n^\rho) = \frac{1}{x_n^\rho} + C\rho$. By induction, one obtains

$$\frac{1}{x_n^\rho} \geq \frac{1}{x_1^\rho} + (n-1)C\rho \geq (n-1)C\rho$$

to deduce the result easily. If $0 < \rho < 1$, then $-C\rho x_n^\rho \geq -1$ for every $n \geq 1$, and the result is deduced by using the Bernoulli inequality and then reasoning by induction.

Proposition 3.2.3. *Let P be such that $\int_{\mathbb{R}^d} \|x\|^r dP(x) < +\infty$. For any distribution ν and Borel function $g_\varepsilon : \mathbb{R}^d \to \mathbb{R}_+$, $\varepsilon \in (0, \frac{1}{3})$, satisfying (3.6),*

$$\forall n \geq 2, \quad e_r(a^{(n)}, P) \leq \varphi_r(\varepsilon)^{-\frac{1}{d}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{1}{d}} \left(\int g_\varepsilon^{-\frac{r}{d}} dP\right)^{\frac{1}{r}} (n-1)^{-\frac{1}{d}} \tag{3.7}$$

where $\varphi_r(u) = \left(\frac{1}{3^r} - u^r\right) u^d$.

Proof. We may assume that $\int g_\varepsilon^{-\frac{r}{d}} dP < +\infty$. Assume $c \in (0, \frac{\varepsilon}{1-\varepsilon}] \cap (0, \frac{1}{2})$ so that $\frac{c}{c+1} \leq \varepsilon$. Moreover $d(x, a^{(n)}) \leq d(x, a_1)$ since $a_1 \in a^{(n)}$. Consequently, for any such c, $\dfrac{c}{c+1}\, d(x, a^{(n)}) \leq \varepsilon\|x - a_1\|$ so that, by (3.6), there exists a function g_ε such that

$$\nu\left(B\left(x, \tfrac{c}{c+1}\, d(x, a^{(n)})\right)\right) \geq V_d \left(\tfrac{c}{c+1}\right)^d d(x, a^{(n)})^d\, g_\varepsilon(x).$$

Then, noting that $\frac{(1-c)^r - c^r}{(1+c)^r} \geq \frac{1}{3^r} - \left(\frac{c}{c+1}\right)^r > 0$, since $c \in (0, \frac{1}{2})$, Proposition 3.2.1 implies that

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq V_d\, \varphi_r\left(\tfrac{c}{c+1}\right) \int g_\varepsilon(x) d(x, a^{(n)})^{d+r} dP(x) \tag{3.8}$$

where $\varphi_r(u) = \left(\frac{1}{3^r} - u^r\right) u^d$, $u \in (0, \frac{1}{3})$. Applying the reverse Hölder inequality with the conjugate Hölder exponents $p = -\frac{r}{d}$ and $q = \frac{r}{r+d}$ yields

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq V_d\, \varphi_r\Big(\tfrac{c}{c+1}\Big)\Big(\int g_\varepsilon(x)^{-\frac{r}{d}} dP(x)\Big)^{-\frac{d}{r}} \big(e_r(a^{(n)}, P)^r\big)^{1+\frac{d}{r}}.$$

Then, applying lemma 3.2.2 to the sequence $x_n = e_r(a^{(n)}, P)^r$ with $C = V_d\, \varphi_r\left(\frac{c}{c+1}\right)\left(\int g_\varepsilon(x)^{-\frac{r}{d}} dP(x)\right)^{-\frac{d}{r}}$ and $\rho = \frac{d}{r}$, one obtains, for every $c \in (0, \frac{1}{2})$,

$$e_r(a^{(n)}, \mathbb{P}) \leq V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{1}{d}} \varphi_r\Big(\tfrac{c}{c+1}\Big)^{-\frac{1}{d}} \left(\int g_\varepsilon^{-\frac{r}{d}} dP\right)^{\frac{1}{r}} (n-1)^{-\frac{1}{d}}.$$

Since in most applications $\varepsilon \mapsto \left(\int g_\varepsilon^{-\frac{r}{d}} dP\right)^{\frac{1}{r}}$ is increasing on $(0, 1/3)$, we are led to study $\varphi_r\big(\frac{c}{c+1}\big)^{-\frac{1}{d}}$ subject to the constraint $c \in \big(0, \frac{\varepsilon}{1-\varepsilon}\big] \cap (0, \frac{1}{2})$. φ_r is increasing in the neighborhood of 0 and $\varphi_r(0) = 0$, so, one has, for every $\varepsilon \in (0, \frac{1}{3})$ small enough, $\varphi_r\big(\frac{c}{c+1}\big) \leq \varphi_r(\varepsilon)$, for $c \in (0, \frac{\varepsilon}{1-\varepsilon}]$. This leads to specify c as $c = \frac{\varepsilon}{1-\varepsilon}$, so that $\frac{c}{c+1} = \varepsilon$, to finally deduce the result.

By specifying the measure ν and the function g_ε, we will obtain two first natural versions of the Pierce Lemma.

Theorem 3.2.4 (Pierce's Lemma). *(a) Assume $\int_{\mathbb{R}^d} \|x\|^r dP(x) < +\infty$. Let $\delta > 0$. Then $e_r(a_1, P) = \sigma_r(P)$ and*

$$\forall n \geq 2, \qquad e_r(a^{(n)}, P) \leq \kappa^{G,P}_{d,\delta,r} \sigma_{r+\delta}(P)(n-1)^{-\frac{1}{d}}$$

where $\kappa^{G,\,P}_{d,\delta,r} \leq V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{1}{d}} \frac{\delta}{r}\Big(1 + \frac{r}{\delta}\Big)^{1+\frac{\delta}{r}} \left(\int_{\mathbb{R}^d} (\|x\| \vee 1)^{-d-\frac{d\delta}{r}} dx\right)^{\frac{1}{d}} \min\limits_{\varepsilon \in (0, \frac{1}{3})} (1+\varepsilon)\varphi_r(\varepsilon)^{-\frac{1}{d}}$.

(b) Assume $\int_{\mathbb{R}^d} \|x\|^r dP(x) < +\infty$. Let $\delta > 0$. Then

$$\forall n \geq 2,\ e_r(a^{(n)}, P) \leq \kappa^G_{d,r,\delta}\Big(\int (\|x - a_1\| \vee 1)^r \left(\log(\|x - a_1\| \vee e)\right)^{\frac{r}{d}+\delta} dP(x)\Big)^{\frac{1}{r}} (n-1)^{-\frac{1}{d}}$$

where $\kappa^G_{d,r,\delta} \leq \left(\frac{r}{dV_d}\right)^{\frac{1}{d}} \min_{\varepsilon \in (0, \frac{1}{3})} (1+\varepsilon)\varepsilon^{\frac{1}{d}+\frac{\delta}{r}} \varphi_r(\varepsilon)^{-\frac{1}{d}} \Big(\int (1 \vee \|x\|)^{-d} \big(\log(\|x\| \vee e)\big)^{-1-\frac{d\delta}{r}}\Big)^{\frac{1}{d}}$.
In particular, if $\int_{\mathbb{R}^d} \|x\|^r (\log^+ \|x\|)^{\frac{r}{d}+\delta} dP(x) < +\infty$, then

$$\limsup_n n^{\frac{1}{d}} \sup\{e_r(a^{(n)}, P) : (a_n)\ L^r\text{-optimal greedy sequence for } P\} < +\infty.$$

Proof. (a) Let $\delta > 0$ be fixed. We set $\nu(dx) = \gamma_{r,\delta}(x)\lambda_d(dx)$ where

$$\gamma_{r,\delta}(x) = \frac{K_{\delta,r}}{(1 \vee \|x - a_1\|)^{d(1+\frac{\delta}{r})}} \qquad \text{with} \qquad K_{\delta,r} = \left(\int \frac{dx}{(1 \vee \|x\|)^{d(1+\frac{\delta}{r})}}\right)^{-1} < +\infty$$

is a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$.
Let $\varepsilon \in (0,1)$ and $t > 0$. For every $x \in \mathbb{R}^d$ such that $\varepsilon\|x - a_1\| \geq t$ and every $y \in B(x,t)$, $\|y - a_1\| \leq \|y - x\| + \|x - a_1\| \leq (1+\varepsilon)\|x - a_1\|$ so that

$$\nu(B(x,t)) \geq \frac{K_{\delta,r} V_d\, t^d}{(1 \vee [(1+\varepsilon)\|x - a_1\|])^{d(1+\frac{\delta}{r})}}.$$

Hence, (3.6) is verified with $g_\varepsilon(x) = \frac{K_{\delta,r}}{(1\vee[(1+\varepsilon)\|x-a_1\|])^{d(1+\frac{\delta}{r})}}$, so we can apply Proposition 3.2.3. We have

$$\int g_\varepsilon(x)^{-\frac{r}{d}} dP(x) \le K_{\delta,r}^{-\frac{r}{d}} \int (1 \vee (1+\varepsilon)\|x-a_1\|)^{r+\delta}\, dP(x)$$

so that, applying $L^{r+\delta}$-Minkowski inequality, one obtains

$$\left(\int g_\varepsilon(x)^{-\frac{r}{d}} dP(x)\right)^{\frac{1}{r}} \le K_{\delta,r}^{-\frac{1}{d}} \left(1 + (1+\varepsilon)\sigma_{r+\delta}\right)^{1+\frac{\delta}{r}}.$$

Consequently, by Proposition 3.2.3, for $\varepsilon \in (0, 1/3)$,

$$e_r(a^{(n)}, P) \le V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{1}{d}} K_{\delta,r}^{-\frac{1}{d}} \left(1 + (1+\varepsilon)\sigma_{r+\delta}\right)^{1+\frac{\delta}{r}} \varphi_r(\varepsilon)^{-\frac{1}{d}} (n-1)^{-\frac{1}{d}} \tag{3.9}$$

Now, we introduce an equivariance argument. For $\lambda > 0$, let $X_\lambda := \lambda(X - a_1) + a_1$ and $(a_{\lambda,n})_{n\ge1} := (\lambda(a_n - a_1) + a_1)_{n\ge1}$. It is clear that $(a_{\lambda,n})_{n\ge1}$ is an L^r-optimal greedy sequence for X_λ and $e_r(a^{(n)}, X) = \frac{1}{\lambda} e_r(a_\lambda^{(n)}, X_\lambda)$. Plugging this in inequality (3.9) yields

$$\begin{aligned} e_r(a^{(n)}, P) \le & V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{1}{d}} K_{\delta,r}^{-\frac{1}{d}} \frac{1}{\lambda} \left(1 + (1+\varepsilon)\,\lambda\,\sigma_{r+\delta}\right)^{1+\frac{\delta}{r}} \varphi_r(\varepsilon)^{-\frac{1}{d}} (n-1)^{-\frac{1}{d}} \\ \le & V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{1}{d}} K_{\delta,r}^{-\frac{1}{d}} \left(\lambda^{-\frac{r}{\delta+r}} + (1+\varepsilon)\,\lambda^{\frac{\delta}{\delta+r}}\,\sigma_{r+\delta}\right)^{1+\frac{\delta}{r}} \varphi_r(\varepsilon)^{-\frac{1}{d}} (n-1)^{-\frac{1}{d}}. \end{aligned}$$

Finally, one deduces the result by setting $\lambda = \dfrac{r}{\delta} \dfrac{1}{(1+\varepsilon)\sigma_{r+\delta}}$.

(*b*) Let $\delta > 0$ be fixed. We set $\nu(dx) = \gamma_{r,\delta}(x)\lambda_d(dx)$ where

$$\gamma_{r,\delta}(x) = \frac{K_{\delta,r}}{(1 \vee \|x-a_1\|)^d \left(\log(\|x-a_1\| \vee e)\right)^{1+\frac{d\delta}{r}}}, \tag{3.10}$$

with $K_{\delta,r} = \left(\int \frac{dx}{(1\vee\|x\|)^d(\log(\|x\|\vee e))^{1+\frac{d\delta}{r}}}\right)^{-1} < +\infty$, is a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$.

Let $\varepsilon \in (0,1)$ and $t > 0$. For every $x \in \mathbb{R}^d$ such that $\varepsilon\|x - a_1\| \ge t$ and every $y \in B(x,t)$, $\|y - a_1\| \le \|y - x\| + \|x - a_1\| \le (1+\varepsilon)\|x - a_1\|$ so that

$$\begin{aligned} \nu(B(x,t)) \ge & \frac{K_{\delta,r} V_d t^d}{(1 \vee (1+\varepsilon)\|x-a_1\|)^d \left(\log((1+\varepsilon)\|x-a_1\| \vee e)\right)^{1+\frac{d\delta}{r}}} \\ \ge & \frac{K_{\delta,r} V_d t^d}{(1+\varepsilon)^d\, \varepsilon^{1+\frac{d\delta}{r}}\, (1 \vee \|x-a_1\|)^d \left(\log(\|x-a_1\| \vee e)\right)^{1+\frac{d\delta}{r}}} \end{aligned}$$

since $\log(1+\varepsilon) \le \varepsilon$. Hence, (3.6) is verified with

$$g_\varepsilon(x) = \frac{K_{\delta,r}}{(1+\varepsilon)^d\, \varepsilon^{1+\frac{d\delta}{r}}\, (1 \vee \|x-a_1\|)^d \left(\log(\|x-a_1\| \vee e)\right)^{1+\frac{d\delta}{r}}},$$

so we can apply proposition 3.2.3. We have

$$\Big(\int g_\varepsilon(x)^{-\frac{r}{d}}dP(x)\Big)^{\frac{1}{r}} \leq \frac{(1+\varepsilon)\varepsilon^{\frac{1}{d}+\frac{\delta}{r}}}{K_{\delta,r}^{\frac{1}{d}}}\Big(\int (1\vee\|x-a_1\|)^r\,(\log(\|x-a_1\|\vee e))^{\delta+\frac{r}{d}}\,dP(x)\Big)^{\frac{1}{r}}.$$

Consequently, one applies Proposition 3.2.3 to deduce the first part. For the second part of the proposition, we start by noticing that

$$(1\vee\|x-a_1\|)^r \leq (1+\|x\|+\|a_1\|)^r \leq 2^{(r-1)_+}\left(\|x\|^r+(1+\|a_1\|)^r\right)$$

and

$$\log(\|x-a_1\|\vee e) \leq \log(\|x\|\vee e)+\frac{\|a_1\|\vee e}{\|x\|\vee e} \leq \log_+\|x\|+1+\frac{\|a_1\|\vee e}{e}$$

where $\log_+ u = \log u\mathbb{1}_{u\geq 1}$, so that

$$\begin{aligned}(1\vee\|x-a_1\|)^r\,(\log(\|x-a_1\|\vee e))^{\delta+\frac{r}{d}} \leq 2^{(r-1)_+ +(\frac{r}{d}+\delta-1)_+}\Big(&\|x\|^r\log_+\|x\|^{\frac{r}{d}+\delta}+A_2\|x\|^r\\ &+A_1\log_+\|x\|^{\frac{r}{d}+\delta}+A_1A_2\Big)\end{aligned}$$

where $A_1=(1+\|a_1\|)^r$ and $A_2=\left(1+\frac{\|a_1\|\vee e}{e}\right)^{\frac{r}{d}+\delta}$. Since $\log\|x\|^{\frac{r}{d}+\delta}=\frac{1}{r}\left(\frac{r}{d}+\delta\right)\log\|x\|^r$, then $\log_+\|x\|^{\frac{r}{d}+\delta}=\frac{1}{r}\left(\frac{r}{d}+\delta\right)\log_+\|x\|^r$. Moreover, $\log_+\|x\|^r\leq\|x\|^r-1$ if $\|x\|^r\geq 1$ and equal to zero otherwise so

$$\log_+\|x\|^{\frac{r}{d}+\delta} \leq \frac{1}{r}\left(\frac{r}{d}+\delta\right)(\|x\|^r-1)_+ \leq \frac{1}{r}\left(\frac{r}{d}+\delta\right)(1+\|x\|^r).$$

Consequently,

$$(1\vee\|x-a_1\|)^r\,(\log(\|x-a_1\|\vee e))^{\delta+\frac{r}{d}} \leq 2^\beta\left(\|x\|^r\log_+\|x\|^{\frac{r}{d}+\delta}+A_1'\|x\|^r+A_2'\right)$$

where $\beta=(r-1)_+ +(\frac{r}{d}+\delta-1)_+$, $A_1'=A_2+\frac{1}{r}\left(\frac{r}{d}+\delta\right)A_2$ and $A_2'=\frac{1}{r}\left(\frac{r}{d}+\delta\right)A_1+A_1A_2$. The result is deduced from the fact that $\sup\{\|a_1\|: a_1\in\operatorname{argmin}_{\xi\in\mathbb{R}^d}e_r(\{\xi\},P)<+\infty$ (see [32, Lemma 2.2]) and $\kappa_{d,r,\delta}$ does not depend on a_1.

Remark 3.2.5. *One checks that φ_r attains its maximum at $\frac{1}{3}\left(\frac{d}{d+r}\right)^{\frac{1}{r}}$ on $(0,\frac{1}{3})$, so one concludes that $\min_{\varepsilon\in(0,\frac{1}{3})}(1+\varepsilon)\varphi_r\left(\varepsilon\right)^{-\frac{1}{d}} \leq \left(1+\frac{1}{3}\left(\frac{d}{d+r}\right)^{\frac{1}{r}}\right)3^{\frac{r}{d}+1}\left(1+\frac{d}{r}\right)^{\frac{1}{d}}\left(1+\frac{r}{d}\right)^{\frac{1}{r}}$ and $\min_{\varepsilon\in(0,\frac{1}{3})}(1+\varepsilon)\varepsilon^{\frac{1}{d}+\frac{\delta}{r}}\varphi_r\left(\varepsilon\right)^{-\frac{1}{d}} \leq \left(1+\frac{1}{3}\left(\frac{d}{d+r}\right)^{\frac{1}{r}}\right)3^{1+\frac{r-1}{d}-\frac{\delta}{r}}\left(1+\frac{d}{r}\right)^{\frac{1}{d}-\frac{1}{r}}\left(1+\frac{r}{d}\right)^{\frac{1}{r}}$.*

At this stage, one can wonder if it is possible to have a kind of hybrid Zador-Pierce result where, if $P=h.\lambda_d$, one has

$$e_r(a^{(n)},P) \leq C\|h\|_{\frac{d}{d+r}}n^{\frac{1}{d}}$$

for some real constant C. To this end, we have to consider

$$\nu=\frac{h^{\frac{d}{d+r}}}{\int h^{\frac{d}{d+r}}d\lambda_d}.\lambda_d.$$

This is related to the following local growth control condition of densities.

Definition 3.2.6. *Let $A \subset \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}_+$ is said to be almost radial non-increasing on A w.r.t. $a \in A$ if there exists a norm $\|.\|_0$ on $\mathbb{R}^d$ and real constant $M \in (0,1]$ such that*

$$\forall x \in A \setminus \{a\}, \quad f_{|B_{\|.\|_0}(a,\|x-a\|_0)\cap(A\setminus\{a\})} \geq Mf(x). \tag{3.11}$$

If (3.11) holds for $M = 1$, then f is called radial non-increasing on A w.r.t. a.

Remark 3.2.7. *(a) (3.11) reads $f(y) \geq Mf(x)$ for all $x, y \in A\setminus\{a\}$ for which $\|y-a\|_0 \leq \|x-a\|_0$.*
(b) If f is radial non-increasing on $\mathbb{R}^d$ w.r.t. $a \in \mathbb{R}^d$ with parameter $\|.\|_0$, then there exists a non-increasing measurable function $g : (0, +\infty) \to \mathbb{R}_+$ satisfying $f(x) = g(\|x - a\|_0)$ for every $x \neq a$.
(c) From a practical point of view, many classes of distributions satisfy (3.11), e.g. the d-dimensional normal distribution $\mathcal{N}(m, \sigma_d)$ for $h(y) = (2\pi)^{-\frac{d}{2}}\det(\sigma_d)^{-\frac{1}{2}}e^{-\frac{y^2}{2}}$ and density $f(x) = h(\|x-m\|_0)$ where $\|x\|_0 = \|\sigma_d^{-\frac{1}{2}}x\|$, and the family of distributions defined by $f(x) \propto \|x\|^c e^{-a\|x\|^b}$, for every $x \in \mathbb{R}^d, a, b > 0$ and $c > -d$, for which one considers $h(u) = u^c e^{-au^b}$. In the one dimensional case, we can mention the Gamma distribution, the Weibull distributions, the Pareto distributions and the log-normal distributions.

Theorem 3.2.8. *Assume $P = h.\lambda_d$ with $h \in L^{\frac{d}{d+r}}(\lambda_d)$ and $\int_{\mathbb{R}^d} \|x\|^r dP(x) < +\infty$. Let a_1 denote the L^r-median of P. Assume that supp$(P) \subset A$ and $a_1 \in A$ for some A star-shaped and peakless with respect to a_1 in the sense that*

$$\mathfrak{p}(A, \|. - a_1\|) := \inf\left\{\frac{\lambda_d(B(x,t)\cap A)}{\lambda_d(B(x,t))}; x \in A, 0 < t \leq \|x - a_1\|\right\} > 0. \tag{3.12}$$

Assume h is almost radial non-increasing on A with respect to a_1 in the sense of (3.11). Then,

$$\forall\, n \geq 2, \quad e_r(a^{(n)}, P) \leq \kappa^{G,Z,P}_{d,r,M,C_0,\mathfrak{p}(A,\|.-a_1\|)}\, \|h\|^{\frac{1}{r}}_{L^{\frac{d}{d+r}}(\lambda_d)} (n-1)^{-\frac{1}{d}},$$

where $\kappa^{G,Z,P}_{d,r,M,C_0,\mathfrak{p}(A,\|.-a_1\|)} \leq \dfrac{2C_0^2\, r^{\frac{1}{d}}}{d^{\frac{1}{d}} M^{\frac{d+r}{d}} V_d^{\frac{r}{d}} \mathfrak{p}(A,\|.-a_1\|)^{\frac{1}{d}}} \min_{\varepsilon\in(0,\frac{1}{3})} \varphi_r(\varepsilon)^{-\frac{1}{d}}$.

Remark 3.2.9. *(a) If $A = \mathbb{R}^d$, then $\mathfrak{p}(A, \|. - a\|) = 1$ for every $a \in \mathbb{R}^d$.*
(b) The most typical unbounded sets satisfying (3.12) are convex cones that is cones $K \subset \mathbb{R}^d$ of vertex 0 with $0 \in K$ ($K \neq \varnothing$) and such that $\lambda x \in K$ for every $x \in K$ and $\lambda \geq 0$. For such convex cones K with $\lambda_d(K) > 0$, we even have that the lower bound

$$\mathfrak{p}(K) := \inf\left\{\frac{\lambda_d(B(x,t)\cap K)}{\lambda_d(B(x,t))}; x \in K,\, t > 0\right\} = \frac{\lambda_d\big(B(0,1)\cap K)\big)}{V_d} > 0.$$

Thus if $K = \mathbb{R}^d_+$, then $\mathfrak{p}(K) = 2^{-d}$.

The proof of Theorem 3.2.8 is based on the following lemma.

Lemme 3.2.10. *Let $\nu = f.\lambda_d$ be a probability measure on $\mathbb{R}^d$ where f is almost radial non-increasing on $A \in \mathcal{B}(\mathbb{R}^d)$ w.r.t. $a_1 \in A$, A being star-shaped relative to a_1 and satisfying (3.12). Then, for every $x \in A$ and positive $t \in (0, \|x - a_1\|]$,*

$$\nu(B(x,t)) \geq M\mathfrak{p}(A, \|. - a_1\|)(2C_0^2)^{-d} V_d f(x) t^d$$

where $C_0 \in [1, +\infty)$ satisfies, for every $x \in \mathbb{R}^d$, $\dfrac{1}{C_0}\|x\|_0 \leq \|x\| \leq C_0\|x\|_0$.

Proof. For every $x \in A$ and $t > 0$,

$$\nu(B(x,t)) \geq \int_{B(x,t)\cap A\cap\{f\geq Mf(x)\}} f d\lambda_d \;\geq\; Mf(x)\lambda_d(B(x,t)\cap A\cap\{f \geq Mf(x)\})$$

and $B(x,t)\cap(A\setminus\{a_1\})\cap B_{\|.\|_0}(a_1,\|x-a_1\|_0) \subset B(x,t)\cap A\cap\{f \geq Mf(x)\}$. Now, assume $0 < t \leq \|x-a_1\| \leq C_0\|x-1\|_0$. Setting $x' := \left(1-\frac{t}{2C_0\|x-a_1\|_0}\right)x + \frac{t}{2C_0\|x-a_1\|_0}a_1 \in A$ (since A is star-shaped with respect to a_1), we notice that, for $y \in B\left(x', \frac{t}{2C_0^2}\right) \subset B_{\|.\|_0}\left(x', \frac{t}{2C_0}\right)$,

$$\|y-x\| \leq \|y-x'\| + C_0\|x'-x\|_0 \leq \frac{t}{2C_0^2} + C_0\left\|\frac{t}{2C_0\|x-a_1\|_0}(x-a_1)\right\|_0 = \frac{t}{2C_0^2} + \frac{t}{2} \leq t$$

and $\|y-a_1\|_0 \leq \|y-x'\|_0 + \|x'-a_1\|_0 \leq \frac{t}{2C_0} + \left\|\left(1-\frac{t}{2C_0\|x-a_1\|_0}\right)(x-a_1)\right\|_0 = \|x-a_1\|_0$, so that, $B\left(x', \frac{t}{2C_0^2}\right) \subset B(x,t)\cap B_{\|.\|_0}(a_1,\|x-a_1\|_0)$. Consequently,

$$\nu(B(x,t)) \geq Mf(x)\lambda_d\left(B\left(x', \frac{t}{2C_0^2}\right)\cap A\right).$$

Moreover, $\frac{t}{2C_0^2} \leq \frac{t}{2} \leq \frac{1}{2}\|x-a_1\| \leq \|x'-a_1\|$. Hence, we have

$$\lambda_d\Big(B\Big(x', \tfrac{t}{2C_0^2}\Big)\cap A\Big) \geq \mathfrak{p}(A,\|.-a_1\|)\lambda_d\Big(B\Big(x', \tfrac{t}{2C_0^2}\Big)\Big) = \mathfrak{p}(A,\|.-a_1\|)(2C_0^2)^{-d}t^d\lambda_d(B(0,1)).$$

Proof of theorem 3.2.8. Consider $\nu = h_r.\lambda_d := \frac{h^{\frac{d}{d+r}}}{\int h^{\frac{d}{d+r}}d\lambda_d}.\lambda_d$. Notice that h_r is alsmost radial non-increasing on A w.r.t. a_1 with parameter $M^{\frac{d}{d+r}}$ so that Lemma 3.2.10 yields for every $x \in A$ and $t \in [0, \|x-a_1\|]$

$$\nu(B(x,t)) \geq M^{\frac{d}{d+r}}\mathfrak{p}(A,\|\cdot-a_1\|)(2C_0^2)^{-d}V_d h_r(x)t^d.$$

Consequently, using the fact that $\int_{\mathbb{R}^d} h_r^{-\frac{r}{d}} dP = \|h\|_{L^{\frac{d}{d+r}}(\lambda_d)}$, the assertion follows from Proposition 3.2.3. □

Remark 3.2.11. *Note that, by applying Hölder inequality with the conjugate exponents $p = 1+\frac{r}{d}$ and $q = 1+\frac{d}{r}$, one has*

$$\int_{\mathbb{R}^d} h(\xi)^{\frac{d}{d+r}}d\xi \leq \left(\int_{\mathbb{R}^d} h(\xi)(1\vee|\xi|)^{r+\delta}d\xi\right)^{\frac{d}{d+r}}\left(\int_{\mathbb{R}^d}\frac{d\xi}{(1\vee|\xi|)^{d(1+\frac{\delta}{r})}}\right)^{\frac{r}{d+r}}.$$

Consequently, since $\int_{\mathbb{R}^d}\frac{d\xi}{(1\vee|\xi|)^{d(1+\frac{\delta}{r})}} < +\infty$, one deduces that $\|h\|_{\frac{d}{d+r}}^{\frac{1}{r}} \asymp \sigma_{r+\delta}^{1+\frac{\delta}{r}}$.

We note that Zador Theorem implies $\liminf_n n^{\frac{1}{d}}e_r(a^{(n)},P) \geq \liminf_n n^{\frac{1}{d}}e_{r,n}(P,\mathbb{R}^d) \geq Q_r(P)^{\frac{1}{r}}$. The next proposition may appear as a refinement of Pierce's Lemma and Theorem 3.2.8 in the sense that it gives a lower convergence rate for the discrete derivative of the quantization error, that is its increment.

Proposition 3.2.12. *Assume $\int_{\mathbb{R}^d} \|x\|^r dP(x) < +\infty$. Then,*

$$\liminf_n n^{1+\frac{r}{d}} \min_{1\le i\le n} \left(e_r(a^{(i)},P)^r - e_r(a^{(i+1)},P)^r\right) > 0.$$

Proof. We start by choosing $N>0$ such that $P(B(0,N))>0$. Proposition 3.2.1 yields, for every probability measure ν on $\mathbb{R}^d$, for every $n\ge n_0$ and $c\in\left(0,\frac{1}{2}\right)$,

$$\begin{aligned} &e_r(a^{(n)},P)^r - e_r(a^{(n+1)},P)^r \\ &\qquad \ge \frac{(1-c)^r-c^r}{(c+1)^r}\int_{B(0,N)\cap supp(P)} \nu\left(B\left(x,\tfrac{c}{c+1}d(x,a^{(n)})\right)\right) d(x,a^{(n)})^r dP. \end{aligned}$$

We choose $\nu=\mathcal{U}(B(0,N))$. Then, for every $x\in B(0,N)$, $t\le N$ and $x'=\left(1-\frac{t}{2N}\right)x$, one has $B\left(x',\frac{t}{2}\right)\subset B(x,t)\cap B(0,N)$ since, for every $y\in B\left(x',\frac{t}{2}\right)$,

$$\|y-x\|\le \|y-x'\|+\|x'-x\|\le \frac{t}{2}+\frac{t}{2N}\|x\|\le t$$

and

$$\|y\|\le \|y-x'\|+\|x'\|\le \frac{t}{2}+\left(1-\frac{t}{2N}\right)\|x\|\le \frac{t}{2}+\left(1-\frac{t}{2N}\right)N\le N.$$

Consequently,

$$\nu(B(x,t))\ge \frac{\lambda_d(B(x',\frac{t}{2}))}{\lambda_d(B(0,N))}=(2N)^{-d}t^d.$$

Moreover, we denote $C:=\sup_{n\ge1}\max_{x\in B(0,N)\cap supp(P)} d(x,a^{(n)})$ which is finite because $a^{(n)}\in \overline{conv}(supp(P))$. Consequently, for every $c\in\left(0,\frac{1}{2}\right)$ such that $\frac{c}{c+1}C\le N$ and every $n\ge n_0$,

$$\begin{aligned} e_r(a^{(n)},P)^r - e_r(a^{(n+1)},P)^r &\ge \frac{(1-c)^r-c^r}{(c+1)^r}\left(\tfrac{c}{c+1}\right)^d (2N)^{-d}\int_{B(0,N)} d(x,a^{(n)})^{d+r}dP(x) \\ &\ge \varphi\left(\tfrac{c}{c+1}\right)(2N)^{-d}P(B(0,N))e^{d+r}_{d+r}(a^{(n)},P(.|B(0,N))). \end{aligned}$$

Finally, one deduces the result using that $\left(e^{d+r}_{d+r}\left(a^{(n)},P(.|B(0,N))\right)\right)_{n\ge1}$ is nonincreasing and relying on Zador's Theorem.

Remark 3.2.13. *For every $m,n\in\mathbb{N}$, if we denote $W_b(a^{(n)})$ the Voronoï cell associated to the sequence $a^{(n)}$ of centroid $b\in a^{(n)}$ and use the fact that $e_r(a^{(n+1)},X)\le e_r(a^{(n)}\cup\{b\},X)$ for every $b\in\mathbb{R}^d$, we deduce*

$$\begin{aligned} e_r(a^{(n)},X)^r - e_r(a^{(n+m)},X)^r &= \sum_{b\in a^{(n)}}\int_{W_b(a^{(n+m)})}\left(d(x,a^{(n)})^r-\|x-b\|^r\right)dP \\ &\quad+\sum_{b\in a^{(n+m)}\setminus a^{(n)}}\int_{W_b(a^{(n+m)})}\left(d(x,a^{(n)})^r-\|x-b\|^r\right)dP \\ &= \sum_{b\in a^{(n+m)}\setminus a^{(n)}}\int_{W_b(a^{(n+m)})}\left(d(x,a^{(n)})^r-d(x,a^{(n)}\cup\{b\})^r\right)dP \\ &\le m\left(e_r(a^{(n)},X)^r-e_r(a^{(n+1)},X)^r\right). \end{aligned}$$

Consequently, since $l\to e_r(a^{(l)},X)$ is non-increasing, one considers $n=i$ and deduces

$$\min_{1\le i\le n}\left(e_r(a^{(i)},X)^r-e_r(a^{(i+1)},X)^r\right)\ge\frac{1}{m}\left(e_r(a^{(n)},X)^r-e_r(a^{(m)},X)^r\right).$$

3.3 Distortion mismatch

We address now the problem of distortion mismatch, i.e. the property that the rate optimal decay property of L^r-quantizers remains true for $L^s(P)$-quantization error for $s \in (0, +\infty)$. This problem was originally investigated in [33] for optimal quantizers. If $s \leq r$, the monotonicity of the L^s-norm as a function of s ensures that any L^r-optimal greedy sequence remains L^s-rate optimal for the L^s-norm. The challenge is when s is larger than r. The problem is solved in [45] for $s \in (0, +\infty)$ relying on an integrability assumption of the b-maximal function Ψ_b. Here, we give an additional nonasymptotic result for $s \in (r, d+r)$, in the same settings as for Theorem 3.2.3, considering auxiliary probability distributions ν satisfying (3.6).

Theorem 3.3.1. *Let P be such that $\int_{\mathbb{R}^d} \|x\|^r dP(x) < +\infty$. Let $s \in (r, d+r)$. Let (a_n) be an L^r-optimal greedy sequence for P. For any distribution ν and Borel function $g_\varepsilon : \mathbb{R}^d \to \mathbb{R}_+$, $\varepsilon \in (0, \frac{1}{3})$, satisfying (3.6), for every $n \geq 3$,*

$$e_s(a^{(n)}, P) \leq \kappa_{d,r,\varepsilon}^{Greedy} \left(\int g_\varepsilon^{-\frac{s}{d+r-s}} dP \right)^{\frac{d+r-s}{s(d+r)}} \left(\int g_\varepsilon^{-\frac{r}{d}} dP \right)^{\frac{1}{d+r}} (n-2)^{-\frac{1}{d}}$$

where $\kappa_{d,r,\varepsilon}^{Greedy} = 2^{\frac{1}{d}} \ (\frac{r}{d})^{\frac{r}{d(d+r)}} \ V_d^{-\frac{1}{d}} \varphi_r(\varepsilon)^{-\frac{1}{d}}$.

Proof. We assume $\frac{1}{g_\varepsilon} \in L^{\frac{s}{d+r-s}}(P)$ so that $\frac{1}{g_\varepsilon} \in L^{\frac{r}{d}}(P)$ since $\frac{s}{d+r-s} \geq \frac{s}{d} \geq \frac{r}{d}$. Inequality (3.8) from the proof of Proposition 3.2.3 still holds, i.e.

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq C \int g_\varepsilon(x) d\left(x, a^{(n)}\right)^{d+r} dP(x).$$

with, for every $c \in (0, \frac{\varepsilon}{1-\varepsilon}] \cap (0, 1/2)$, $C = V_d \, \varphi_r\left(\frac{c}{c+1}\right)$ where $\varphi_r(u) = \left(\frac{1}{3^r} - u^r\right) u^d$. The reverse Hölder inequality applied with $p = \frac{s}{d+r} \in (0,1)$ and $q = -\frac{s}{d+r-s} \in (-\infty, 0)$ yields that

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq C_1 e_s(a^{(n)}, P)^{d+r}$$

where $C_1 = C \left(\int g_\varepsilon^{-\frac{s}{d+r-s}} dP \right)^{-\frac{d+r-s}{s}}$. Hence, knowing that $k \mapsto e_s(a^{(k)}, P)$ is non-increasing and summing between n and $2n-1$, we obtain for $n \geq 1$

$$n \, e_s(a^{(2n-1)}, P)^{d+r} \leq \sum_{k=n}^{2n-1} e_s(a^{(k)}, P)^{d+r} \leq \frac{1}{C_1} \sum_{k=n}^{2n-1} e_r(a^{(k)}, P)^r - e_r(a^{(k+1)}, P)^r \leq \frac{1}{C_1} e_r(a^{(n)}, P)^r.$$

Finally, since $2\left\lceil \frac{n}{2} \right\rceil - 1 \leq n$, we have $e_s(a^{(n)}, P) \leq e_s\left(a^{2\lceil \frac{n}{2} \rceil - 1}, P\right)$ and we derive that

$$\frac{n}{2} e_s(a^{(n)}, P)^{d+r} \leq \left\lceil \frac{n}{2} \right\rceil e_s(a^{(n)}, P)^{d+r} \leq \left\lceil \frac{n}{2} \right\rceil e_s\left(a^{2\lceil \frac{n}{2} \rceil - 1}, P\right)^{d+r} \leq \frac{1}{C_1} e_r\left(a^{\lceil \frac{n}{2} \rceil}, P\right)^r.$$

Consequently, plugging in C_1,

$$e_s(a^{(n)}, P) \leq 2^{\frac{1}{d+r}} V_d^{-\frac{1}{d+r}} \varphi_r\left(\frac{c}{c+1}\right)^{-\frac{1}{d+r}} \left(\int g_\varepsilon^{-\frac{s}{d+r-s}} dP \right)^{\frac{d+r-s}{s(d+r)}} n^{-\frac{1}{d+r}} e_r\left(a^{\lceil \frac{n}{2} \rceil}, P\right)^{\frac{r}{d+r}}.$$

Consequently, one can deduce from Proposition 3.2.3, for $n \geq 3$,

$$e_s(a^{(n)}, P) \leq \frac{2^{\frac{1}{d}} r^{\frac{r}{d(d+r)}}}{V_d^{\frac{1}{d}} d^{\frac{r}{d(d+r)}}} \left(\int g_\varepsilon^{-\frac{s}{d+r-s}} dP \right)^{\frac{d+r-s}{s(d+r)}} \left(\int g_\varepsilon^{-\frac{r}{d}} dP \right)^{\frac{1}{d+r}} \varphi_r\left(\tfrac{c}{c+1}\right)^{-\frac{1}{d}} (n-2)^{-\frac{1}{d}}.$$

Hence, the result is owed to the fact that $\varphi_r\left(\frac{c}{c+1}\right) \leq \varphi_r(\varepsilon)$ for $c \in (0, \frac{\varepsilon}{1-\varepsilon}]$.

Corollary 3.3.2. *Let $s \in (r, d+r)$. Assume that $\int \|x\|^{\frac{ds}{d+r-s}} (\log^+\|x\|)^{\frac{s}{d+r-s}+\delta} dP(x) < +\infty$, for $\delta > 0$, then*

$$\limsup_n n^{\frac{1}{d}} \sup\{e_s(a^{(n)}, P) : (a_n) L^r\text{-optimal greedy sequence for } P\} < +\infty.$$

Proof. The proof is divided in two steps.
Step 1: Let $\delta > 0$ be fixed and $\beta = 1 + \frac{(d+r-s)\delta}{s}$. We consider $\nu(dx) = \gamma_{r,\delta}(x)\lambda_d(dx)$ where $\gamma_{r,\delta}(x)$ is a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$ defined by (3.10) in the proof of Theorem 3.2.4(b). The density $\gamma_{r,\delta}$ is radial non-increasing on the whole $\mathbb{R}^d$ w.r.t. a_1 (and $\|\cdot\|_0 = \|\cdot\|$) so that $\mathfrak{p}(\|\cdot - a_1\|) = 1$ by Remark 3.2.9(a) and, in turn, Lemma 3.2.10 yields for every $x \in \mathbb{R}^d$ and $t \le \|x - a_1\|$

$$\nu(B(x,t)) \ge 2^{-d} V_d \gamma_{r,\delta}((x) t^d.$$

Consequently, Theorem 3.3.1 yields, for $n \ge 3$,

$$e_s(a^{(n)}, P) \le C_{d,r,\delta} \left(\int (1 \vee \|x - a_1\|)^r \left(\log(\|x - a_1\| \vee e)\right)^{\beta \frac{r}{d}} dP(x) \right)^{\frac{1}{d+r}}$$
$$\times \left(\int (1 \vee \|x - a_1\|)^{\frac{sd}{d+r-s}} \left(\log(\|x - a_1\| \vee e)\right)^{\delta + \frac{s}{d+r-s}} dP(x) \right)^{\frac{d+r-s}{s(d+r)}} (n-2)^{-\frac{1}{d}}$$

where $C_{d,r,\delta} \le 2^{1+\frac{1}{d}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} K_{\delta,r}^{-\frac{1}{d}} \min_{\varepsilon \in (0,\frac{1}{3})} (1+\varepsilon)^d \varepsilon^{\frac{\beta}{d}} \varphi_r(\varepsilon)^{-\frac{1}{d}}$.
Step 2: Just as in the proof of Theorem 3.2.4(b), we have

$$(1 \vee \|x - a_1\|)^r \left(\log(\|x - a_1\| \vee e)\right)^{\beta \frac{r}{d}} \le 2^{(r-1)_+ + (\beta \frac{r}{d} - 1)_+} \left(\|x\|^r \log_+ \|x\|^{\beta \frac{r}{d}} + A_1 \|x\|^r + A_2 \right)$$

and, denoting $\beta' = \delta + \frac{s}{d+r-s}$,

$$(1 \vee \|x - a_1\|)^{\frac{sd}{d+r-s}} \left(\log(\|x - a_1\| \vee e)\right)^{\beta'} \le 2^{(\frac{ds}{d+r-s} - 1)_+ + (\beta' - 1)_+}$$
$$\times \left(\|x\|^{\frac{ds}{d+r-s}} \log_+ \|x\|^{\beta'} + B_1 \|x\|^r + B_2 \right)$$

where A_1, A_2, B_1 and B_2 are constants depending only on r, d, s, δ and a_1. Since, $\frac{s}{d+r-s} \ge \frac{r}{d}$, one has $\frac{ds}{d+r-s} > r$ and $\delta + \frac{s}{d+r-s} \ge \beta \frac{r}{d}$, so that the two above quantities are finite (by the assumption made in the theorem).The result is deduced from the fact that $\sup\{\|a_1\| : a_1 \in \operatorname{argmin}_{\xi \in \mathbb{R}^d} e_r(\{\xi\}, P)\} < \infty$.

3.4 Algorithmics

An important application of quantization is numerical integration. Let us consider the quadratic case $r = 2$ and an L^2-optimal greedy quantization sequence $a^{(n)}$ for a random variable X with distribution $\mathbb{P}_X = P$. Since we know that $e_2(a^{(n)}, X) = \|X - a^{(n)}\|_2$ converges to 0 when n goes to infinity, this means that $a^{(n)}$ converges towards X in L^2 and hence in distribution. So, one can approximate $\mathbb{E}[f(X)]$, for every continuous function $f : \mathbb{R}^d \to \mathbb{R}$, by the following cubature formula

$$I(f) := \mathbb{E}[f(X)] \approx \sum_{i=1}^n p_i^n f(a_i^{(n)}) \tag{3.13}$$

where, for every $i \in \{1, \dots, n\}$, $p_i^n = P(X \in W_i(a^{(n)}))$ represents the weight of the i^{th} Voronoï cell corresponding to the greedy quantization sequence $a^{(n)} = \{a_1^{(n)}, \dots, a_n^{(n)}\}$. When the function f satisfies certain regularities, one establishes error bounds for this quantization-based cubature formula and obtains an $\mathcal{O}(n^{-\frac{1}{d}})$ rate of convergence, we refer to [57] for details.

When working on the unit cube $[0,1]^d$, it is natural to compare an optimal greedy sequence of the Uniform distribution $\mathcal{U}([0,1]^d)$ and a uniformly distributed sequence with low discrepancy used in the quasi-Monte Carlo method (QMC). A $[0,1]^d$-valued sequence $\xi = (\xi_n)_{n\geq 1}$ is uniformly distributed if $\mu_n = \frac{1}{n}\sum_{k=1}^n \delta_{\xi_k}$ converges weakly to $\lambda_{d_{|[0,1]^d}}$ (where λ_d denotes the Lebesgue measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$). It is well known (see [40] for example) that $(\xi_n)_{n\geq 1}$ is uniformly distributed if and only if

$$D_n^*(\xi) = \sup_{u\in[0,1]^d} \Big| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\xi_i \in [0,u]^d} - \lambda_d([0,u]^d) \Big| \ \to \ 0 \quad \text{as} \quad n \to +\infty. \tag{3.14}$$

The above modulus is known as the star-discrepancy of ξ at order n and can be defined, for fixed $n \in \mathbb{N}$, for any n-tuple $(\xi_1, \dots, \xi_n)$ whose components ξ_k lie in $[0,1]^d$. There exists many sequences (Halton, Kakutani, Faure, Niederreiter, Sobol', see [6, 57] for example) achieving a $\mathcal{O}\Big(\frac{(\log(n+1))^d}{n}\Big)$ rate of decay for their star-discrepancy and it is a commonly shared conjecture that this rate is optimal, such sequences are called sequences with low discrepancy. By a standard so-called Hammersley argument, one shows that if a $[0,1]^{d-1}$-valued sequence $\zeta = (\zeta_n)_{n\geq 1}$ has low discrepancy i.e. there exists a real constant $C(\zeta) \in (0, +\infty)$ such that $D_n^*(\zeta) \leq C(\zeta)\frac{\big(\log(n+1)\big)^d}{n}$, for every $n \geq 1$, then, for every $n \geq 1$, the $[0,1]^d$-valued n-tuple $\Big((\zeta_k, \frac{k}{n})\Big)_{1\leq k\leq n}$ satisfies

$$D_n^*\left(\Big((\zeta_k, \tfrac{k}{n})\Big)_{1\leq k\leq n}\right) \leq C(\zeta)\frac{\big(\log(n+1)\big)^{d-1}}{n}.$$

The QMC method finds its gain in the following error bound for numerical integration. Let $(\xi_1, \dots, \xi_n)$ be a fixed n-tuple in $([0,1]^d)^n$, then, for every $f : [0,1]^d \to \mathbb{R}$ with finite variation (in the Hardy and Krause sense, see [53] or in the measure sense see [6, 57]),

$$\Big| \frac{1}{n}\sum_{i=1}^n f(\xi_k) - \int_{[0,1]^d} f(u)du \Big| \leq V(f) D_n^*(\xi_1, \dots, \xi_n). \tag{3.15}$$

where $V(f)$ denotes the (finite) variation of f. So a $\mathcal{O}\Big(\frac{(\log(n+1))^d}{n}\Big)$ or $\mathcal{O}\Big(\frac{(\log(n+1))^{d-1}}{n}\Big)$ rate of convergence can be achieved, for this class of functions, depending on the composition of the sequence. However, the class of functions with finite variation becomes sparser in the space of functions defined from $[0,1]^d$ to $\mathbb{R}$ and it seems natural to evaluate the performance of the low-discrepancy sequences or n-tuples on a more natural space of test functions like the Lipschitz functions. This is the purpose of Proïnov's Theorem reproduced below.

Theorem 3.4.1. *(Proinov, see [70]) Let $(\mathbb{R}^d, \|.\|_\infty)$. Let $\xi = (\xi_1, \dots, \xi_n)$ a sequence of $[0,1]^d$. For every continuous function $f : [0,1]^d \to (\mathbb{R}, |.|_\infty)$, we define the uniform continuity modulus of f by $w(f,\delta) = \sup_{\xi,\xi' \in [0,1]^d, |\xi-\xi'|_\infty \leq \delta} |f(\xi) - f(\xi')|$ where $|u|_\infty = \max_{1\leq i\leq d} |u_i|$ if $u = (u_1, \dots, u_d)$. Then, for every $n \geq 1$,*

$$\left| \frac{1}{n}\sum_{i=1}^n f(\xi_i) - \int_{[0,1]^d} f(x)dx \right| \leq C_d\, w(f, D_n^*(\xi)^{\frac{1}{d}}),$$

where C_d is a constant lower than 4 and depending only on the dimension d. In particular, if f is $[f]_{\mathrm{Lip}}-$Lipschitz and ξ has low discrepancy, one has

$$\left|\frac{1}{n}\sum_{i=1}^{n} f(\xi_i) - \int_{[0,1]^d} f(x)dx\right| \leq C_d\,[f]_{\mathrm{Lip}} D_n^*(\xi)^{\frac{1}{d}} \leq C_d\,[f]_{\mathrm{Lip}} \frac{\log(n+1)}{n^{\frac{1}{d}}}.$$

This suggests that, at least for a commonly encountered class of regular functions, the curse of dimensionality is more severe with QMC than with quantization due to the extra $(\log(n+1))^{1-\frac{1}{d}}$ factor in QMC. This is the price paid by QMC for considering uniform weights $p_i = \frac{1}{n}, i = 1, \ldots, n$.

With greedy quantization sequences, we will show that it is possible to keep the $n^{-\frac{1}{d}}$ rate of decay for numerical integration but also keep the asset of a sequence which is a recursive formula for cubatures.

3.4.1 Optimization of the algorithm and the numerical integration in the 1-dimensional case

Quadratic optimal greedy quantization sequences are obtained by implementing algorithms such as Lloyd's I algorithm, also known as k-means algorithm, or the Competitive Learning Vector Quantization (CLVQ) algorithm, which is a stochastic gradient descent algorithm associated to the distortion function. We refer to [46] (an extended version of [45] on ArXiv) where greedy variants of these procedures are explained in detail. According to Lloyd's algorithm, the construction of the sequences is recursive in the sense that, at the iteration n, we add one point a_n to $\{a_1, \ldots, a_{n-1}\}$, and we denote $\{a_1^{(n)}, \ldots, a_n^{(n)}\}$ an increasing reordering of $\{a_1, \ldots, a_n\}$ where the new added point is denoted by $a_{i_0}^{(n)}$. Since the other points are frozen, we can notice that the local inter-point inertia σ_i^2 defined by

$$\sigma_i^2 := \int_{a_i^{(n-1)}}^{a_{i+\frac{1}{2}}^{(n-1)}} |a_i^{(n-1)} - \xi|^2 P(d\xi) + \int_{a_{i+\frac{1}{2}}^{(n-1)}}^{a_{i+1}^{(n-1)}} |a_{i+1}^{(n)-1} - \xi|^2 P(d\xi), \quad i = 0, \ldots, n-1 \tag{3.16}$$

(where $a_{i+\frac{1}{2}}^{(n-1)} = \frac{a_i^{(n-1)} + a_{i+1}^{(n-1)}}{2}$ with $a_{\frac{1}{2}}^{(n-1)} = a_0^{(n-1)} = -\infty$ and $a_{n-\frac{1}{2}}^{(n-1)} a_n^{(n-1)} = +\infty$) remains untouched for every $i \in \{0, \ldots, n-1\}$ except $\sigma_{i_0}^2$ (the inertia between the point $a_{i_0}^{(n)}$ added at the n-th iteration and the following point) and $\sigma_{i_0-1}^2$ (the inertia between $a_{i_0}^{(n)}$ and the preceding point). Thus, at each iteration, the computation of n inertia can be reduced to the computation of only 2, thereby reducing the cost of the procedure. Likewise, the weights $p_i^n = P(W_i(a^{(n)}))$ of the Voronoï cells remain mostly unaffected. The only cells that change from one step to another are the cell $W_{i_0}(a^{(n)})$ having for centroid the new point $a_{i_0}^{(n)}$ and the two neighboring cells $W_{i_0-1}(a^{(n)})$ and $W_{i_0+1}(a^{(n)})$. Thus, the online computation of cell weights just needs 3 calculations instead of n (or 2 in case the added point is the first or last point in the reordered sequence). The utility of the weights of the Voronoï cells is featured in the approximation of $\mathbb{E}[f(X)]$ for $f : \mathbb{R}^d \to \mathbb{R}$ by the quadrature formula (3.13) using the reordered sequence $a^{(n)}$. Thus, based on the fact that only 3 Voronoï cells are modified at each iteration, one can deduce an iterative formula for the approximation of $I(f)$ by $I_n(f)$, requiring the storage of only 2

weights and 2 indices, as follows

$$\begin{aligned} I_n(f) &= I_{n-1}(f) - p_-^n f(a_{i_0-1}^{(n)}) - p_+^n f(a_{i_0+1}^{(n)}) + (p_+^n + p_-^n) f(a_{i_0}^{(n)}) \\ &= I_{n-1}(f) - p_-^n \Big(f(a_{i_0-1}^{(n)}) - f(a_{i_0}^{(n)}) \Big) - p_+^n \Big(f(a_{i_0+1}^{(n)}) - f(a_{i_0}^{(n)}) \Big), \end{aligned} \tag{3.17}$$

where
• $a_{i_0}^{(n)}$ is the point added to the greedy sequence at the n-th iteration, in other words, it is the point a_n,
• $a_{i_0-1}^{(n)}$ and $a_{i_0+1}^{(n)}$ are the points lower and greater than $a_{i_0}^{(n)}$, i.e. $a_{i_0-1}^{(n)} < a_{i_0}^{(n)} < a_{i_0+1}^{(n)}$,
•

$$p_-^n = P\Big([a_{i_0-\frac{1}{2}}^{(n)}, a_{\text{mil}}^{(n)}] \Big) \qquad \text{and} \qquad p_+^n = P\Big([a_{\text{mil}}^{(n)}, a_{i_0+\frac{1}{2}}^{(n)}] \Big). \tag{3.18}$$

where $a_{i_0\pm\frac{1}{2}}^{(n)} = \frac{a_{i_0}^{(n)} + a_{i_0\pm1}^{(n)}}{2}$ and $a_{\text{mil}}^{(n)} = \dfrac{a_{i_0+1}^{(n)} + a_{i_0-1}^{(n)}}{2}$, with $a_0 = -\infty$ and $a_n = +\infty$.
Practically, this numerical iterative method can be applied without storing the whole ordered greedy quantization sequence nor computing the weights of the Voronoï cells, which could appear as significant drawbacks for quantization. Instead, it requires the possession of 2 indices of 2 particular points of the non-ordered greedy quantization sequence and 2 weights. In fact, one can start by determining the indices of the points preceding and following a_n in the ordered sequence, in other words, the indices of the points in the non-ordered sequence corresponding to $a_{i_0-1}^{(n)}$ and $a_{i_0+1}^{(n)}$. Then, it becomes possible to compute the weights p_-^n et p_+^n.

3.4.2 Product greedy quantization ($d > 1$)

In higher dimensions, greedy quantization has always the recursive properties, so it gets interesting to apply the same numerical improvements as in the one-dimensional case. However, the construction of multidimensional greedy quantization sequences is complex and expensive since it relies on complicated stochastic optimization algorithms. As an alternative, one can use one-dimensional greedy quantization grids as tools to obtain multidimensional greedy quantization sequences in some cases.

How to build multi-dimensional greedy product grids

Multidimensional greedy quantization sequences can be obtained as a result of the tensor product of one-dimensional sequences, when the target law is a tensor product of its independent marginal laws. These grids are, of course, not optimal nor asymptotically optimal but they allow to approach the multidimensional law.
Let $X_1, \dots, X_d$ be d independent L^2-random variables taking values in $\mathbb{R}$ with respective distributions $\mu_1, \dots, \mu_d$ and $a^{1,(n_1)}, \dots, a^{d,(n_d)}$ the corresponding greedy quantization sequences. By computing the tensor product of these d one-dimensional greedy sequences, we obtain the d-dimensional greedy quantization grid $a^{1,(n_1)} \otimes \ldots \otimes a^{d,(n_d)}$ of the product law $\mu = \mu_1 \otimes \ldots \otimes \mu_d$, given by $(a_{\underline{j}}^{(n)})_{1 \le \underline{j} \le n} = (a_{j_1}^{1,(n_1)}, \dots, a_{j_d}^{d,(n_d)})_{1 \le j_1 \le n_1, \dots, 1 \le j_d \le n_d}$ of size $n = \prod_{i=1}^d n_i$. The corresponding quantization error is given by

$$e_r(a^{1,(n_1)} \otimes \ldots \otimes a^{d,(n_d)}, X_1 \otimes \ldots \otimes X_d)^r = \sum_{k=1}^d e_r(a^{k,(n_k)}, X_k)^r. \tag{3.19}$$

The weights $p_{\underline{j}}^{(n)}$ of the d-dimensional Voronoï cells $(W_{\underline{j}}(a^{(n)}))_{1\leq \underline{j}\leq n}$ are deduced from the one-dimensional Voronoï weights $(p_j^{k,n_k})_{1\leq j\leq n_d}$, $k=1,\dots,d$, corresponding to the one-dimensional greedy sequences, via

$$p_{\underline{j}} = p_{j_1}^{1,n_1}\times\ldots\times p_{j_d}^{d,n_d} \qquad \forall j_k\in\{1,\dots,n_k\}, \forall k\in\{1,\dots,d\}, \forall \underline{j}\in\{1,\dots,n\}.$$

The implementation of d-dimensional grids is not a point-by-point implementation. In fact, at each iteration n, having the d one-dimensional sequences, one must add a point to one one-dimensional sequence, generating this way several points of the multidimensional sequence. One must choose between d possibilities: add one point to only one sequence $a^{k,(n_k)}$ among the d marginal sequences to obtain $a^{(n_1\times\ldots\times n_{k-1}\times(n_k+1)\times n_{k+1}\times\ldots\times n_d)}$. These d cases are not similar since each one produces a different error quantization. So, the implementation is not a random procedure. To make the right decision, one must compute in each case, using (3.19), the quantization error E_k obtained if we add a point to $a^{k,(n_k)}$ for a $k\in\{1,\dots,d\}$. In other words, we compute, for $k=1,\dots,d$

$$E_k = e_r(a^{k,(n_k+1)},\mu_k)^r + \sum_{l\in\{1,\dots,d\}\setminus\{k\}} e_r(a^{l,(n_l)},\mu_l)^r.$$

Then, one chooses the index i such that $E_i=\min_{1\leq k\leq d}E_k$, adds a point to the sequence $a^{i,(n_i)}$ and obtains the grid $a^{(n_1\times\ldots\times n_{i-1}\times(n_i+1)\times n_{i+1}\times\ldots\times n_d)}$. We note that if the marginal laws $\mu_1,\dots,\mu_d$ are identical, this step is not necessary and the choice of the sequence to which a point is added, at each iteration, is systematically done in a periodic manner.

Numerical integration

Similarly to the 1-dimensional case, the majority of the Voronoï cells do not change while passing from an iteration n to an iteration $n+1$. At the n-th iteration, having $n_1\times\ldots\times n_d$ points in the sequence, one adds a new point to $a^{(i,n_i)}$. Hence, we will have $n_1\times\ldots\times n_{i-1}\times n_{i+1}\times\ldots\times n_d$ new created cells having for centroids the new points added to the d-dimensional sequence $a^{(n)}$, and another $2(n_1\times\ldots\times n_{i-1}\times n_{i+1}\times\ldots\times n_d)$ modified cells, corresponding to all the neighboring cells of the new added cells. In total, there is $3(n_1\times\ldots\times n_{i-1}\times n_{i+1}\times\ldots\times n_d)$ new Voronoï cells, while the rest remains unchanged. This leads to an iterative formula for quantization-based numerical integration (where the same principle as in the one dimensional case is applied) as follows: we denote, for the sake of simplicity $f_{i_0}=f(a_{j_1}^{1,(n_1)},\dots,a_{i_0}^{i,(n_i+1)},\dots,a_{j_d}^{d,(n_d)})$, $f_{i_0-1}=f(a_{j_1}^{1,(n_1)},\dots,a_{i_0-1}^{i,(n_i+1)},\dots,a_{j_d}^{d,(n_d)})$ and $f_{i_0+1}=f(a_{j_1}^{1,(n_1)},\dots,a_{i_0+1}^{i,(n_i+1)},\dots,a_{j_d}^{d,(n_d)})$

$$\begin{aligned} I_{n+1}(f) = I_n(f) &- p_-^{i,n_i+1}\sum_{\substack{j_k=1\\ k\in\{1,\dots,d\}\setminus\{i\}}}^{n_k}\prod_{\substack{k=1,\dots,d\\ k\neq i}} p_{j_k}^{k,(n_k)}(f_{i_0-1}-f_{i_0}) \\ &- p_+^{i,n_i+1}\sum_{\substack{j_k=1\\ k\in\{1,\dots,d\}\setminus\{i\}}}^{n_k}\prod_{\substack{k=1,\dots,d\\ k\neq i}} p_{j_k}^{k,(n_k)}(f_{i_0+1}-f_{i_0}) \end{aligned} \tag{3.20}$$

Note that in the d-dimensional case, the weights $p^{k,(n_k)}$, $k\in\{1,\dots,d\}\setminus\{i\}$ of the Voronoï cells of the other marginal sequences obtained at the previous iteration are needed, as well as the ordered one-dimensional greedy sequences $a^{k,(n_k)}$.

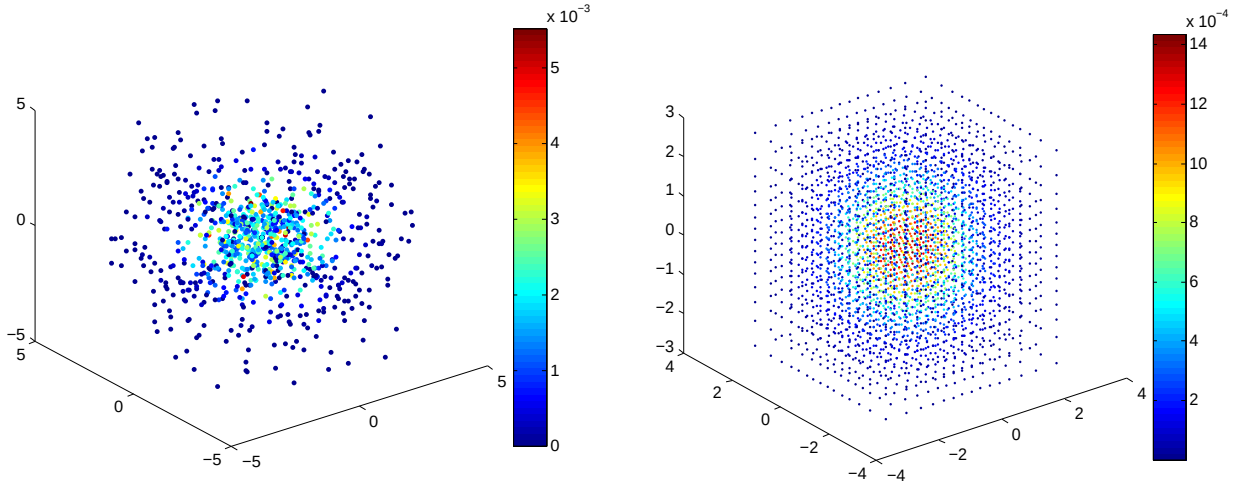

Figure 3.1: Greedy quantization sequences of the distribution $\mathcal{N}(0, I_3)$ of size $N = 15^3$ designed by Box Müller method (left) and greedy product quantization (right).

3.5 Numerical applications and examples

3.5.1 Greedy quantization of $\mathcal{N}(0, I_d)$ via Box-Müller

The Box-Müller method allows to generate a random vector with normal distribution $\mathcal{N}(0, I_2)$, actually two independent one-dimensional random variables Z_1 and Z_2 with distribution $\mathcal{N}(0, 1)$ by considering two independent random variables E and U with respective distributions $\mathcal{E}(1)$ and $\mathcal{U}([0, 1])$. Then, $2E \sim \mathcal{E}(\frac{1}{2})$ and $2\pi U \sim \mathcal{U}([0, 2\pi])$, so, the two variables

$$Z_1 = \sqrt{2E}\,\cos(2\pi U) \qquad \text{et} \qquad Z_2 = \sqrt{2E}\,\sin(2\pi U)$$

are independent and with normal distribution $\mathcal{N}(0, 1)$.
We use greedy quantization sequences $\varepsilon^{(n_1)}$ and $u^{(n_2)}$ of respective distributions $\mathcal{E}(1)$ and $\mathcal{U}[0, 1]$ to design two $\mathcal{N}(0, 1)$-distributed independent sequences $z_1^{(n)}$ et $z_2^{(n)}$, of size $n = n_1 \times n_2$, via the previous formulas so we can get a greedy sequence $z^{(n)}$ of the two-dimensional normal distribution $\mathcal{N}(0, I_2)$. The procedure is implemented as described in section 3.4.2. At each iteration, we must choose the one-dimensional distribution to which we should add a point. Thus, we compute the error induced if we add a point to $u^{(n_2)}$

$$E_u = e_2\big(u^{(n_2+1)}, \mathcal{U}[0, 2\pi]\big)^2 + e_2\big(\varepsilon^{(n_1)}, \mathcal{E}\big(\tfrac{1}{2}\big)\big)^2 = 4\pi^2 e_2\big(u^{(n_2+1)}, \mathcal{U}[0, 1]\big)^2 + 4e_2\big(\varepsilon^{(n_1)}, \mathcal{E}(1)\big)^2$$

and the error induces if we add a point to $\varepsilon^{(n_1)}$

$$E_\varepsilon = e_2\big(u^{(n_2)}, \mathcal{U}[0, 2\pi]\big)^2 + e_2\big(\varepsilon^{(n_1+1)}, \mathcal{E}\big(\tfrac{1}{2}\big)\big)^2 = 4\pi^2 e_2\big(u^{(n_2)}, \mathcal{U}[0, 1]\big)^2 + 4e_2\big(\varepsilon^{(n_1+1)}, \mathcal{E}(1)\big)^2$$

and we add a point to $u^{(n_2)}$ if $E_u < E_\varepsilon$ and a point to $\varepsilon^{(n_1)}$ if $E_\varepsilon < E_u$.

To design sequences in dimension $d > 2$, one uses several couples (E_i, U_i) to get several pairs (Z_i, Z_j) and uses the wanted number of $(Z_k)_k$ to obtain multidimensional sequences. In Figure

3.1, we compare two greedy quantization sequences of the distribution $\mathcal{N}(0, I_3)$ of size $N = 15^3$, one is obtained using the Box-Müller method based on two greedy exponential sequences $\mathcal{E}(1)$ and two greedy uniform sequences $\mathcal{U}([0,1])$, and the other obtained by greedy product quantization based on 3 one-dimensional Gaussian greedy sequences. The weights of the Voronoï cells in both cases are represented by a color scale (growing from blue to red). Note that, even if the greedy product quantization of a Normal distribution takes the shape of a cube (which is unusual for such distribution), the low values of the Voronoï weights at the edges of this cube allow to consider such a sequence as a valid approximation of the Gaussian distribution.

3.5.2 Pricing of a 3-dimensional basket of European call options

We consider a Call option on a basket of 3 positive risky assets, with strike price K and maturity T, with payoff $h_T = \left(\sum_{i=1}^3 w_i X_i - K\right)_+$ where (X_1, X_2, X_3) represent the prices of the 3 traded assets of the market and w_i are positive weights such that $\sum_{i=1}^3 w_i = 1$. We consider a 3-dimensional correlated Black-Scholes model where the prices of the assets are given,for every $i \in \{1,2,3\}$, by

$$X_i = X_{0,i} \exp\left((r - \tfrac{\sigma_i^2}{2})t + \sum_{j=1}^{q} \sigma_{ij} W_{j,t}\right), \quad t \in [0,T]$$

where r is the interest rate, σ_i the volatility of X_i and$(W_i)_i$ represent a correlated 3-dimensional Brownian motion, i.e. $(W_i, W_j) = \rho_{ij} t$ with $\rho_{1,1} = \rho_{1,2} = \rho_{1,3} = 0.5$ and all the others $\rho_{i,j}$'s are equal to 0. First, we compute $V_0 = e^{-rT}\mathbb{E}\left[h_T(X_1, X_2, X_3)\right]$ by a quadrature formula, according to (3.13), using a 3-dimensional greedy quantization sequences of $\mathcal{N}(0, I_3)$ obtained, on one hand, by the Box-Müller algorithm explained in the previous section and, on the other hand, by greedy product quantization of 3 one-dimensional sequences. Then, we estimate V_0 by the recursive formula (3.20) for $d = 3$ using the greedy product quantization sequence. We build sequences of size 32 000 and consider the following parameters

$$r = 0.1\ ,\ \sigma_i = 0.3\ ,\ X_{i,0} = 100\ ,\ T = 1 \text{ and } K = 100.$$

The reference price is given by a large Monte Carlo simulation with control variate of size $M = 2.10^7$. We consider the control variate

$$k_T = \left(e^{\sum_{i=1}^3 w_i \log(X_i)} - K\right)_+$$

which is positive and lower than h_T (owing to the convexity of the exponential). Since $e^{-rT}\mathbb{E}k_t$ has a normal distribution with mean $(r - \frac{1}{2}\sum_{i=1}^3 w_i\sigma_i^2)T$ and variance $w^t\sigma\sigma^t wT$, it admits a closed form given by

$$e^{-rT}\mathbb{E}k_t = \text{Call}_{\text{BS}}\left(\prod_{i=1}^3 X_{i,0}^{w_i} e^{-\frac{1}{2}T(\sum_{i=1}^3 w_i\sigma_i^2 - w^t\sigma\sigma^t w)}, K, r, \sqrt{w^t\sigma\sigma^t w}, T\right).$$

We compare the three methods in Table 3.1 where we expose the errors obtained by each method for particular number of points. We deduce that the recursive numerical integration gives the same results as the quadrature formula-based numerical integration making quantization-based numerical integration less expensive and more advantageous by reducing the cost in time and storage. Moreover, one notices that the Box-Müller algorithm is more accurate than the greedy

Table 3.1: Approximation of a 3-dimensional basket of call options in a BS model by Box-Müller with quadrature formula (BM), greedy product quantization with quadrature formula (GPQ) and with recursive formula (GPI).

n	BM	GPQ	GPI
100	1.72	1.68	1.84
1 000	0.07	0.42	0.42
8 000	0.04	0.08	0.08
15 000	0.07	0.08	0.08

product quantization, this can be explained by the fact that Box-Müller sequences fill the space in a way that resembles more to the normal distribution, we can notice a kind of ball different than the cube observed when implementing greedy product sequences (see Figure 3.1).

3.6 Further properties and numerical remarks

In this section, we present, based on numerical experiments, some properties of the one-dimensional quadratic greedy quantization sequences. We recall that $a^{(n)} = \{a_1^{(n)}, \dots, a_n^{(n)}\}$ denotes the re-ordered greedy sequence of the n first elements $\{a_1, \dots, a_n\}$ of $(a_n)_{n\geq 1}$.

3.6.1 Sub-optimality of greedy quantization sequences

The implementation of a greedy quantization sequence $(a_n)_{n\geq 1}$ of a distribution P and the computation of the corresponding weights p_i^n of the Voronoï cells $W_i(a^{(n)})$ for $i \in \{1, \dots, n\}$ defined by (3.1) is, in general, not optimal. However, numerical implementations and graphs representing $a_i \mapsto p_i^n = P(X \in W_i(a^{(n)}))$ for different number of points n show that, for certain distributions, the weights of the Voronoï cells converge towards the density curve of the corresponding distribution when the greedy sequence has a certain number of points.
For the normal distribution, this is observed when $n = 2^k - 1$, for every integer $k \geq 1$. So, we can say that the greedy quantization sequence is sub-optimal since the subsequence

$$\alpha^{(n)} = \alpha^{(2^k-1)} \text{ t.q. } \quad n = 2^k - 1,\ k \in \mathbb{N}^* \tag{3.21}$$

is itself optimal. Regarding the Uniform distribution on $[0, 1]$, we can check that there exist two sub-optimal sequences of the greedy sequence $\alpha^{(n)} = a^{(k_i)}$ for values of k_i defined by

$$\begin{cases} k_0 = 3, & \\ k_i = 2k_{i-1} + 1 & \text{if } i \equiv 1 \pmod 3, \\ k_i = 2(k_{i-1} - 2) + 1 & \text{if } i \equiv 2 \pmod 3, \\ k_i = 2(k_{i-1} + 2) + 1 & \text{if } i \equiv 0 \pmod 3. \end{cases} \qquad \begin{cases} k_0 = 11, & \\ k_i = 2k_{i-1} + 1 & \text{if } i \equiv 1 \pmod 3, \\ k_i = 2(k_{i-1} - 2) + 1 & \text{if } i \equiv 2 \pmod 3, \\ k_i = 2(k_{i-1} + 2) + 1 & \text{if } i \equiv 0 \pmod 3. \end{cases}$$

Some results for the normal distribution are represented in Figure 3.2 where we observe the unimodal weights for $n = 255 = 2^8 - 1$ and non-unimodal weights for $n = 400$. Similarly, the greedy quantization sequence of the Laplace distribution $L(0, 1)$ admits optimal subsequences of the form $a^{(2^k-1)}, k \in \mathbb{N}^*$. These observations allow to conjecture the sub-optimality of such subsequences for symmetrical distributions around 0.

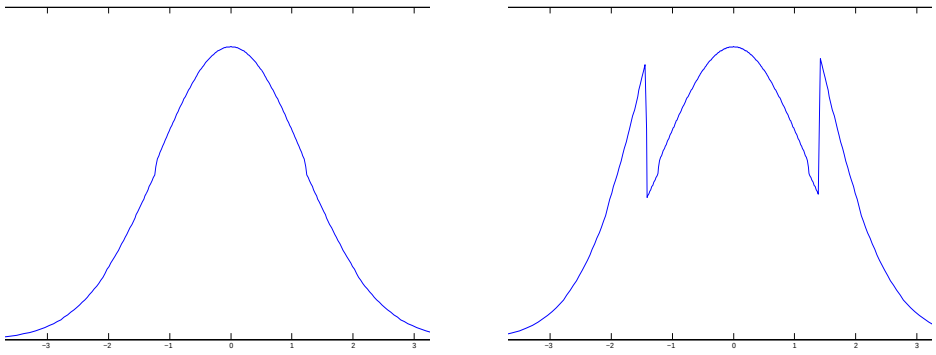

Figure 3.2: Representation of $a_i \mapsto p_i^n$ where $(p_i^n)_{1\leq i\leq n}$ denote the Voronoï weights of the greedy quantization sequence of $\mathcal{N}(0,1)$ for $n = 255$ (left), $n = 400$ (right).

3.6.2 Convergence of standard and weighted empirical measures

Sequences of asymptotically optimal n-quantizers $(\Gamma_n)_{n\geq 1}$ of P satisfy some empirical measure convergence theorems established in [32] (see Theorem 7.5 p. 96) and [19] and recalled below, where

$$\widehat{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i^n} \qquad \text{and} \qquad \widetilde{P}_n = \sum_{i=1}^{n} P(W_i(\Gamma_n))\delta_{x_i^n}$$

designate, respectively, the empirical standard measure and the empirical weighted measure associated to $\Gamma_n = \{x_1^n, \ldots, x_n^n\}$.

Theorem 3.6.1. *Assume P is absolutely continuous w.r.t the Lebesgue measure on $\mathbb{R}^d$ with density f. Let Γ_n be an asymptotically optimal n-quantizer of P. Then, denoting $C = \left(\int_{\mathbb{R}} f^{\frac{d}{d+p}}(u)du\right)^{-1}$, one has*

$$\widetilde{P}_n \underset{n\to+\infty}{\Rightarrow} P \qquad \text{and} \qquad \widehat{P}_n \underset{n\to+\infty}{\Rightarrow} C\, f^{\frac{d}{p+d}}(u)du. \tag{3.22}$$

Due to the existence of suboptimal greedy quantization sequences, detailed previously, we hope to obtain such results for greedy sequences or, at least, for sub-optimal greedy sequences defined in the previous section. To this end, we *"divide"* the two limits mentioned in (3.22), along the sequence $(W_i(a^{(n)}))_{1\leq i\leq n}$ and we obtain that, for every $i \in \{1, \ldots, n\}$, the limiting measure of the Voronoï cells of the greedy sequence is given by

$$P_l(W_i(a^{(n)})) \simeq \frac{f^{\frac{p}{d+p}}(a_i^{(n)})}{Cn} \tag{3.23}$$

In other words, if the greedy sequences satisfy the convergence of the empirical measures, then the weights of the Voronoï cells, computed using the c.d.f of P, must converge to the limit weights $P_l(W_i(a^{(N)}))$ given in (3.23).

Numerical experiments were established for the one-dimensional standard Normal, Uniform, Exponential and Laplace distribution. We observe that the exact weights of the Voronoï cells

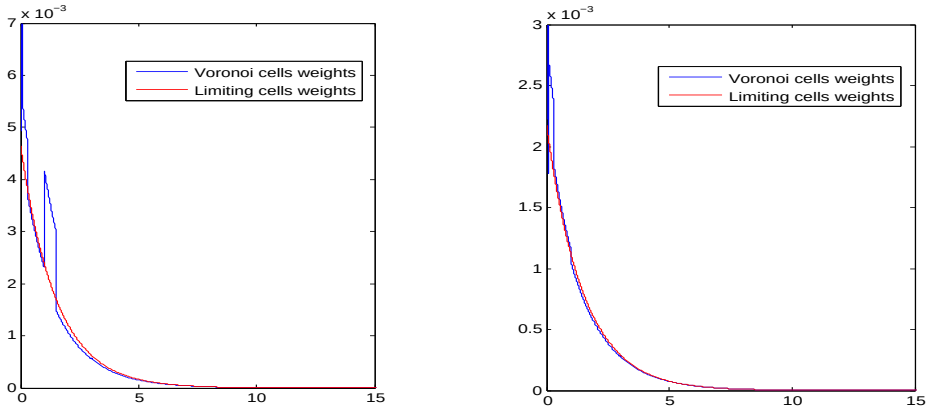

Figure 3.3: Comparison of the exact Voronoï weights (blue) and the limit weights (red) for the exponential distribution $\mathcal{E}(1)$ for $n = 645$ (left) and $n = 1\,379$ (right).

computed online get closer to the limit weights P_l when n increases. For the Gaussian distribution, we observe a more important convergence for the subsequences $a^{(2^k-1)}$ (as predicted). We present, in Figure 3.3 the obtained results for the exponential distribution where we compare the exact weights (blue) and the limit weights (3.23) (red) for different number of points n.

3.6.3 Stationarity and ρ-quasistationarity

An interesting question is to see if the greedy sequences are stationary i.e. satisfy

$$a_i^{(n)} = \mathbb{E}(X|X \in W_i(a^{(n)})), i = 1, \ldots, n,$$

or can be close to stationarity, a property shared by quadratic optimal quantizers. Numerical experiments conducted for several probability distributions yield that, unfortunately, greedy sequences are not stationary in this sense. In fact, one can prove that the greedy quantization sequence $a^{(n)}$ of a symmetric unimodal distribution is not stationary, expect for $n \in \{1; 3\}$. A proof is available in Chapter 4.

However, further different numerical observations show that most greedy quantization sequences satisfy a certain criteria that we call ρ-quasi-stationarity, approaching the stationary property and defined, for $r \in \{1, 2\}$ and $\rho \in [0, 1]$, by

$$\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_r = o(\|\widehat{X}^{a^{(n)}} - X\|_{1+\rho}^{1+\rho}), \quad \text{or} \quad \frac{\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_r}{\|\widehat{X}^{a^{(n)}} - X\|_{1+\rho}^{1+\rho}} \underset{n\to+\infty}{\longrightarrow} 0. \quad (3.24)$$

It is satisfied by greedy sequences for ρ lower than certain optimal values ρ_l depending on r and the distribution P. We expose, in Table 3.2, these values of ρ_l for $r \in \{1; 2\}$ for the Normal, Uniform and exponential distribution. This property is important because it brings improvements to quantization-based numerical integration. In fact, if f is $\mathcal{C}^1$ with ρ-hölder gradient with coefficient $[\nabla f]_\rho$, the classical error bound is given by (see [56])

$$|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})| \leq [\nabla f]_\rho \|X - \widehat{X}^{a^{(n)}}\|_{1+\rho}^{1+\rho}.$$

Table 3.2: Values of optimal ρ_l for different distributions and for $r \in \{1;2\}$.

	$\mathcal{N}(0,1)$	$\mathcal{U}([0,1])$	$\mathcal{E}(1)$
$r=1$	$\rho_l = 0.92$	$\rho_l = \frac{3}{4}$	$\rho_l = \frac{2}{3}$
$r=2$	$\rho_l = 0.47$	$\rho_l = \frac{3}{8}$	$\rho_l = \frac{1}{3}$

However, one has

$$\begin{aligned}\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}}) \leq & \mathbb{E}\big(\nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}\big) \\ & + \mathbb{E}\left[\int_0^1 \Big(\nabla f\Big(\widehat{X}^{a^{(n)}} + t(X - \widehat{X}^{a^{(n)}})\Big) - \nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}\Big)dt\right]\end{aligned}$$

where the second expectation in the right term of the above inequality is bounded by $[\nabla f]_\rho \mathbb{E}|X - \widehat{X}^{a^{(n)}}|^{1+\rho} \int_0^1 t^{1+\rho}dt$ and

$$\begin{aligned}\mathbb{E}(\nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}) &= \mathbb{E}(\nabla f(\widehat{X}^{a^{(n)}})|X) - \mathbb{E}(\nabla f(\widehat{X}^{a^{(n)}})|\widehat{X}^{a^{(n)}}) \\ &= \mathbb{E}\left(\nabla f(\widehat{X}^{a^{(n)}})|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\right).\end{aligned}$$

So, if (3.24) is satisfied, then one obtains

$$\begin{aligned}|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})| \leq & \|\nabla f(\widehat{X}^{a^{(n)}})\|_2 \|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\|_2 + \tfrac{1}{1+\rho}[\nabla f]_\rho \|X - \widehat{X}^{a^{(n)}}\|_{1+\rho}^{1+\rho} \\ \leq & \frac{1}{1+\rho}[\nabla f]_\rho \|X - \widehat{X}^{a^{(n)}}\|_{1+\rho}^{1+\rho}.\end{aligned}$$

3.6.4 Discrepancy of greedy sequences

The comparison established, in the beginning of section 3.4, between greedy quantization-based numerical integration and quasi-Monte Carlo methods, showing a gain of $\log(n)$-factor with greedy quantization in terms of convergence rate, drives us to build a relation, based on Proïnov's Theorem 3.4.1, between the error quantization and the discrepancy. In fact, for every n-tuple $\Xi = (\xi_1, \ldots, \xi_n) \in [0,1]^n$, noticing that a Lipschitz function f has always a finite variation and considering the function $f : u \to \min_{1\leq i\leq n} |u - \xi_i|$ which is 1-Lipschitz (since $|\min_i a_i - \min_i b_i| \leq \max_i |a_i - b_i|$) and satisfies $f(\xi_i) = 0$ for every $i \in \{1, \ldots, n\}$ and $\displaystyle\int_0^1 f(u)du = e_1(X, \mathcal{U}([0,1]))$, one applies the Koksma-Hlawka inequality (3.15) to f to deduce that

$$e_1(\Xi, \mathcal{U}([0,1])) \leq D_n^*(\Xi). \tag{3.25}$$

This motivates us to study the discrepancy of greedy sequences hoping that they can be comparable to low discrepancy sequences. We compute this quantity for $d \in \{1,2,3\}$, using formulas given in [20] and deduce that, in the one-dimensional case, greedy sequences can be used as a low discrepancy sequence. But, when d becomes larger than 1, the situation becomes less convincing: The discrepancy of pure greedy sequences, designed by implementing Lloyd's algorithm, and that of low discrepancy sequences (Niederreiter sequences for example) are comparable, but the problem that arises is the complexity of the computations making greedy sequences less practical. On the other hand, if we build greedy product sequences, the computations will be

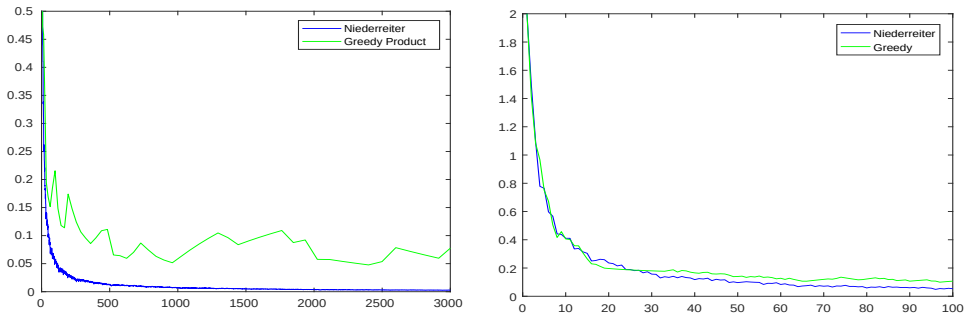

Figure 3.4: Comparisons of the star discrepancy of the Niederreiter sequence to a greedy product quantization sequence of the Uniform distribution $\mathcal{U}([0,1]^2)$ (left) and to a pure greedy quantization sequence (right) for $d = 2$.

less expensive but there is no gain in terms of discrepancy. Figure 3.4 shows a comparison of the discrepancy of a Niederreiter sequence in dimension 2 to that of a product greedy quantization sequence of $\mathcal{U}([0,1]^2)$ on the one hand, and to that of pure greedy quantization sequence of $\mathcal{U}([0,1]^2)$ on the other hand.

The positive results obtained for $d = 1$ encourage us to try and manipulate low discrepancy sequences, such as Van der Corput sequences, in order to be able to use them as greedy quantization sequences. In other words, we will assign to them a Voronoï diagram, compute the weights of the corresponding Voronoï cells instead of considering uniform weights and observe the impact this brings to numerical integration. To this end, we consider a basic example where we compute the price of a European call $C_0 = \mathbb{E}[(X_T - K)_+]$ for a maturity T and a strike price K where the price of the asset X_t at a time t is given by

$$X_t = x_0 \exp\left((r - \tfrac{\sigma^2}{2})t + \sigma\sqrt{t}Z_t\right)$$

where r is the interest rate, σ the volatility and $(Z_t)_{0\leq t\leq T}$ is an i.i.d. sequence of random variables with distribution $\mathcal{N}(0,1)$. We consider

$$T = 1,\ K = 9,\ x_0 = 10,\ \mu = 0.06,\ \sigma = 0.1.$$

The exact price is given by a closed formula known in the Black-Scholes case and is approximately equal to 1.5429. We compute the price C_0 first via a classical quadrature formula using the new weights p_i^n assigned to the VdC sequence, then by a classical quasi-Monte Carlo simulation (using uniform weights of a VdC sequence) and finally by a quantization-based quadrature formula based on a greedy quantization sequence of $\mathcal{U}([0,1])$. We compare the errors induced by these three methods in Figure 3.5 and deduce that the procedure using the greedy quantization sequence converges faster than the ones using the Van der Corput sequence. Consequently, one can say that greedy sequences are more advantageous than low discrepancy sequences, even if we assign to them non-uniform weights.

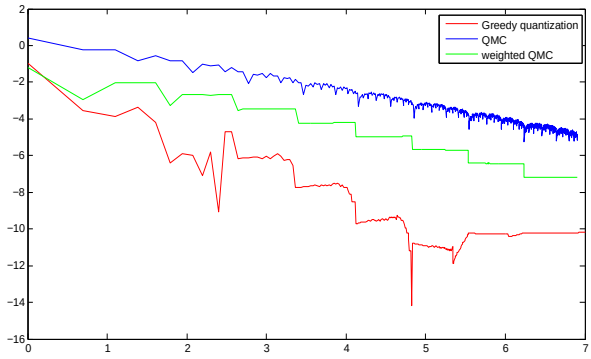

Figure 3.5: Price of a European call in a Black-Scholes model via a usual QMC method (blue), greedy quantization-based quadrature formula (red) and quadrature formula using VdC sequence with non-uniform weights (logarithmic scale).

Chapter 4

Greedy vector quantization: Detailed numerical studies

Let X be a random variable with probability P defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. In Chapter 3, we studied theoretical aspects of greedy vector quantization first developed extensively in [45] and consisting in building a *sequence* of points $(a_n)_{n\geq 1}$ in $\mathbb{R}^d$ which is recursively optimal step by step, in the sense that it minimizes the L^r-quantization error at each iteration. In other words, at each iteration of the implementation, one adds a point a_{n+1} solution to

$$a_{n+1} \in \operatorname{argmin}_{\xi\in\mathbb{R}^d} e_r(a^{(n)} \cup \{\xi\}, X). \tag{4.1}$$

We also presented a new numerical approach and established a new iterative formula for quantization-based numerical integration, based on the fact that, at each iteration, while adding a new point to the greedy sequence, most of the Voronoï cells remain untouched.
Furthermore, to study to what extent greedy sequences can be close to optimality, we exposed some numerical experiments that led us to extend some properties to greedy sequences. Our interest in this chapter will be to develop what was briefly presented in the first chapter. We will explain the algorithms that allow to obtain greedy quantization sequences and give further details and new experimental results related to the properties already deduced in Chapter 3.

Throughout this chapter, we consider a random variable X with distribution P and $a^{(n)}$ a corresponding greedy quantization sequence and we assume that $\mathbb{R}^d$ is equipped with the canonical Euclidean norm and that $p = 2$ (except when mentioned otherwise).

4.1 Algorithms of computation of greedy sequences

Practical computation of an optimal greedy quantization sequence relies on variants of usual algorithms such as CLVQ and Lloyd's algorithm used for building sequences of optimal quantizers. For greedy quantization, the implementation is recursive, in the sense that, in order to switch from the $(n-1)$-th to the n-th iteration, one adds, in a way to be specified, a n-th point to the existing $(n-1)$-tuple $(a_1, \ldots, a_{n-1})$ computed during the first $n-1$ iterations of the algorithm. This way, one possesses the starting n-tuple for the modified CLVQ or Lloyd procedure. One implements these optimization procedures keeping in mind that all formerly computed components $(a_i)_{1\leq i\leq n-1}$ are kept frozen, and only the new added point is moved following the standard rules

of the algorithm. In other words, the new point added at each iteration of the greedy procedure is the only centroïd updated during the algorithms. Note that the first point of the sequence is the L^p-median of the distribution P, i.e. $a_1 = \mathbb{E}[X]$.

4.1.1 One-dimensional case

In the one-dimensional case, when the distribution is absolutely continuous with a continuous positive probability density φ, deterministic Lloyd and CLVQ algorithms can easily be extended to greedy versions. We present the details in the following.

Greedy Lloyd's algorithm As already mentioned, the computation of greedy sequences is recursive. So, we assume that the first $n-1$ points $a_1, \ldots, a_{n-1}$ have been computed and we compute the n^{th} point a_n according to the following steps.
• Sort the first $n-1$ points of the sequence $a_1, \ldots, a_{n-1}$ in an increasing order:

$$a_1^{(n-1)} < \cdots < a_{n-1}^{(n-1)}.$$

• Compute the n inter-point local inertia given by

$$\sigma_i^2 := \int_{a_i^{(n-1)}}^{a_{i+\frac{1}{2}}^{(n-1)}} |a_i^{(n-1)} - \xi|^2 P(d\xi) + \int_{a_{i+\frac{1}{2}}^{(n-1)}}^{a_{i+1}^{(n-1)}} |a_{i+1}^{(n-1)} - \xi|^2 P(d\xi), \quad i = 0, \ldots, n-1 \tag{4.2}$$

where $a_0^{(n-1)} = -\infty$, $a_n^{(n-1)} = +\infty$ and $a_{i+\frac{1}{2}}^{(n-1)}$ is the mid-point of $[a_i^{(n-1)}, a_{i+1}^{(n-1)}]$:

$$a_{\frac{1}{2}}^{(n-1)} = -\infty, \quad a_{i+\frac{1}{2}}^{(n-1)} = \frac{a_i^{(n-1)} + a_{i+1}^{(n-1)}}{2}, \quad a_{n-\frac{1}{2}}^{(n-1)} = +\infty.$$

• Determine the index $i_0 = i_0(n-1)$ corresponding to the maximal local inertia, i.e. such that $\sigma_{i_0}^2 = \max_{0 \le i \le n-1} \sigma_i^2$, and choose a random point $\bar{a}_0 \in (a_{i_0}^{(n-1)}, a_{i_0+1}^{(n-1)})$.
• Define a recursive sequence $a_{[l]} = a_{n,[l]}$ starting at $a_{n,0} = \bar{a}_0$ by

$$a_{[l+1]} = \mathbb{E}(X | X \in W_{n,l}) = \frac{K_X\left(\frac{a_{i_0+1}^{(n-1)} + a_{[l]}}{2}\right) - K_X\left(\frac{a_{i_0}^{(n-1)} + a_{[l]}}{2}\right)}{F_X\left(\frac{a_{i_0+1}^{(n-1)} + a_{[l]}}{2}\right) - F_X\left(\frac{a_{i_0}^{(n-1)} + a_{[l]}}{2}\right)}, \quad l \ge 0, \tag{4.3}$$

where $F_X(x) = P((-\infty, x])$ is the cumulative distribution function of the distribution P of X, K_X is its cumulative first moment function defined by $K_X(x) = \int_{(-\infty,x]} \xi dP(\xi), \quad x \in \mathbb{R}$ and $W_{n,l}$ is the Voronoï cell of centroid $a_{n,[l]}$ corresponding to the sequence $a^{(n-1)} \cup \{a_{n,[l]}\}$. One can easily check that, at each iteration, $a_{n,[l]} \in (a_{i_0}^{(n-1)}, a_{i_0+1}^{(n-1)})$ which makes the procedure well defined.

Relying on classical arguments called upon in the proofs of the convergence of the standard Lloyd I procedure (see [11, 39]), one can show that if $P = \varphi.\lambda_1$ where $\varphi : \mathbb{R} \to \mathbb{R}$ is a log-concave function, the sequence $a_{n,[l]}$ converges towards the unique solution $a_{n,\infty} \in (a_{i_0}^{(n-1)}, a_{i_0+1}^{(n-1)})$ of the fixed-point equation

$$a_n = \mathbb{E}\left(X | X \in W_n\right)$$

where W_n is the closed Voronoï cell of centroid a_n in the Voronoï diagram corresponding to $a^{(n-1)} \cup \{a_n\}$.

Remark 4.1.1. • *The integrals involved in the algorithm can be computed using higher order quadrature formulas, or, for example for the standard Normal distribution, using the closed form for* $\int_{-\infty}^{x} \xi e^{-\frac{\xi^2}{2}} \frac{d\xi}{\sqrt{2\pi}} = -\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$.
• *The* log*-concave assumption, which implies the uniqueness of the fixed point for equation* (4.3)*, is satisfied by many families of distributions like the Gaussian distributions* $\mathcal{N}(m, \sigma^2)$*, the exponential and Laplace distributions, the* $\gamma(\alpha, \beta)$*-distributions,* $\alpha \geq 1$*,* $\beta > 0$ *which are strongly unimodal.*

Greedy CLVQ algorithm Assume $P = \varphi.\lambda_d$. This is a gradient descent algorithm, also known as k-means algorithm, defined by the following recursion

$$a_{[l+1]} = a_{[l]} - \Big(\gamma_{l+1} \wedge \frac{1}{\rho(a_{[l]})}\Big) \int_{\frac{a_{i_0}^{(n-1)} + a_{[l]}}{2}}^{\frac{a_{i_0+1}^{(n-1)} + a_{[l]}}{2}} (a_{[l]} - \xi) P(d\xi) \tag{4.4}$$

where γ_{l+1} is a $(0,1)$-valued sequence that goes to 0 when l goes to $+\infty$ and such that $\sum_l \gamma_l = +\infty$, and $\rho(a) > 0$ is the second derivative of the function $a \mapsto \mathbb{E}\big(\min |X - a_i|^2 \wedge |X - a|^2\big)$ given by

$$\rho(a) = P\Big(\Big[\frac{a_{i_0}^{(n-1)} + a}{2}, \frac{a_{i_0+1}^{(n-1)} + a}{2}\Big]\Big) + \frac{a - a_{i_0}^{(n-1)}}{2} \varphi\Big(\frac{a + a_{i_0}^{(n-1)}}{2}\Big) + \frac{a_{i_0+1}^{(n-1)} - a}{2} \varphi\Big(\frac{a + a_{i_0+1}^{(n-1)}}{2}\Big).$$

This recursion is well defined and consistent since it lives in the interval $(a_{i_0}^{(n-1)}, a_{i_0+1}^{(n-1)})$, this is due to the fact that $\gamma_{l+1} \in (0,1)$. Similarly to the Lloyd's algorithm, the computation of integrals involved can be performed by higher order quadrature formulas and closed forms of certain integrals in some cases.
In case P is not absolutely continuous, one has only to replace the term involving the second derivative with a step γ_l satisfying the standard *decreasing step assumption* that is $\sum_l \gamma_l = +\infty$ and $\sum_l \gamma_l^2 < +\infty$, provided one can compute the P-integrals of interest.

Greedy quantization of the $\mathcal{N}(0,1)$ distribution The symmetry of the one-dimensional standard Gaussian distribution allows to simplify, in a certain way, the computation of its quadratic optimal greedy sequence. In other terms, we consider the distribution $\tilde{P} = P_{|\mathbb{R}_+}$ (which is clearly strongly unimodal) and we compute its quadratic optimal greedy sequence $(\widetilde{a}_n)_{n \geq 1}$ by the greedy Lloyd's procedure starting at $\widetilde{a}_1 = 0$. Consequently, the sequence given by

$$a_0 = 0,\ a_{2n-1} = \widetilde{a}_n,\ a_{2n} = -\widetilde{a}_n,\ n \geq 1,$$

is a quadratic optimal greedy sequence for the Gaussian distribution.

In order to proceed with the computation of the greedy sequence, note that the integrals involved in the algorithms explained above are computed using the following functions

$$F_X(x) = P((-\infty, x])\ , \quad K_X(x) = \int_{(-\infty, x]} \xi dP(\xi) = -\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

and the cumulative second moment function

$$L_X(x) = \int_{]-\infty,x]} \xi^2 P(d\xi) = F_X(x) + xK_X(x).$$

At the n-th iteration of the algorithm, the inter-point inertia are given as follows, where we replace $a_i^{(n-1)}$ by a_i for simplicity,

$$\begin{aligned}
\sigma_0^2 &= (a_1^2+1)F_X(a_1) + a_1\frac{e^{-\frac{a_1^2}{2}}}{\sqrt{2\pi}},\\
\sigma_i^2 &= F_X(a_{i+1})(1+a_{i+1}^2) + F_X(a_{i+\frac{1}{2}})(a_i^2 - a_{i+1}^2) - F_X(a_i)(1+a_i^2)\\
&\quad + a_{i+1}\frac{e^{-\frac{a_{i+1}^2}{2}}}{\sqrt{2\pi}} - a_i\frac{e^{-\frac{a_i^2}{2}}}{\sqrt{2\pi}} + 2(a_i - a_{i+1})\frac{e^{-\frac{a_{i+\frac{1}{2}}^2}{2}}}{\sqrt{2\pi}}, \quad 1 \le i \le n-2,\\
\sigma_{n-1}^2 &= (1-F_X(a_{n-1}))(1+a_{n-1}^2) - a_{n-1}\frac{e^{-\frac{a_{n-1}^2}{2}}}{\sqrt{2\pi}}.
\end{aligned}$$

The Voronoï diagram corresponding to this sequence is given by $W_i(a^{(n)}) = (a^{(n)}_{i-\frac{1}{2}}, a^{(n)}_{i+\frac{1}{2}})$, $i \in \{1,\dots,n\}$ and the corresponding Voronoï weights are given by

$$p_i^n = P(X \in W_i(a^{(n)})) = F_X(a^{(n)}_{i+\frac{1}{2}}) - F_X(a^{(n)}_{i-\frac{1}{2}}).$$

We study the convergence of the quadratic quantization error $e_2(a^{(n)}, X)$ induced by the greedy quantization of the distribution $\mathcal{N}(0,1)$. This error is given by

$$e_2(a^{(n)}, X)^2 = \int_{\mathbb{R}^d} \min_{1\le i\le n} |a_i^{(n)} - \xi|^2 dP(\xi) = \sum_{i=1}^n \int_{W_i(a^{(n)})} |a_i^{(n)} - \xi|^2 dP(\xi) = \sum_{i=1}^n \sigma_i^2,$$

where $\sigma_i^2, i = 0,\dots,n-1$ are the inter-point local inertia already computed. We reproduce in Figure 4.1 the graph representing $n \mapsto n\, e_2(a^{(n)}, P)$, $n = 4,\dots,20\,000$, for $P = \mathcal{N}(0,1)$ and observe that

$$\liminf_n n e_2(a^{(n)}, \mathcal{N}(0,1)) \approx 1.6534\cdots > \sqrt{\frac{3}{2}}\,\pi^{\frac{1}{4}} = \lim_n n e_{2,n}(\mathcal{N}(0,1))$$

and that

$$\limsup_n n e_2(a^{(n)}, \mathcal{N}(0,1)) \approx 1.8921 < 2 \times \sqrt{\frac{3}{2}}\,\pi^{\frac{1}{4}} \approx 3.2611.$$

The real constant in the right hand side of the inequality easily follows from Zador's Theorem.

Here, we can highlight the fact that the quantization error attains its lowest values when $n = 2^{k-1}, k \ge 1$. This can be explained by the existence, emphasized by numerical experiments in Section 3.6.1 of Chapter 3, of sub-optimal sequences of the greedy quantization grids of the standard Gaussian distribution. In fact, the graphs representing the weights of the Voronoï cells of a greedy quantization sequence $a^{(n)}$ of $\mathcal{N}(0,1)$ appeared to be uni-modal when the number of points is $n = 2^{k-1}, k \ge 1$, which led us to conjecture that the sequences $a^{(2k-1)}$ are sub-optimal and thus, produce the lowest values of the quantization error.

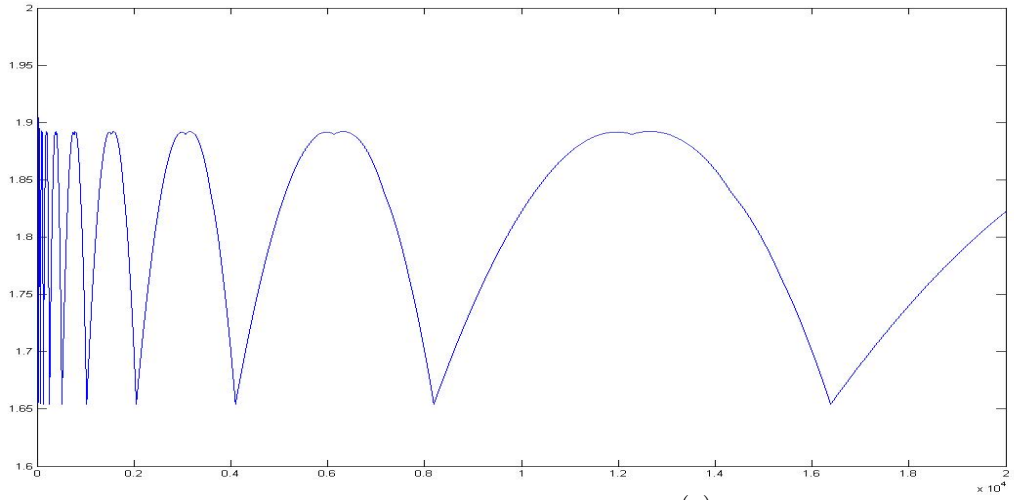

Figure 4.1: Quadratic greedy quantization error $n \longmapsto ne_2(a^{(n)}, X)$ associated to the Gaussian distribution $\mathcal{N}(0,1)$ for $n = 2, \ldots, 20\,000$.

Greedy quantization of the $\mathcal{U}([0,1])$ distribution The computation of a greedy quantization sequence of the Uniform distribution $\mathcal{U}([0,1])$ is implemented via the greedy Lloyd's algorithm starting at $a_1 = \frac{1}{2}$. Then, considering that, at the n^{th}-iteration, the $n-1$ first points $a_1^{(n-1)}, \ldots, a_{n-1}^{(n-1)}$ are already computed and reordered increasingly, we compute the $n-1$ inter-point inertia given by the following (where we replace $a_i^{(n-1)}$ by a_i for every $i \in \{1, \ldots, +\infty\}$ for simplicity)

$$\begin{aligned}
\sigma_0^2 &= \frac{a_1^3}{3}, \\
\sigma_i^2 &= \frac{(a_{i+1} - a_{i+\frac{1}{2}})^3}{3} - \frac{(a_i - a_{i+\frac{1}{2}})^3}{3}, \quad i = 1, \ldots, n-2, \\
\sigma_{n-1}^2 &= \frac{(1 - a_{n-1})^3}{3}.
\end{aligned}$$

Then, the algorithm is implemented as described previously, having in mind that we consider $a_0 = -\infty$ and $a_{n+1} = +\infty$ even if the support of P is $[0,1]$. The Voronoï cells are given by

$$W_1(a^{(n)}) = (0, a_{1+\frac{1}{2}}), \quad W_n(a^{(n)}) = (a_{n-\frac{1}{2}}, 1) \quad \text{and} \quad W_i(a^{(n)}) = (a_{i-\frac{1}{2}}, a_{i+\frac{1}{2}}), \quad i = 2, \ldots, n-1.$$

The weights of the Voronoï cells are computed easily using the cumulative distribution function F_X. Figure 4.2 represents the graph of the quadratic quantization error $n \longmapsto ne_2(a^{(n)}, X)$ where we observe that

$$\liminf_n ne_2(a^{(n)}, X) \approx 0.295 > \lim_n ne_{2,n}(X),$$

and

$$\limsup_n ne_2(a^{(n)}, X) \approx 0.32 < 2 \lim_n ne_{2,n}(X).$$

Just as in the Gaussian distribution case, we notice that the quantization error induced by the approximation of the Uniform distribution attains its lowest values for two sub-sequences. As

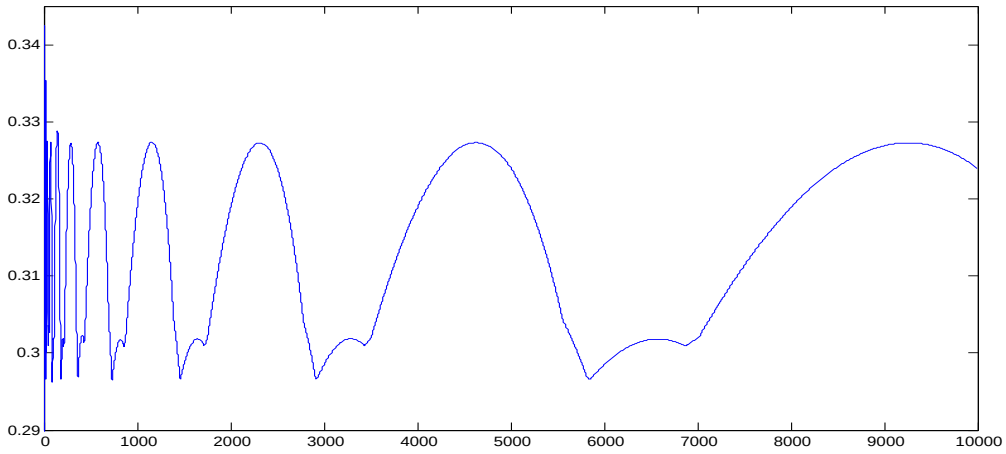

Figure 4.2: Quadratic greedy quantization error corresponding to the Uniform distribution $\mathcal{U}([0,1])$ for $n = 1, \ldots, 10\,000$.

mentioned in Section 3.6.1 of Chapter 3, there exists two optimal sub-sequences for which the weights of the Voronoï cells approach well the distribution curve and thus allowing to obtain the lowest values of the quantization error.

Greedy quantization of the $\mathcal{E}(1)$ distribution Let X be a random variable with exponential distribution $P = \mathcal{E}(1)$. The same procedure as previously is adopted to compute a greedy quantization sequence $a^{(n)}$ of X. The integrals can be computed relying on the functions $F_X(x) = P(]-\infty, x])$ and $K_X(x) = 1 - e^{-x} - xe^{-x}$. We start the algorithm at $a_1 = \mathbb{E}(X) = 1$ and, at the n-th iteration, we compute the n inter-point inertia as follows (where we replace $a_i^{(n-1)}$ by a_i for every $i \in \{1, \ldots, +\infty\}$ for simplicity)

$$\sigma_0^2 = 2 - 2a_1 - 2e^{-a_1} + a_1^2,$$
$$\sigma_i^2 = e^{-a_{i-\frac{1}{2}}}(a_i - a_{i+\frac{1}{2}})(2 - a_i + a_{i+\frac{1}{2}}) + 2(e^{-a_i} - e^{-a_{i+1}}) - e^{-a_{i+\frac{1}{2}}}(a_{i+1} - a_{i+\frac{1}{2}})(2 - a_{i+1} - a_{i+\frac{1}{2}}),$$
$$1 \leq i \leq n-1$$
$$\sigma_{n-1}^2 = 2e^{-a_{n-1}}.$$

We observe, in Figure 4.3, the quadratic error $n \mapsto ne_2(a^{(n)}, X)$ induced by the greedy quantization of the exponential distribution $\mathcal{E}(1)$ for $n = 4, \ldots, 10\,000$ points.

Greedy quantization of the Laplace distribution with parameters $\alpha = 0$ and $\beta = 1$
Let X be a random variable with a Laplace distribution with parameters 0 and 1. The computation of a greedy quantization sequence $a^{(n)}$ of X is implemented via Lloyd's algorithm starting at $a_1 = \mathbb{E}[X] = 0$ and in which the local inter-point inertia at the n-th iteration are

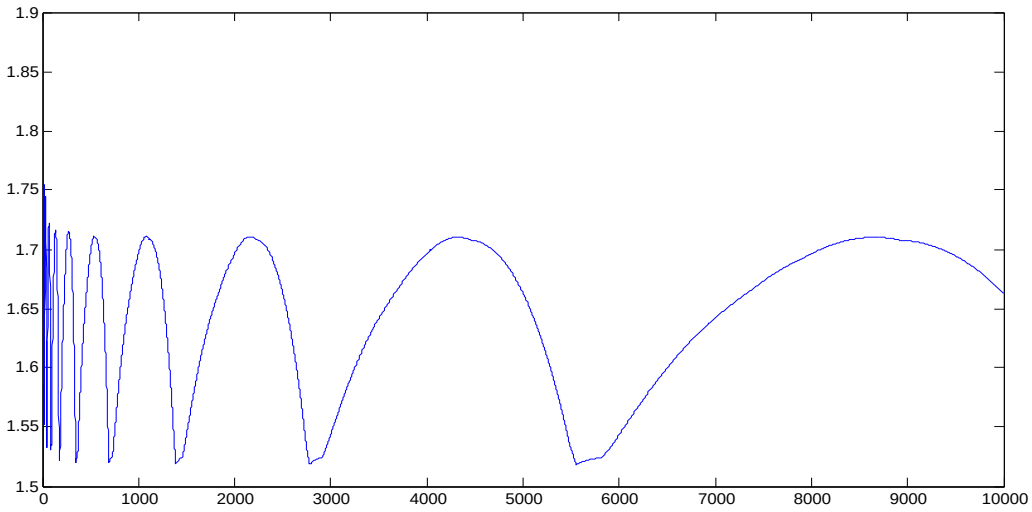

Figure 4.3: Quadratic greedy quantization error $n \mapsto ne_2(a^{(n)}, X)$ of the exponential distribution $\mathcal{E}(1)$ for $n = 4, \ldots, 10\,000$ points.

given by

$$\sigma_0^2 = \frac{e^{a_1}}{2},$$

$$\sigma_i^2 = \begin{cases} e^{-a_{i+\frac{1}{2}}}(a_i - a_{i+\frac{1}{2}})(1 - \frac{1}{2}(a_i - a_{i+\frac{1}{2}})) + e^{-a_{i+\frac{1}{2}}}(a_{i+1} - a_{i+\frac{1}{2}})\Big(\frac{1}{2}(a_{i+1} - a_{i+\frac{1}{2}}) - 1\Big) \\ +e^{-a_i} - e^{-a_{i+1}} \qquad \text{if } a_{i+1} < 0 \\ e^{a_{i+\frac{1}{2}}}(a_i - a_{i+\frac{1}{2}})(1 + \frac{1}{2}(a_i - a_{i+\frac{1}{2}})) - e^{a_{i+\frac{1}{2}}}(a_{i+1} - a_{i+\frac{1}{2}})\Big(\frac{1}{2}(a_{i+1} - a_{i+\frac{1}{2}}) + 1\Big) \\ +e^{a_{i+1}} - e^{a_i} \qquad \text{if } a_{i+1} > 0, \end{cases}$$

$$i = 1, \ldots, n-2,$$

$$\sigma_{n-1}^2 = e^{-a_{n-1}}.$$

The quantization error thus obtained is illustrated in Figure 4.4 where we represent $n \mapsto ne_2(a^{(n)}, X)$ for $n = 1, \ldots, 10\,000$.

4.1.2 Multi-dimensional case

When the dimension becomes higher ($d \geq 2$), deterministic greedy Lloyd's and greedy CLVQ algorithms become too demanding due to several computations of integrals over the Voronoï cells of the quantization sequence. So, it becomes necessary to switch to stochastic optimization procedures which are adaptations of the stochastic procedures introduced to compute optimal n-quantizers. The convergence results of these procedures remain partial, especially if the distribution P is not compactly supported. For more details about these original stochastic optimization procedures, mostly devised in the 1950's, we refer *e.g.* to [8, 62] for *CLVQ* and [21, 39, 68] for (randomized) Lloyd's I procedure or more applied textbooks like [22]. In practice, the computation of integrals on the Voronoï cells is replaced, in both procedures, by

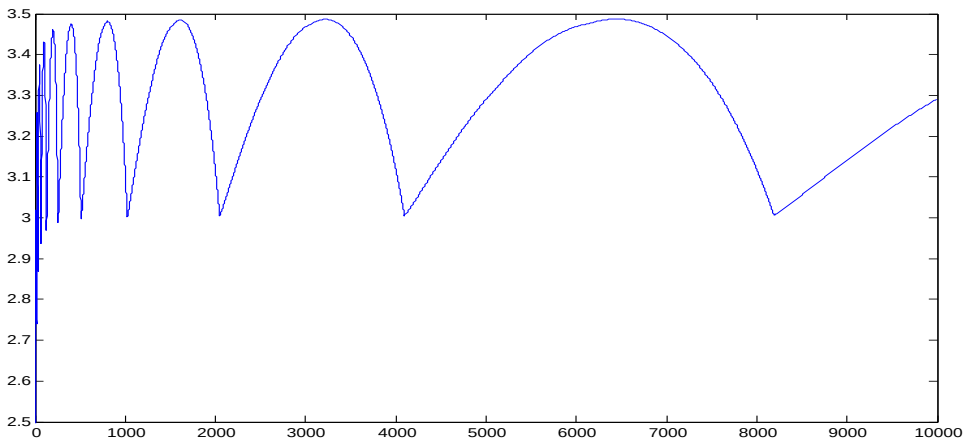

Figure 4.4: Quadratic greedy quantization error $n \mapsto ne_2(a^{(n)}, X)$ of the Laplace distribution with parameters 0 and 1 for $n = 1, \ldots, 10\,000$ points.

repeated nearest neighbor searches among the components of the current n-quantizers. We present in the following their greedy variants.

Multi-dimensional greedy randomized Lloyd's procedure

Just as in the one-dimensional case, the greedy Lloyd's I procedure to compute a_n, assuming that $a^{(n-1)}$ is already known, is defined, in the quadratic case, by the following recursion

$$a_{n,[l+1]} = \mathbb{E}(X \,|\, X \in W_{n,[l]}), \quad a_{n,[0]} \in \mathbb{R}^d \setminus \{a^{(n-1)}\}, \tag{4.5}$$

where $W_{n,[l]}$ is the closed Voronoï cell of $a_{n,[l]}$ with respect to the quantizer $a^{(n-1)} \cup \{a_{n,[l]}\}$.

From a practical point of view, the conditional expectations are computed by a Monte Carlo simulation (provided X can be simulated at a reasonable cost). In other words, by the Strong Law of Large Numbers

$$a_{n,[l+1]} = \lim_{M \to +\infty} \frac{\sum_{m=1}^{M} X^m \mathbb{1}_{\{X^m \in W_{n,[l]}\}}}{\sum_{m=1}^{M} \mathbb{1}_{\{X^m \in W_{n,[l]}\}}} \quad \mathbb{P}\text{-}a.s.$$

where $(X^m)_{m \geq 1}$ is an i.i.d. sequence of copies of X (with distribution P) defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.
The existence of the above limit is given in the proposition below, at least for absolutely continuous distributions with convex support.

Proposition 4.1.2. *Assume the distribution P of X is strongly continuous with a convex support. Then the sequence $(a_{n,[l]})_{l \geq 0}$ is bounded and there exists $\ell \in \big[e_2(a^{(n)}, P),\, e_2(a^{(n-1)} \cup \{a_{[0]}\}, P)\big]$ such that the set $\mathcal{A}_\infty(a_{[0]})$ of its limiting points is a connected compact subset of the set Λ_ℓ of ℓ-*stationary points *defined by*

$$\Lambda_\ell = \Big\{ a \in \mathbb{R}^d \,|\, e_{2,n}(a^{(n-1)} \cup \{a\}) = \ell \;\; \textit{and} \;\; a = \mathbb{E}(X \,|\, X \in W_{n,a}) \Big\}$$

where $W_{n,a}$ denotes the closed Voronoï cell of centroid a induced by the n-quantizer $a^{(n-1)} \cup \{a\}$. In particular, $e_2(a^{(n-1)} \cup \{a_{[l]}\}, X) \to \ell$ as $l \to +\infty$.

*Furthermore, if the ℓ-*stationary set Λ_ℓ *is locally finite (i.e. with a finite trace on compact sets of $\mathbb{R}^d$), then $a_{n,[l]}$ a.s. converges to some point in Λ_ℓ.*

In higher dimensions, the equilibrium point is not unique. So, in theory, this limit may be just a local minimizer and not the solution to our greedy optimization problem. However, in practice, the results turn out to be satisfying.

Multidimensional Competitive Learning Vector Quantization procedure

This stochastic gradient descent algorithm (zero search algorithm) is defined by the following recursion

$$a_{n,[l+1]} = a_{n,[l]} - \gamma_{l+1} \mathbb{1}_{\left\{|X^{l+1} - a_{n,[l]}| < \min_{a \in a^{(n-1)}} |X^{l+1} - a|\right\}} \left(a_{n,[l]} - X^{l+1}\right),\ a_{l,[0]} \in \mathbb{R}^d.$$

where $(\gamma_l)_{l \geq 1}$ is a sequence of $(0,1)$-valued step parameters satisfying the so-called *decreasing step assumption*, namely $\sum_l \gamma_l = +\infty$ and $\sum_l \gamma_l^2 < +\infty$.
Numerical experiments show that $\lim_{l \to +\infty} a_{n,[l]} = a_n$, at least for distributions with compact convex support. Furthermore, one can speed up the convergence of the procedure by applying the so-called *Ruppert-Polyak principle* which consists in choosing a slowly decreasing step of the form $\gamma_l = \frac{c}{c+l^\alpha}$, $\frac{1}{2} < \alpha < 1$, and averaging the procedure by setting

$$\bar{a}_{n,[l]} = \frac{1}{l}\left(a_{[n,0]} + \cdots + a_{[n,l-1]}\right),\ l \geq 1,$$

In other words, it will satisfy a Central Limit Theorem at rate $\sqrt{n}$ with the lowest possible asymptotic variance (see *e.g.* [43, 57] for details).

However, these procedures are very demanding. There is so many integrals that intervene especially in the computation of the local inter-point inertia needed to decide in which cell we should add the new point. The improvements applied to the algorithm, as explained in Chapter 3, allow a reduction in the cost of the implementations, but the remaining computations are still too demanding and expensive. That is why we presented, in Chapter 3, what is called *greedy product quantization* consisting in obtaining multi-dimensional greedy sequences by computing the tensor product of several one-dimensional sequences, when the target law is a tensor product of its independent marginal laws. We give, in this chapter, one further example of this technique and compute the greedy product quantization sequence of the Normal distribution $P = \mathcal{N}(0, I_2)$. Noting that $P = P_1 \otimes P_2$ where $P_1 = P_2 = \mathcal{N}(0,1)$, we use two identical copies of the one-dimensional greedy quantization sequence $a^{(n)}$ of $\mathcal{N}(0,1)$ of same size n (already computed and stocked) and build the two-dimensional sequence $x^{(n^2)}$ of $\mathcal{N}(0, I_2)$. In Figure 4.5, we expose the weights of the Voronoï cells of the sequence $x^{(n^2)}$ for $n = 170$ and $n = 127 = 2^7 - 1$, where we clearly observe the bell curve in the case of $n = 127$, hence highlighting the existence of optimal sub-sequences of the form $x^{(n^2)}$ for $n = 2^k - 1, k \in \mathbb{N}^*$, even when we design product sequences. This means that, for these sequences, the empirical weighted measure $\sum_{i=1}^{n^2} p_i^{(n^2)} \delta_{x_i^{(n^2)}}$ approximates best the distribution $\mathcal{N}(0, I_2)$.

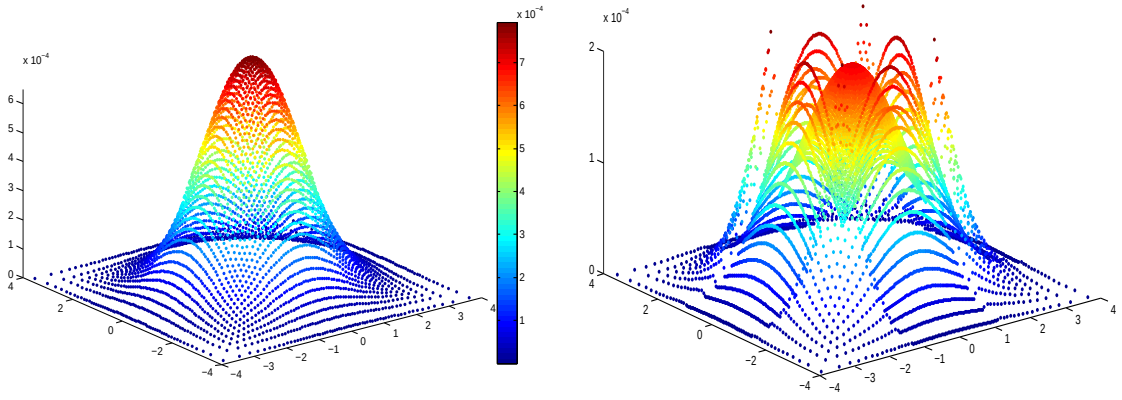

Figure 4.5: Representation of the Voronoï weights associated to the greedy product quantization sequence of the Normal distribution $\mathcal{N}(0, I_2)$ obtained using two 1-dimensional grids of size $n = 127$ (left) and $n = 170$ (right).

4.2 Deterministic algorithm in the two-dimensional case

We consider a random variable X taking values in $\mathbb{R}^2$, with absolutely continuous distribution P with density φ and we work in a quadratic framework. Deterministic variants of greedy Lloyd and greedy CLVQ can be extended to the two-dimensional framework. When $d = 1$, the key to having deterministic procedures is that the Voronoï cells corresponding to the greedy quantization sequence and the domains over which we need to integrate specific functions are intervals of $\mathbb{R}$ of the form $[a, b]$ which allowed the exact computation of the expectations and probabilities, in the expression of the local inertias and the recurrence of the algorithm. When $d = 2$, the corresponding Voronoï cells and the domains over which we need to integrate are convex polygonial sets. Integrals over these sets cannot be computed exactly but there exist numerical techniques able to approach them in a very effective and deterministic way. The idea is to decompose each (polygonial) domain into several triangles and use quadrature formulas to integrate the desired functions over these triangles.

The steps to follow for the greedy procedures when $d = 2$ are the same as in the one-dimensional case. Starting at $a_0 = \mathbb{E}[X]$ and assuming that $a^{(n)} = \{a_1, \ldots, a_n\}$ is already computed, we add the $(n + 1)$-th point while the others remain frozen. The procedure is detailed below where we denote by $T = (x, y, z)$ the triangle whose vertices are the three points x, y and z of $\mathbb{R}^2$.

First note that if the support of X is not compact, we start by truncating it into a bounded domain. Take, for example, the Gaussian distribution $\mathcal{N}(0, I_2)$ whose support is $\mathbb{R}^2$ but satisfies, for a certain $L > 0$ large enough, $P(X \notin [-L, L] \times [-L, L]) \approx 0$, so it is natural to truncate the support and consider the square $[-L, L] \times [-L, L]$ as a workspace.

Computation of the local inter-point inertias

Consider three neighboring points a_i, a_j and a_k of the sequence $a^{(n)}$ and let $T_\ell = (a_i, a_j, a_k)$ be a triangle whose vertices are these three points. The local interpoint inertia between them is

given by

$$\sigma_\ell^2 = \int_{W_i(a^{(n)})\cap T_\ell} (a_i-\xi)^2\varphi(\xi)d\xi + \int_{W_j(a^{(n)})\cap T_\ell} (a_j-\xi)^2\varphi(\xi)d\xi + \int_{W_k(a^{(n)})\cap T_\ell} (a_k-\xi)^2\varphi(\xi)d\xi. \quad (4.6)$$

To compute these inertias, one needs to handle the computation of integrals of the form $\int_D f(\xi)d\xi$ where D is the intersection between T_ℓ and the Voronoï cell and takes the form of a convex quadrilater. Such a problem is solved as follows: Let v_m be the vertice common to the three cells $W_i(a^{(n)})$, $W_j(a^{(n)})$ and $W_k(a^{(n)})$, i.e.

$$v_m = W_i(a^{(n)}) \cap W_j(a^{(n)}) \cap W_k(a^{(n)}).$$

We denote

$$\begin{aligned} v_{m+1} &= W_i(a^{(n)}) \cap W_j(a^{(n)}) \setminus \{v_m\} \\ v_{m+2} &= W_i(a^{(n)}) \cap W_k(a^{(n)}) \setminus \{v_m\} \\ v_{m+3} &= W_j(a^{(n)}) \cap W_k(a^{(n)}) \setminus \{v_m\}. \end{aligned}$$

Furthermore, we denote $c_{m+1} = (a_ia_j) \cap (v_mv_{m+1})$ the intersection between the line formed by a_i and a_j and the line formed by v_m and v_{m+1}, and $c_{m+2} = (a_ia_k) \cap (v_mv_{m+2})$ the intersection between the line formed by a_i and a_k and the line formed by v_m and v_{m+2}, and $c_{m+3} = (a_ia_k) \cap (v_mv_{m+3})$ the intersection between the line formed by a_i and a_k and the line formed by v_m and v_{m+3} (see Figure 4.6).
We start by decomposing each quadrilater D, over which we want to integrate, into 2 triangles $D = t_{m_1} \cup t_{m_2}$, which means that the triangle T_ℓ is itself divided in a total of 6 triangles $t_1, \dots, t_6$. For example, $D = W_i(a^{(n)}) \cap T_\ell$ is divided into 2 triangles $t_1 = (a_i, v_m, c_{m+1})$ and $t_6 = (a_i, v_m, c_{m+2})$, as showed in Figure 4.6.
This way, the integral over D will be of the form

$$\int_D f(\xi)d\xi = \int_{t_{m_1}} f(\xi)d\xi + \int_{t_{m_2}} f(\xi)d\xi$$

and the local inter-point inertia is given by the sum of 6 integrals (not necessarily of the same function), each over a triangle $t_m, m \in \{1, \dots, 6\}$, which are approximated by

$$\int_{t_m} f(\xi)d\xi \approx A_{t_m} \sum_{k=1}^{K} \omega_k f(x_k) \quad (4.7)$$

where A_{t_m} is the area of the triangle t_m and the points x_k and the weights ω_k are given by the quadrature formulas over triangles as provided in [69].

When working close to the sides of the square $[-L, L] \times [-L, L]$, the vertices of the triangle T_ℓ are no longer three neighboring points of the sequence. Instead, T_ℓ takes one of the two following forms: The first possibility is $T_\ell = (a_i, a_j, v_m)$ where a_i and a_j are two neighboring centroids at the edge and $v_m = W_i(a^{(n)}) \cap W_j(a^{(n)}) \cap [-L, L]^2$ (see Figure 4.7 (left)). In this case, the inertia is given by

$$\sigma_\ell^2 = \int_{W_i(a^{(n)})\cap T_\ell} (a_i - \xi)^2\varphi(\xi)d\xi + \int_{W_j(a^{(n)})\cap T_\ell} (a_j - \xi)^2\varphi(\xi)d\xi \quad (4.8)$$

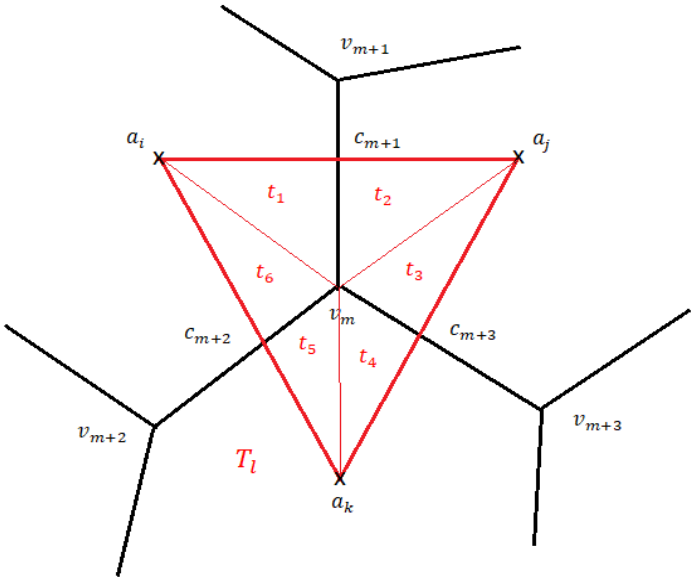

Figure 4.6: Decomposition of T_ℓ in the center in order to compute the local inter-point inertia.

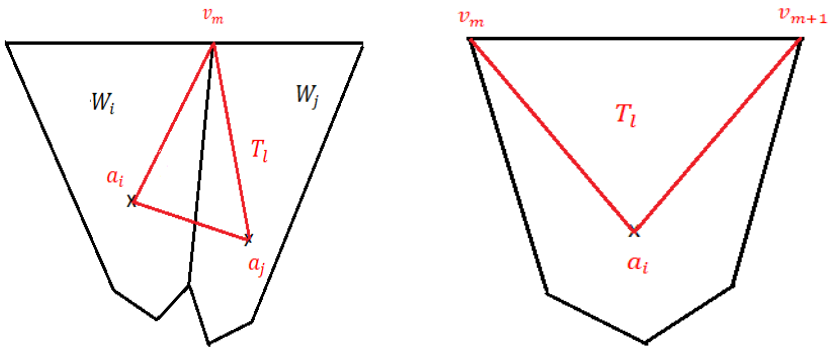

Figure 4.7: Decomposition of T_ℓ at the edge in order to compute the local inter-point inertia.

and the domain $D = W_i \cap T_\ell$ over which we need to integrate is a triangle so there is no need to divide it. It suffices to denote $T_\ell = t_{m_1} \cup t_{m_2}$ and apply (4.7) twice.
The second possibility is $T_\ell = (a_i, v_m, v_{m+1})$ where $\{v_m, v_{m+1}\} = W_i(a^{(n)}) \cap [-L, L]^2$ (see Figure 4.7 (right)). In this case,

$$\sigma_\ell^2 = \int_{T_\ell} (a_i - \xi)^2 \varphi(\xi) d\xi \tag{4.9}$$

and the integral over the triangle T_ℓ is performed via (4.7).

Addition of a new point

After computing the local inertias, we proceed by choosing the triangle T_{ℓ_0} having the maximal inter-point inertia among all the triangles T_ℓ. Then, we add a new point a_0 to the greedy quantization sequence as the barycenter of T_{ℓ_0} w.r.t. the underlying probability distribution P,

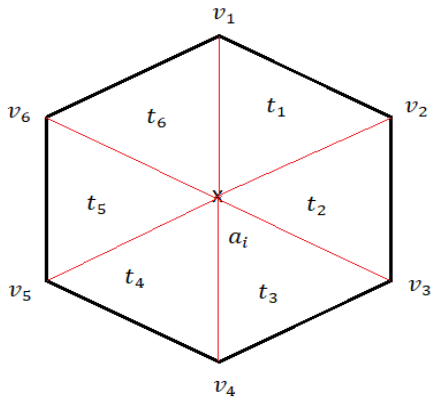

Figure 4.8: Decomposition of a Voronoï cell $W_i(a^{(n)})$ with $s = 6$ vertices.

i.e.

$$a_0 = \frac{\mathbb{E}\left(X \mathbb{1}_{X \in T_{\ell_0}}\right)}{P(X \in T_{\ell_0})} = \frac{\int_{T_{\ell_0}} \xi \varphi(\xi) d\xi}{\int_{T_{\ell_0}} \varphi(\xi) d\xi}. \tag{4.10}$$

To compute these integrals, we decompose T_{ℓ_0} in the same way as explained for the computation of the local inertia and apply (4.7).

Lloyd's algorithm

The point added in the previous step is the starting point of a fixed-point search defined by the following recursion: $a_{[0]} = a_0$ given by (4.10) and, for every $l \geq 1$,

$$a_{[l]} = \frac{\mathbb{E}\left(X \mathbb{1}_{X \in W_{[l]}}\right)}{P(X \in W_{[l]})} = \frac{\int_{W_{[l]}} \xi \varphi(\xi) d\xi}{\int_{W_{[l]}} \varphi(\xi) d\xi}$$

where $W_{[l]}$ is the Voronoï cell of centoid $a_{[l]}$ in the Voronoï diagram corresponding to the sequence $a^{(n)} \cup \{a_{[l]}\}$. This recurrence converges well to the solution $a_{[\infty]}$ of the fixed-point search problem allowing us to obtain the $(n+1)$-th point a_{n+1} of the greedy quantization sequence.

Since $W_{[l]}(a^{(n)})$ is a convex polygon with s vertices $v_1, \ldots, v_s$, $s \geq 3$, we proceed as before to compute the above integrals. We start by dividing the cell $W_{[l]}(a^{(n)})$ into s triangles $t_m = (a_{[l]}, v_m, v_{m+1})$, $m \in \{1, \ldots, s\}$, as shown in Figure 4.8 so that

$$\int_{W_{[l]}(a^{(n)})} f(\xi) d\xi = \sum_{m=1}^{s} \int_{t_m} f(\xi) d\xi$$

and the integrals over the triangles t_m are computed by (4.7).

In Figure 4.9, we observe the Voronoï diagram of a greedy quantization sequence $a^{(n)}$ of the Normal distribution $\mathcal{N}(0, I_2)$ designed by a deterministic Lloyd's algorithm. We expose the sequences obtained at several steps of the algorithm, i.e. sequences of different sizes n between $n = 6$ and $n = 100$ to emphasize the recursive dynamic of greedy quatization. The represented sizes are $n = 6, 7, 11, 16, 18, 24, 28, 32, 39, 51, 86, 100$.

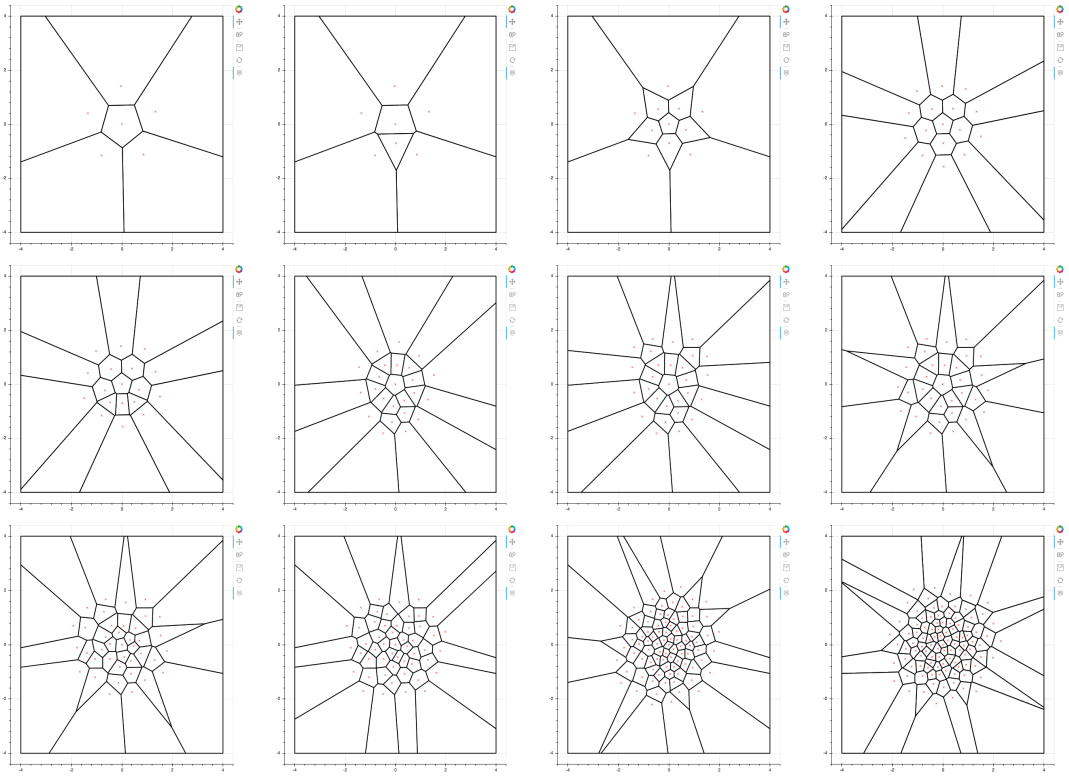

Figure 4.9: Greedy quantization sequences of $\mathcal{N}(0, I_2)$ obtained by a deterministic Lloyd's algorithm of sizes $n = 6, 7, 11, 16, 18, 24, 28, 32, 39, 51, 86, 100$ (starting from the upper left corner).

4.3 Low discrepancy sequences viewed as quantization sequences

In Section 3.6.4 of Chapter 3, we explain the interest in comparing greedy quantization sequences to low discrepancy sequences. The most important advantage of quantization is the gain of a $\log(n)$-factor in the rate of convergence of quantization-based numerical integration error. Furthermore, based on Proïnov's Theorem (3.4.1), we were motivated to notice a link between the discrepancy of a sequence Ξ and the quantization error induced by this sequence with respect to the Uniform distribution given by

$$e_1(\Xi, \mathcal{U}([0,1]^d)) \leq D_n^*(\Xi)^{\frac{1}{d}}.$$

This led us to study the discrepancy of greedy quantization sequences which gave non-drastic but also non-reliable results (see Section 3.6.4 of Chapter 3). That pushed us to tackle the opposite problem which is trying to manipulate low discrepancy sequences, such as Van der Corput, Halton sequences, Niederreiter and others, in order to be able to use them as greedy quantization sequences. In other words, we assign to these particular sequences a Voronoï diagram, give weights to the corresponding Voronoï cells, compute the quantization error hence obtained and observe their behavior. We expose here further details of this study.

Van der Corput sequence We consider the dyadic Van der Corput *(VdC)* sequence $\Xi = (\xi_n)_n$ defined by

$$\xi_n = \sum_{k=0}^{r} \frac{a_k}{2^{k+1}} \quad \text{where} \quad n = a_r 2^r + \cdots + a_0,\ a_i \in \{0,1\},\ i = 1, \ldots, r.$$

When $d = 1$, the Voronoï diagram is trivial and given by $(W_i)_{1 \le i \le n}$ where

$$W_1(\Xi) = [0, \xi_{1+\frac{1}{2}}) \qquad W_n(\Xi) = [\xi_{n-\frac{1}{2}}, 1] \qquad W_i(\Xi) = [\xi_{i-\frac{1}{2}}, \xi_{i+\frac{1}{2}}],\ 1 < i < n$$

where $\xi_{i+\frac{1}{2}} = \frac{\xi_i + \xi_{i+1}}{2},\ i \in \{1, \ldots, n-1\}$. We manipulate this sequence as a quantization sequence of the Uniform distribution $\mathcal{U}([0,1])$ and start by studying the corresponding quadratic quantization error

$$e_2^2(X, \mathcal{U}([0,1])) = \int_0^1 \min_{1 \le i \le n} |\xi_i - u|^2 du = \sum_{i=1}^{n} \int_{\xi_{i-\frac{1}{2}}}^{\xi_{i+\frac{1}{2}}} |\xi_i - u|^2 du$$

We observe, in Figure 4.10, the graph representing $n \to ne_2(\Xi, \mathcal{U}([0,1]))$ where we notice that

$$\liminf_n ne_2(\Xi, \mathcal{U}([0,1])) = \frac{1}{2\sqrt{3}} = \widetilde{J}_{2,1} \quad \text{and} \quad \limsup_n ne_2(\Xi, \mathcal{U}([0,1])) = \frac{3\sqrt{5}}{4} \times \widetilde{J}_{2,1}$$

keeping in mind that $\widetilde{J}_{2,1} = \lim_n ne_{2,n}(\mathcal{U}([0,1])) = \inf_n ne_{2,n}(\mathcal{U}([0,1]))$ is the sharp limiting constant in Zador's Theorem (1.5). The convergence rate of this error towards 0 is of $\mathcal{O}(n^{-1})$, similar to that of a real greedy quantization sequence.

The same phenomenons are observed in the L^1-case where

$$\liminf_n ne_1(\Xi, \mathcal{U}([0,1])) = \frac{1}{4} = \widetilde{J}_{1,1} \quad \text{and} \quad \limsup_n ne_1(\Xi, \mathcal{U}([0,1])) = \frac{9}{32} = \frac{9}{8}\,\widetilde{J}_{1,1}$$

where $\widetilde{J}_{1,1}$ is as well the constant given by Zador Theorem for $p = d = 1$. This lim inf is achieved by sub-sequences of Ξ of size 2^{k-1}, $k \ge 1$, and the lim sup achieved by sub-sequences of Ξ of size $\frac{3}{2}.2^k = 3.2^{k-1}$, $k \ge 1$. This leads us to claim that there exist rate optimal sequences, i.e. whose corresponding quantizaton error ocnverges to 0 with an $\mathcal{O}(n^{-\frac{1}{d}})$-rate of decay, which are not solutions to the greedy problem (4.1).

From another point of view, we consider, instead of uniform Voronoï weights equal to $\frac{1}{n}$, new weights (easy to compute) given, for every $i \in \{1, \ldots, n\}$, by

$$p_i^n = \int_{W_i(\Xi)} dP = P(W_i(\Xi)).$$

Numerical implementations show that, when the size of the sequence is equal to $2^k, k \ge 1$, the weights of the Voronoï cells induced by the VdC sequence are uniform. For comparison purposes, an example was established in Chapter 3 to study the difference brought by the use of such weights instead of uniform weights where we consider a basic example of pricing a european call $C_0 = \mathbb{E}[(X_T - K)_+]$ for a maturity T and a strike price K where the price of the asset X_t at a time t evolves following a Black-Scholes model.

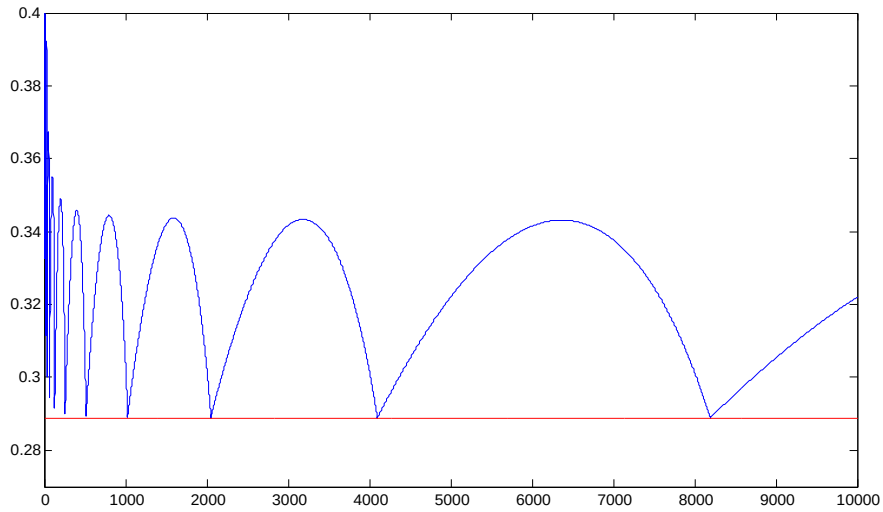

Figure 4.10: Quadratic quantization error of the Van der Corput sequence viewed as a quantization sequence (logarithmic scale).

Niederreiter sequence in dimension 2 We consider a two-dimensional Niederreiter sequence $\Xi = (\xi_i)_{1\leq i\leq n}$ and we aim to apply the same study already established for the Van der Corput sequence previously. When $d = 2$, the corresponding Voronoï cells are harder to define thus the computation of the quantization error corresponding to Ξ becomes more complicated and a deterministic computation seems to be impossible, this is due to the integrals appearing in this case. To this end, one needs to do a stochastic computation based on large Monte Carlo simulations coupled with a nearest neighbor search procedure. We compute the quadratic quantization error $e_2(\Xi, \mathcal{U}([0,1]^2)$ and expose it in Figure 4.11 in a logarithmic scale where we observe an $\mathcal{O}(n^{-1})$-rate of convergence.

Furthermore, we assign non-uniform weights for the cells of the Voronoï diagram induced by the 2-dimensional Niederreiter sequence, via Monte carlo simulations. To study the utility of such non-uniform weights, we consider the example of a European Best-of-Call Vanilla option of maturity T and strike price K given by

$$V_0 = e^{-rT}\mathbb{E}[\big(\max(X_T^1, X_T^2) - K\big)_+]$$

where r is the interest rate and X_T^1 and X_T^2 are 2 risky assets in a 2-dimensional Black-Scholes model given as follows

$$X_0^1 = X_0^2 = e^{-rT}, \qquad X_t^i = X_0^i \exp\Big((r - \frac{\sigma_i^2}{2})t + \sigma_i W_t^i\Big), \quad i = 1, 2,$$

where(W_t^1, W_t^2) is a correlated Brownian motion, i.e. $W_t^2 = \rho W_t^1 + \sqrt{1-\rho^2}\,\widetilde{W}_t^2$ where $(W_t^1, \widetilde{W}_t^2)$ is a standard Brownian motion. We consider

$$T = 1\,, \quad K = 100\,, \quad X_0^1 = X_0^2 = 100\,, \quad \rho = 0.5\,, \quad \sigma_1 = \sigma_2 = 0.2\,, \quad r = 0.1$$

and we compute the price of V_0 via a classical quadrature formula using the new weights p_i^n assigned to the Niederreiter sequence instead of uniform weights. The benchmark is given in

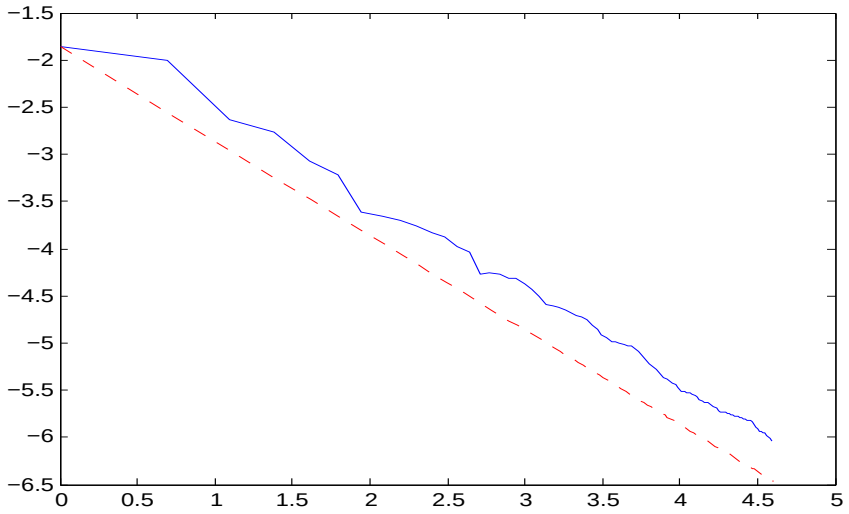

Figure 4.11: Quadratic quantization error of the two-dimensional Niederreiter sequence viewed as a quantization sequence for the $\mathcal{U}([0,1]^2)$ distribution. (logarithmic scale).

[57].
We compare, in Figure 4.12, the error induced by this approximation to the one obtained by a classical quasi-Monte Carlo method (i.e. where we use the uniform weights of the Niederreiter sequence) and to the one obtained by a quantization-based numerical integration quadrature formula using a greedy quantization sequence of the $\mathcal{U}([0,1])$-distribution. We conclude with the same observations as in the one-dimensional case (see Chapter 3), the convergence of greedy quantization-based procedures is more important than Niederreiter-based procedures.

4.4 To what extent are greedy quantization sequences optimal?

Based on the studies established so far in this chapter, one wonders if there is a method to produce, for any distribution P, a rate optimal sequence for L^p-quantization. In fact, one checks that it is possible by *concatenating* L^p-optimal grids of size 2^ℓ. We consider a sequence $(b_n)_{n\geq 1}$ made up with (L^p,P)-optimal quantizers at level 2^ℓ, $\ell\geq 0$ *i.e.* in a way that $\left\{b_{2^\ell},\ldots,b_{2^{\ell+1}-1}\right\}$ is an (L^p,P)-optimal quantizer at level 2^ℓ. For every $n\geq 1$, let $k=k(n)$ be such that $2^k-1\leq n\leq 2^{k+1}$ so that, by monotony of the L^p-quantization error, one has, for every $k\geq 1$,

$$e_p(b^{(n)},P)\leq e_p(b^{(2^k-1)},P)\leq e_p(\{b_{2^{k-1}},\ldots,b_{2^k-1}\})=e_{p,2^{k-1}}(P)$$

so that

$$\limsup_n n^{\frac{1}{d}}e_p(b^{(n)},P)\leq \limsup_n\left(\frac{n}{2^{k(n)}}\right)^{\frac{1}{d}}\lim_n n^{\frac{1}{d}}e_{p,n}(P)=2^{\frac{1}{d}}\lim_n n^{\frac{1}{d}}e_{p,n}(P).$$

Numerical observations
• If $P=U([0,1])$ and $p=1$, one checks by induction that the dyadic *VdC* sequence can be obtained as a reordered sequence $(b_n)_{n\geq 1}$ from the L^p-optimal quantizers at level n given by

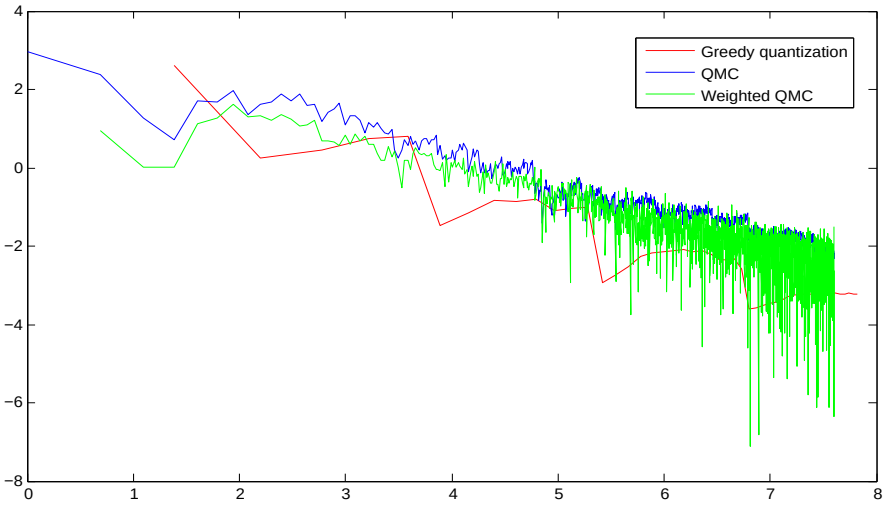

Figure 4.12: Price of a European Best-of-Call Vanilla option in a Black-Scholes model via a usual QMC method (blue), greedy quantization-based quadrature formula (red) and quadrature formula using 2-dimensional Niederreiter sequence with non-uniform weights (logarithmic scale).

$\left\{\frac{2k-1}{2n},\, 1 \le k \le n\right\}$ when $n = 2^\ell$, $\ell \ge 0$. In this situation, as seen in the previous section, the factor $2^{\frac{1}{d}} = 2$ is replaced by $\frac{9}{8} = 1.125$ and the L^1-optimal greedy quantization sequence keeps the lead, since we have already seen that

$$\limsup_n \frac{e_1(a^{(n),1}, P)}{e_{1,n}(P)} \approx 1.09 < \frac{9}{8} \approx 1.125$$

• If $P = U([0,1])$ and $p = 2$, once again, the quadratic optimal greedy quantization sequence keeps the lead, since

$$\limsup_n \frac{e_2(a^{(n),2}, P)}{e_{2,n}(P)} \approx 1.13401 < \frac{3\sqrt{5}}{4} \approx 1.67706 < 2.$$

• If $P = \mathcal{N}(0, I_2)$ and $d = p = 2$, numerical experiments suggest for the third time that a quadratic optimal greedy quantization sequence (or, actually, the sub-optimal sequences) has a lower constant than $2^{\frac{1}{d}} \times \lim_n n^{\frac{1}{2}} e_{2,n}(\mathcal{N}(0; I_2))$.

These experiments lead us to wonder if optimal greedy quantization sequences produce the lowest value for $\limsup_n n^{\frac{1}{d}}\, e_{p,n}(a^{(n)}, P)$ or if the strict inequality $\limsup_N \frac{e_p(a^{(N),p}, P)}{e_{p,N}(P)} < 2^{\frac{1}{d}}$ is always satisfied.

4.5 Quasi-stationarity and ρ-quasi stationarity

Quadratic optimal quantizers share a property called *stationarity* that is very important in most applications, especially since most algorithms devised to compute optimal n-quantizers are

based on this stationarity property. Moreover, its importance is emphasized in the quantization-based numerical integration. In fact, if $\widehat{X}^{\Gamma_n}$ is an optimal quantizer of X induced by the grid $\Gamma_n = \{x_1^n, \ldots, x_n^n\}$, then it is already known that

$$\|X - \widehat{X}^{\Gamma_n}\|_p = e_p(\Gamma_n, X) \underset{n\to+\infty}{\longrightarrow} 0.$$

then, we have the convergence of $\widehat{X}^{\Gamma_n}$ towards X in $L^p(\mathbb{P})$ when $n \to +\infty$ and consequently the convergence in distribution. In particular, if $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous bounded function, then one deduces that $\mathbb{E}f(\widehat{X}^{\Gamma_n}) \to \mathbb{E}f(X)$ when $n \to +\infty$. Consequently, using the weights $(p_i^n)_{1\leq i\leq n}$ of the Voronoï cells corresponding to Γ_n, one approaches $\mathbb{E}f(X)$ by

$$\mathbb{E}f(\widehat{X}^{\Gamma_n}) = \sum_{i=1}^{n} p_i^n f(x_i^n). \tag{4.11}$$

Error bounds induced by this approximation are established for various classes of functions f and, in most cases, it is mainly due to the stationarity property shared by optimal quadratic quantizers. For more details, we refer to [56].

It was mentioned in the first chapter that we tried to extend this property to greedy quantizaton sequences, i.e. to see if

$$a_i^{(n)} = \mathbb{E}(X|X \in W_i(a^{(n)})), \qquad i = 1, \ldots, n.$$

Unfortunately, numerical experiments computing the error $\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_1$ under the standard empirical measure $\widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{a_i^{(n)}}$ gave negative results. In fact, we show below that when the distribution is symmetrical and unimodal, the corresponding greedy quantization sequence cannot be stationary except for $n \in \{1; 3\}$. For the proof, we rely on a result given in [39].

Theorem 4.5.1. *(J.C. Kieffer) Let $d = 1$ and P a probability distribution with log-concave density. Then, there exists a unique stationary quantizer of P.*

Proposition 4.5.2. *Let X be a random variable with distribution P which is symmetric and unimodal (log-concave density) and $a^{(n)}$ a corresponding greedy quantization sequence. Then, for every $n \in \mathbb{N} \setminus \{1, 3\}$, the sequence $a^{(n)}$ is not stationary.*

Proof. We suppose that $\mathbb{E}[X] = 0$ (symmetric around 0). If it is not the case, a translation gives the same results. We will detail the proof in 3 cases
▷ **For $n = 3$:** Since $\mathbb{E}[X] = 0$, the first point is $a_1 = 0$. A second point is given by

$$a_2 = \text{argmin}_{a\in\mathbb{R}}\mathbb{E}X^2 \wedge (X-a)^2 = \{\nabla_{a_2}\, \mathbb{E}X^2 \wedge (X-a_2)^2 = \int_{W_2(a^{(i)})} (\xi - a_2)dP(\xi) = 0\}.$$

Hence, $a_2 = \frac{\int_{W_2(a^{(n)})}\xi dP(\xi)}{P(W_2(a^{(n)}))}$ is stationary. The third point is $a_3 = -a_2$ by symmetry of P so a_3 is also stationary. Finally, $a_1 = 0$ is also stationary since $\int_{a_2/2}^{a_3/2}\xi dP(\xi) = \int_{a_2/2}^{-a_2/2}\xi dP(\xi) = 0$. Consequently, the sequence $a^{(3)} = \{-a_2; a_1; a_2\}$ is stationary.
▷ **For $n = 2k$ even:** Since P is unimodal, the stationary quantizer is unique, let $x^{(n)}$ be

this quantizer, which is the n-optimal quantizer of P because we know it is stationary. The symmetry of P lets us know that the quantizer $(x^{(n)}_{n+1-l})_{1\leq l\leq n}$ of P is also stationary, so, for every $l\in\{1,\dots,n\}$, $x^{(n)}_l + x^{(n)}_{n+1-l} = 0$. Since $n = 2k$ is even, we have, in particular,

$$x^{(n)}_k = -x^{(n)}_{n+1-k} = -x^{(n)}_{n+1-\frac{n}{2}} = -x^{(n)}_{k+1},$$

so, $x^{(n)}_k < 0 < x^{(n)}_{k+1}$ and, since, $x^{(n)}_k$ et $x^{(n)}_{k+1}$ are two consecutive terms of the grid, we deduce that 0 is not an element of $x^{(n)}$, and hence can not be a point of a stationary quantizer. Consequently, the greedy sequence starting at $a_1 = 0$ can not be stationary.
▷ **For $n = 2k+1$ odd:** First, notice that, in the greedy non-stationary sequence $a^{(2k)}$, there exists, at least, two non-stationary Voronoï cells, a first non-stationary cell $W_i(a^{(2k)})$ and its symmetric cell $W_{2k+1-i}(a^{(2k)})$ which is also non-stationary due to the symmetry of P. To build the sequence $a^{(2k+1)}$, we add a new point in one of the Voronoï cells without modifying the others. If the new point is added in one of the non-stationary cells, we know that the second one will remain untouched, having, at least, one non-stationary cell in $a^{(2k+1)}$. And, if the new point is not in these cells, then they will remain untouched and there will be, at least, 2 non-stationary cells in $a^{(2k+1)}$. □

However, we indicated that greedy quantization sequences satisfy a ρ-quasi-stationarity property approaching the stationary property and defined, for $r\in\{1,2\}$ and $\rho\in[0,1]$, by

$$\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_r = o(\|\widehat{X}^{a^{(n)}} - X\|^{1+\rho}_{1+\rho}), \quad \text{or} \quad \frac{\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_r}{\|\widehat{X}^{a^{(n)}} - X\|^{1+\rho}_{1+\rho}} \underset{n\to+\infty}{\longrightarrow} 0. \quad (4.12)$$

We detail in the following the study that allowed us to conclude with this conjecture.

We start by evaluating the error between $\widehat{X}^{a^{(n)}}$ and $\mathbb{E}(X|\widehat{X}^{a^{(n)}})$ under the weighted empirical measure $\widetilde{P}_n = \sum_{i=1}^n p_i^n \delta_{a_i^{(n)}}$, hoping that a change of measures will induce positive results. We compute the quadratic $L^2(\mathbb{R})$-error

$$\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_2 = \left(\sum_{i=1}^n p_i^{(n)} \left|a_i^{(n)} - \mathbb{E}(X|X\in W_i(a^{(n)}))\right|^2\right)^{\frac{1}{2}} \quad (4.13)$$

and the $L^1(\mathbb{R})$-error

$$\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_1 = \sum_{i=1}^n p_i^{(n)} \left|a_i^{(n)} - \mathbb{E}(X|X\in W_i(a^{(n)}))\right| \quad (4.14)$$

for the one-dimensional Gaussian $\mathcal{N}(0,1)$, Uniform $\mathcal{U}([0,1])$ and Exponential $\mathcal{E}(1)$ distributions.

First numerical observation The conducted experiments allow us to deduce that both errors (4.13) and (4.14) converge to 0 when n goes to infinity, for the 3 mentioned probability distributions. This can be explained by the convergence of the quadratic greedy quantization error towards 0 (that is already a well known result) and by the fact that

$$\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_1 \leq \|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_2 = \|\mathbb{E}(\widehat{X}^{a^{(n)}} - X|\widehat{X}^{a^{(n)}})\|_2 \leq \|\widehat{X}^{a^{(n)}} - X\|_2.$$

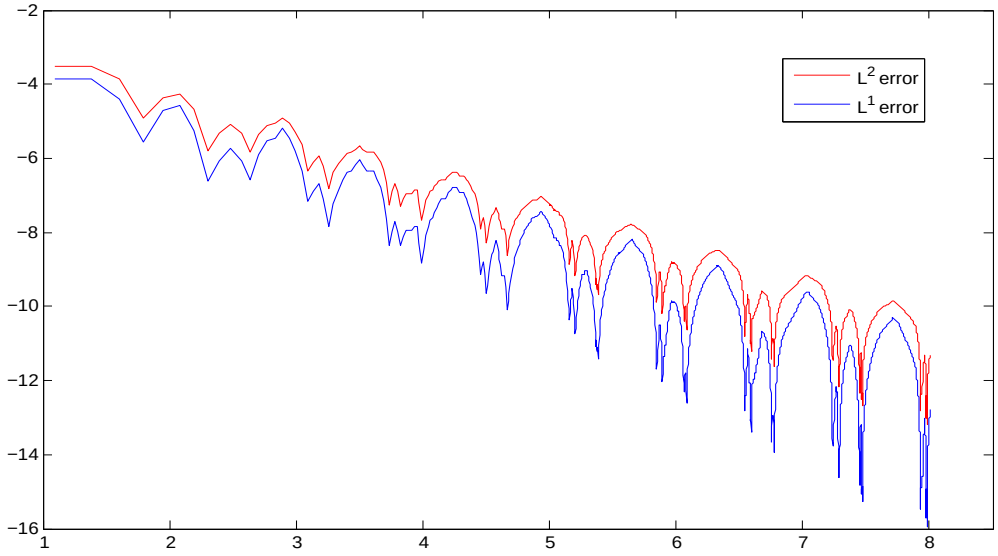

Figure 4.13: The errors $\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_2$ and $\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_1$ induced by a greedy quantization sequence $a^{(n)}$ corresponding to the distribution $\mathcal{U}([0,1])$ for $n = 1, \ldots, 1\,000$ (logarithmic scale).

We expose, in Figure 4.13, the convergence of the errors (4.13) and (4.14) induced by the greedy quantization sequence of the Uniform distribution of size n varying between 1 and 1 000, where we observe a faster convergence for the optimal sub-sequences of the greedy quantization sequence of $\mathcal{U}([0,1])$, given in Section 3.6.1 of the previous chapter, than for the greedy quantization sequence itself.

Based on the above results, one wonders if this convergence affects, in some way, the quantization-based numerical integration errors, or if maybe one needs to study some different (but in a way related) property than (4.13) and (4.14), to achieve improvements. Our motivation is the following.

Motivation If $f : \mathbb{R}^d \to \mathbb{R}$ is a function with a Lipschitz gradient and $[\nabla f]_{\text{Lip}}$ its Lipschitz coefficient, then, using the same notations as previously and noting $(.|.)$ a scalar product, one has

$$f(X) - f(\widehat{X}^{a^{(n)}}) - (\nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}) = \int_0^1 \left(\nabla f\left(\widehat{X}^{a^{(n)}} + t(X - \widehat{X}^{a^{(n)}})\right) - \nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}\right) dt.$$

Taking the expectation yields

$$\begin{aligned}&\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}}) - \mathbb{E}(\nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}})\\ &= \mathbb{E}\left[\int_0^1 \left(\nabla f\left(\widehat{X}^{a^{(n)}} + t(X - \widehat{X}^{a^{(n)}})\right) - \nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}\right) dt\right].\end{aligned}$$

Since

$$\mathbb{E}(\nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}) = \mathbb{E}(\nabla f(\widehat{X}^{a^{(n)}})|X) - \mathbb{E}(\nabla f(\widehat{X}^{a^{(n)}})|\widehat{X}^{a^{(n)}}) = \mathbb{E}\left(\nabla f(\widehat{X}^{a^{(n)}})|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\right), \tag{4.15}$$

and

$$\int_0^1 \left(\nabla f\left(\widehat{X}^{a^{(n)}} + t(X - \widehat{X}^{a^{(n)}})\right) - \nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}\right) dt \leq [\nabla f]_{\text{Lip}}\mathbb{E}|X - \widehat{X}^{a^{(n)}}|^2 \int_0^1 t dt,$$

then, owing to Minkowski and Cauchy-Schwarz inequalities, one obtains

$$|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})| \leq \|\nabla f(\widehat{X}^{a^{(n)}})\|_2 \|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\|_2 + \frac{1}{2}[\nabla f]_{\text{Lip}}\|X - \widehat{X}^{a^{(n)}}\|_2^2. \quad (4.16)$$

At this stage, one note that, since ∇f is lipschitz, then

$$\|\nabla f(\widehat{X}^{a^{(n)}})\|_2 \leq [\nabla f]_{\text{Lip}}\|\widehat{X}^{a^{(n)}}\|_2 + |\nabla f(0)|,$$

and, since $a^{(n)}$ is a greedy quantization sequence, then

$$\|\widehat{X}^{a^{(n)}}\|_2 \leq \|X - \widehat{X}^{a^{(n)}}\|_2 + \|X\|_2 = \| \min_{1\leq i\leq n} |X - a_i^{(n)}|\|_2 + \|X\|_2 \leq \|X - a_1^{(n)}\|_2 + \|X\|_2,$$

so that

$$\|\nabla f(\widehat{X}^{a^{(n)}})\|_2 \leq [\nabla f]_{\text{Lip}}(\|X - a_1^{(n)}\|_2 + \|X\|_2) + |\nabla f(0)| < +\infty.$$

Hence, we can hope that, if

$$\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_2 = o(\|\widehat{X}^{a^{(n)}} - X\|_2^2), \quad (4.17)$$

then

$$\limsup_n \frac{|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})|}{\|X - \widehat{X}^{a^{(n)}}\|_2^2} \leq \frac{1}{2}[\nabla f]_{\text{Lip}}.$$

This result provides an upper bound of the greedy quantization-based numerical integration error better than the one adopted till now. But this is true only if the sequence $a^{(n)}$ is asymptotically L^2-quasi stationary, i.e. satisfies (4.17).

Remark 4.5.3. *It is clear that the sequence $a^{(n)}$ satisfies (4.17) if, and only if,*

$$\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_2 = o(\|X - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_2^2). \quad (4.18)$$

In fact, for every $n \geq 1$, one has

$$\|\widehat{X}^{a^{(n)}} - X\|_2^2 = \|X - \mathbb{E}(X|\widehat{X}^{a^{(n)}}) + \mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\|_2^2.$$

Noting that $(\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})) \in L^2(\Omega, \sigma(\widehat{X}^{a^{(n)}}), P)$ and by definition of the conditional expectation $\mathbb{E}(.|\widehat{X}^{a^{(n)}})$ as the orthogonal projection in the space generated by the variable $\widehat{X}^{a^{(n)}}$, pythagoras Theorem yields

$$\|\widehat{X}^{a^{(n)}} - X\|_2^2 = \|X - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_2^2 + \|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\|_2^2,$$

So, condition (4.17) can also be read as

$$\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_2 = o\left(\|X - \mathbb{E}(X|\widehat{X}^{a^{(n)}}\|_2^2\right) + o\left(\|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\|_2\right)$$

which yields (4.18).

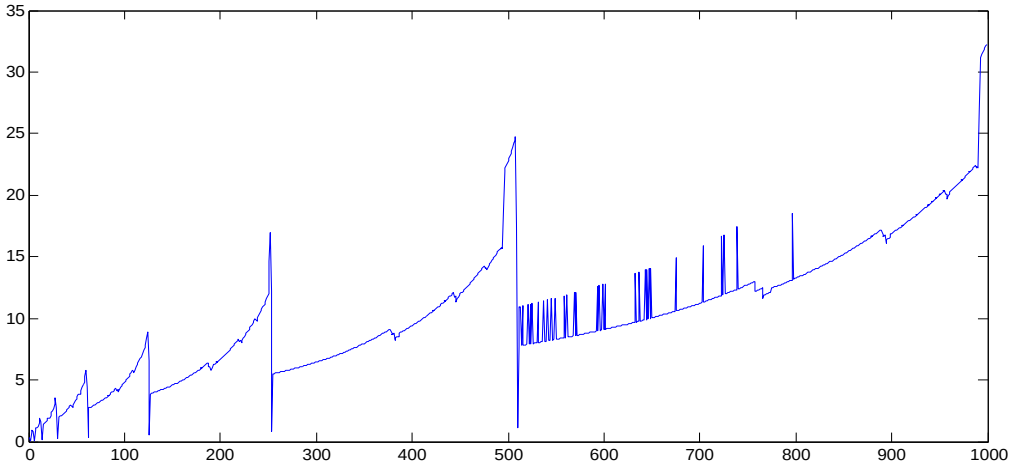

Figure 4.14: The error $\frac{\|\widehat{X}^{a^{(n)}}-\mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_2}{\|\widehat{X}^{a^{(n)}}-X\|_2^2}$ with $a^{(n)}$ a greedy sequence of the $\mathcal{N}(0,1)$ distribution for $n = 1, \dots, 1\,000$.

Likewise, one notes that if ∇f is simply bounded, then equation (4.16) becomes

$$|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})| \leq \|\nabla f(\widehat{X}^{a^{(n)}})\|_\infty \|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\|_1 + \frac{1}{2}[\nabla f]_{\text{Lip}} \|X - \widehat{X}^{a^{(n)}}\|_2^2, \quad (4.19)$$

where we can replace $\|\nabla f(\widehat{X}^{a^{(n)}})\|_\infty = [\nabla f]_{\text{Lip}}$ since ∇f is bounded. Then, the same arguments as previously lead us to hope that

$$\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_1 = o(\|\widehat{X}^{a^{(n)}} - X\|_2^2), \quad (4.20)$$

in order to obtain the upper bound

$$\limsup_n \frac{|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})|}{\|X - \widehat{X}^{a^{(n)}}\|_2^2} \leq \frac{1}{2}[\nabla f]_{\text{Lip}}.$$

Second numerical observation To test if a greedy quantization sequence is asymptotically L^p-quasi-stationary for $p \in \{1; 2\}$, i.e. if it satisfies (4.17) for $p = 2$ and (4.20) for $p = 1$, we compute the ratio

$$R_{p,2} = \frac{\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_p}{\|\widehat{X}^{a^{(n)}} - X\|_2^2}, \qquad p = 1; 2$$

for the probability distributions studied in this section, and we observe its behavior with respect to the size n of the sequence. The results show that the ratio does not converge towards 0. A divergence to $+\infty$ is observed for the whole greedy quantization sequence $a^{(n)}$ and for the optimal sub-sequences as well. Figure 4.14 depicts the behaviour of $R_{2,2}$ for a greedy quantization sequence $a^{(n)}$ of the Gaussian standard distribution of size n varying between 1 and 1 000.

Motivation Since the previous required result is not achieved, another motivation (different but somehow similar to the previous one) leads us to set another definition of an asymptotically quasi-stationary sequence, in the hopes of winning in terms of convergence of quantization-based numerical integration errors. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function with a continuous gradient. A Taylor-Young formula yields

$$f(X) - f(\widehat{X}^{a^{(n)}}) - (\nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}) = (X - \widehat{X}^{a^{(n)}})\varepsilon(X - \widehat{X}^{a^{(n)}})$$

where $\varepsilon(X - \widehat{X}^{a^{(n)}})$ is a function that converges to 0 when $\widehat{X}^{a^{(n)}}$ converges to X. Taking the expectation, one has

$$\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}}) - \mathbb{E}(\nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}) = \mathbb{E}[(X - \widehat{X}^{a^{(n)}})\varepsilon(X - \widehat{X}^{a^{(n)}})]$$

Since, $\varepsilon(X - \widehat{X}^{a^{(n)}})$ converges to 0, then there exists a small constant $c > 0$ such that $\varepsilon(X - \widehat{X}^{a^{(n)}}) < c$. Consequently, using (4.15), one has

$$|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})| \leq \|\nabla f(\widehat{X}^{a^{(n)}})\|_2 \|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\|_2 + c\|X - \widehat{X}^{a^{(n)}}\|_1. \tag{4.21}$$

Moreover, if ∇f is bounded, then (4.21) can also be written as

$$|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})| \leq \|\nabla f(\widehat{X}^{a^{(n)}})\|_\infty \|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\|_1 + c\|X - \widehat{X}^{a^{(n)}}\|_1. \tag{4.22}$$

Hence, a similar reasoning to the one established in the previous motivation pushes us to hope that

$$\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_p = o(\|\widehat{X}^{a^{(n)}} - X\|_1), \tag{4.23}$$

for $p \in \{1; 2\}$, in order to obtain the following upper error bound

$$\limsup_n \frac{|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})|}{\|X - \widehat{X}^{a^{(n)}}\|_1} \leq c.$$

Remark 4.5.4. *Before we move on to the numerical results, let us note that the constant c in the upper bound is not controlled. Thus, even if* (4.23) *is verified, the gain in numerical integration is not very remarkable, but it would be interesting to study this case to get an additional idea.*

Third numerical observation We are interested in the study of the convergence of the ratio

$$R_{p,1} = \frac{\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_p}{\|\widehat{X}^{a^{(n)}} - X\|_1}, \qquad \text{for } p, q \in \{1; 2\}.$$

Numerical experiments conducted for different distributions give interesting results. When considering Gaussian and Exponential distributions, the ratio $R_{p,1}$ converges to 0 with an $\mathcal{O}(n^{-\frac{1}{2}})$-rate of decay when $p = 2$ and an $\mathcal{O}(n^{-1})$-rate of decay when $p = 1$. However, for the Uniform distribution, similar observations are made only with optimal sub-sequences of the greedy sequence. Figures 4.15, 4.16 and 4.17 present some graphs showing the behaviour of $R_{p,q}$ with respect to the size of the greedy quantization sequence $a^{(n)}$, in a logarithmic scale, for different probability distributions.

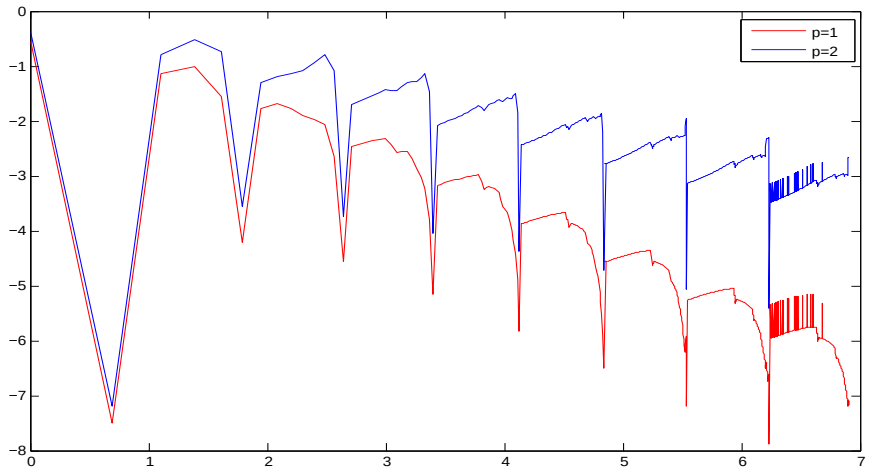

Figure 4.15: The ratios $R_{1,1}$ et $R_{2,1}$ where $a^{(n)}$ is a greedy quantization sequence of the $\mathcal{N}(0,1)$ distribution for $n = 1, \ldots, 1\,000$ (logarithmic scale).

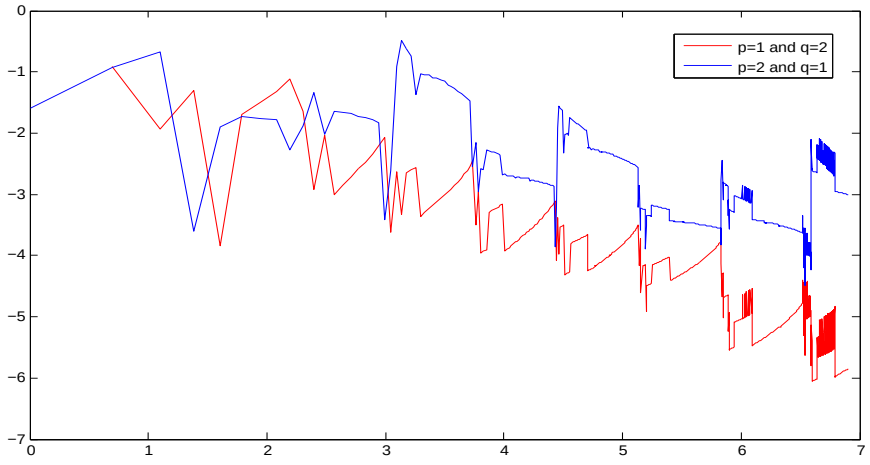

Figure 4.16: The ratios $R_{1,2}$ et $R_{2,1}$ where $a^{(n)}$ is a greedy quantization sequence of the $\mathcal{E}(1)$ distribution for $n = 1, \ldots, 1\,000$ (logarithmic scale).

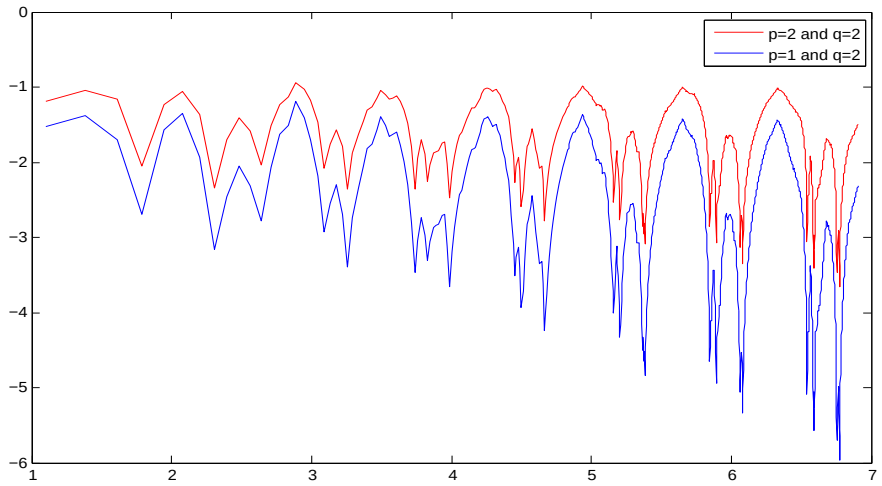

Figure 4.17: The ratios $R_{1,2}$ et $R_{2,2}$ where $a^{(n)}$ is a greedy quantization sequence of the $\mathcal{U}([0,1])$ distribution for $n = 1, \ldots, 1\,000$ (logarithmic scale).

After comparing the various observations presented so far, we wish to put a general definition of quasi-stationarity satisfied by the majority of the probability distributions. Therefore, we are now interested in the behaviour of the ratio

$$R_{p,\rho} = \frac{\|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\|_p}{\|X - \widehat{X}^{a^{(n)}}\|_{1+\rho}^{1+\rho}} \tag{4.24}$$

for $p \in \{1; 2\}$ and $\rho \in [0,1]$, to check whether or not

$$\|\mathbb{E}(X|\hat{X}^{a^{(N)}}) - \hat{X}^{a^{(N)}}\|_p = o(\|X - \hat{X}^{a^{(N)}}\|_{1+\rho}^{1+\rho}). \tag{4.25}$$

We have already seen that, for $\rho = 0$, the ratio converges to 0 when the number of points n increases (see the third numerical observation), while for $\rho = 1$, we get the contrary (see the second numerical observation). Consequently, one wonders if there exists a *limit value* $\rho_l \in]0,1[$, such that, for $\rho \leq \rho_l$, $R_{p,\rho}$ satisfies the requested criteria, and for $\rho > \rho_l$, it does not.

The convergence of the ratio $R_{p,\rho}$ to 0 will cause improvements in the quantization-based numerical integration, in this case, for ρ-Hölder functions. In fact, if $\rho \in [0,1]$ and f is a continuous function with ρ-Hölder gradient with Hölder coefficient $[\nabla f]_\rho$, one has

$$\begin{aligned} f(X) - f(\widehat{X}^{a^{(n)}}) \leq & (\nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}) \\ & + \int_0^1 \left(\nabla f\left(\widehat{X}^{a^{(n)}} + t(X - \widehat{X}^{a^{(n)}})\right) - \nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}\right) dt. \end{aligned}$$

Taking the expectation yields

$$\begin{aligned} \mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}}) \leq & \mathbb{E}(\nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}) \\ & + \mathbb{E}\left[\int_0^1 \left(\nabla f\left(\widehat{X}^{a^{(n)}} + t(X - \widehat{X}^{a^{(n)}})\right) - \nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}\right) dt\right]. \end{aligned}$$

	$\mathcal{N}(0,1)$	$\mathcal{U}([0,1])$	$\mathcal{E}(1)$
$p=1$	$\rho_l = 0.92$	$\rho_l = \frac{3}{4}$	$\rho_l = \frac{2}{3}$
$p=2$	$\rho_l = 0.47$	$\rho_l = \frac{3}{8}$	$\rho_l = \frac{1}{3}$

Table 4.1: Optimal values ρ_l for which different probability distributions satisfy the ρ-quasi-stationarity criterion for $p \in \{1; 2\}$.

At this stage, one notices that

$$\int_0^1 \left(\nabla f\left(\widehat{X}^{a^{(n)}} + t(X - \widehat{X}^{a^{(n)}})\right) - \nabla f(\widehat{X}^{a^{(n)}})|X - \widehat{X}^{a^{(n)}}\right) dt \leq [\nabla f]_\rho \mathbb{E}|X - \widehat{X}^{a^{(n)}}|^{1+\rho} \int_0^1 t^{1+\rho} dt,$$

and uses (4.15) to obtain

$$|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})| \leq \|\nabla f(\widehat{X}^{a^{(n)}})\|_2 \|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\|_2 + \frac{1}{1+\rho}[\nabla f]_\rho \|X - \widehat{X}^{a^{(n)}}\|_{1+\rho}^{1+\rho}.$$

Moreover, if ∇f is bounded, then the above equation can be rewritten as

$$|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})| \leq \|\nabla f(\widehat{X}^{a^{(n)}})\|_\infty \|\mathbb{E}(X|\widehat{X}^{a^{(n)}}) - \widehat{X}^{a^{(n)}}\|_1 + \frac{1}{1+\rho}[\nabla f]_\rho \|X - \widehat{X}^{a^{(n)}}\|_{1+\rho}^{1+\rho}.$$

Hence, if (4.25) is satisfied, then, in both cases, on can conclude with a new upper bound to the error induced by the approximation of $\mathbb{E}[f(X)]$ by $\mathbb{E}[f(\widehat{X}^{a^{(n)}})]$ given by

$$\limsup_n \frac{|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{a^{(n)}})|}{\|X - \widehat{X}^{a^{(n)}}\|_{1+\rho}^{1+\rho}} \leq \frac{1}{1+\rho}[\nabla f]_\rho. \tag{4.26}$$

We study numerically the behavior of the $R_{p,\rho}$ defined by (4.24) and hope to observe a convergence towards 0, at least for certain values of ρ. The conducted experiments yield different results depending on the underlying probability distribution. Let us give some details: For the Gaussian distribution, $R_{p,\rho}$ converges to 0 for the optimal sub-sequences $a^{(2^k-1)}, k \in \mathbb{N}^*$, seen in Section 3.6.1 of the previous Chapter 3, for certain values of ρ. In the case of the Uniform distribution, we observe a convergence of $R_{p,\rho}$ for the optimal sub-sequences, given in Section 3.6.1 of Chapter 3, up to a particular ρ depending on whether $p = 1$ or $p = 2$. The ratio $R_{p,\rho}$ remains bounded for the whole greedy sequence for $\rho < 0.1$. Finally, the convergence is not very clear in the case of the exponential distribution, this can be explained by the fact that we did not find sub-optimal sequences. However, $R_{p,\rho}$ remains bounded for certain values of ρ.

These particular values are exposed in the *two*-entries Table 4.1. Moreover, Figure 4.18 represents the convergence of $R_{1,\frac{1}{4}}$ for a greedy quantization sequence of $\mathcal{N}(0,1)$ of size $n = 400$ ($\rho = \frac{1}{4} < \rho_l = 0.92$). Figure 4.19 shows the divergence of $R_{2,\frac{2}{3}}$ for a greedy sequence of $\mathcal{U}([0,1])$ of size $n = 1\,000$. Finally, we observe consistent results with Table 4.1 in Figure 4.20 where we illustrate the behaviour of $R_{1,\frac{1}{2}}$ for a greedy sequence of $\mathcal{E}(1)$ of size $n = 1\,000$.

These observations allow us to propose the following definition of a ρ-quasi stationary sequence, that is satisfied by greedy quantization sequences for certain values ρ_l given in Table 4.1.

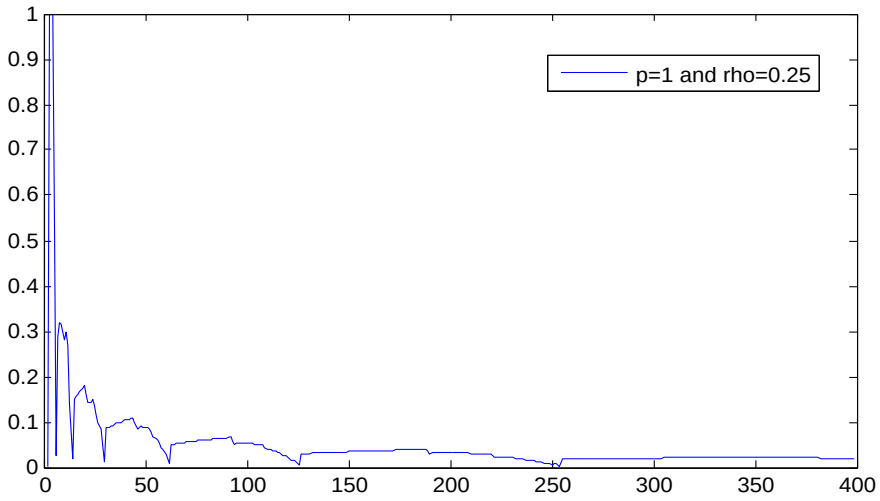

Figure 4.18: The behavior of $R_{1,\frac{1}{4}}$ for a greedy quantization sequence of the Gaussian distribution $\mathcal{N}(0,1)$ of size $n = 400$.

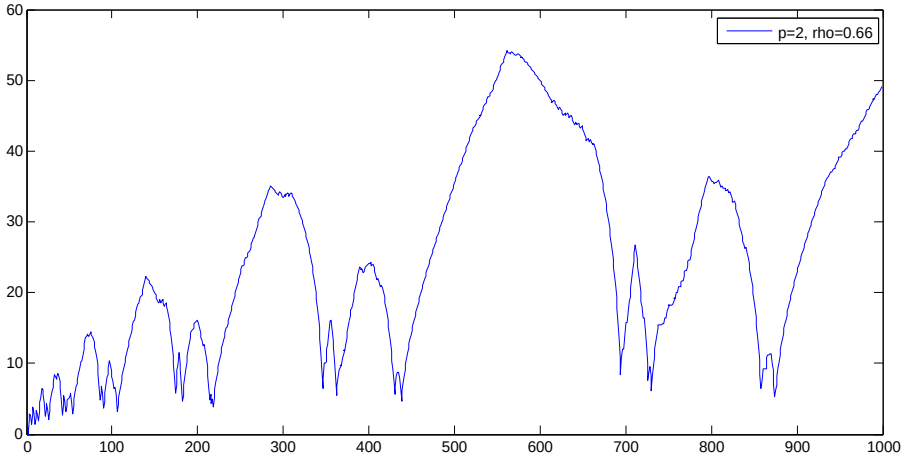

Figure 4.19: The behavior of $R_{2,\frac{2}{3}}$ for a greedy quantization sequence of the Uniform distribution $\mathcal{U}([0,1])$ of size $n = 1\,000$.

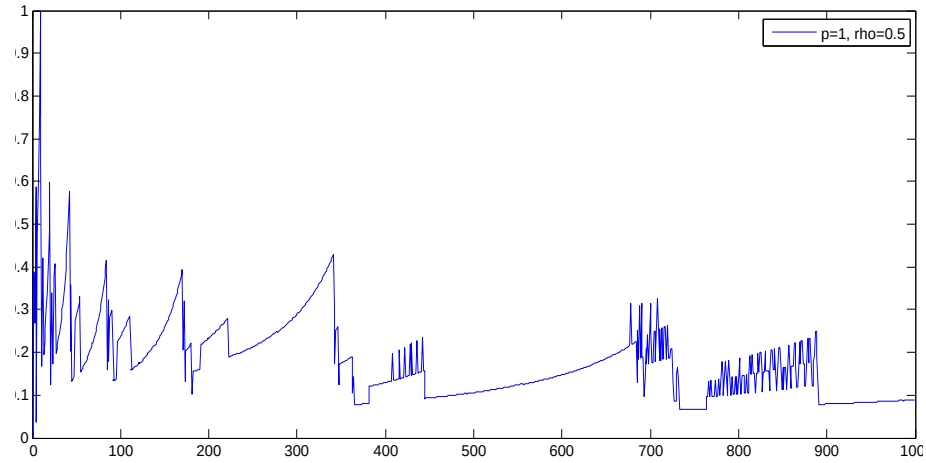

Figure 4.20: The behavior of $R_{1,\frac{1}{2}}$ for a greedy quantization sequence of the exponential distribution $\mathcal{E}(1)$ of size $n = 1\,000$.

Definition 4.5.5. *Let $p \in \{1; 2\}$, $\rho \in [0, 1]$. A greedy quantization sequence $a^{(n)}$ of a random variable X is ρ-quasi-stationary if*

$$\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_p = o(\|\widehat{X}^{a^{(n)}} - X\|_{1+\rho}^{1+\rho}),$$

or, in other words,

$$\frac{\|\widehat{X}^{a^{(n)}} - \mathbb{E}(X|\widehat{X}^{a^{(n)}})\|_p}{\|\widehat{X}^{a^{(n)}} - X\|_{1+\rho}^{1+\rho}} \underset{n\to+\infty}{\longrightarrow} 0.$$

Remark 4.5.6. *(a) Definition 4.5.5 can clearly be extended to greedy quantization sub-sequences.*
(b) Although it is interesting to find an optimal value of ρ common to all the distributions, numerical experiments show that this would not be possible.
(c) The particular case of ρ-Hölder functions is not very practical since this class of functions is not very frequent. Nevertheless, the positive numerical results obtained are interesting and should not be overlooked.

4.6 Construction of sequences with minimal L^*-discrepancy

In Section 3.6.4 of Chapter 3 and Section 4.3 of this chapter, we studied a relation between greedy quantization sequences and low discrepancy sequences such as Van der Corput or Niederreiter sequences. These sequences are known to have a low discrepancy because they present an $\mathcal{O}(\frac{(\log(n))^d}{n})$-rate of convergence of their star discrepancy D_n^* defined by (3.14) as the L^∞-norm of the Uniform distribution of a sequence $\Xi = (\xi_i)_{1\le i\le N}$ of size N on the unit cube $[0,1]^d$. By replacing the L^∞-norm by the L^2-norm, one obtains another modulus, known as the L^*-discrepancy at the origin defined by

$$L_N^*(\Xi) = \left[\int_{[0,1]^d} \left(\frac{A(E,\Xi)}{N} - \lambda_d(E)\right)^2 du\right]^{\frac{1}{2}} = \left[\int_{[0,1]^d} \left(\frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\xi_i\le u} - u_1 \dots u_d\right)^2 du\right]^{\frac{1}{2}},$$

where $A(E,\Xi) = \text{card}\{i;\, \xi_i \in E\}$, $E = \prod_{i=1}^{d}[0,u_i)$ and $u = (u_1,\ldots,u_d)$.

In this section, we are interested in searching for sequences with minimal L^*-discrepancy. This study is motivated by the fact that the manipulation of the L^*-discrepancy is somehow simple due to its definition. We start by finding a one point sequence for any dimension d, then a two points sequence when $d=2$ and finally switch the study to a sequence of size larger than 2.

4.6.1 One point-sequence with minimal L_1^* discrepancy for $d \geq 1$

In this paragraph, the goal is to find a sequence consisting of a single point in any dimension d that admits the lowest L_1^*-discrepancy. Clearly, a minimization problem is a problem of resolving derivatives. Let us denote $\Xi = (\xi_1,\ldots,\xi_d)$ the point we are looking for and start by studying the continuity of L_N^*.

Proposition 4.6.1. *If $d=1$, the discrepancy L_N^* at the origin is a continuous function.*

Proof. We consider two sequences $\Xi = (\xi_k)_{k\geq 1}$ and $\Xi' = (\xi_k')_{k\geq 1}$ of the same size N.

$$\begin{aligned}
|L_N^*(\Xi) - L_N^*(\Xi')| &= \left| \left\| \frac{A(E,\xi)}{N} - \lambda_d(E) \right\|_2 - \left\| \frac{A(E,\xi')}{N} - \lambda_d(E) \right\|_2 \right| \\
&\leq \left\| \frac{A(E,\xi)}{N} - \lambda_d(E) - \frac{A(E,\xi')}{N} + \lambda_d(E) \right\|_2 \\
&\leq \frac{1}{N} \left\| \left| \sum_{k=1}^{N} \mathbb{1}_{\xi_k \leq u} - \sum_{k=1}^{N} \mathbb{1}_{\xi_k' \leq u} \right| \right\|_2 \\
&\leq \frac{1}{N} \left\| \sum_{k=1}^{N} \left| \mathbb{1}_{\xi_k \leq u} - \mathbb{1}_{\xi_k' \leq u} \right| \right\|_2 .
\end{aligned}$$

One easily checks that $\mathbb{1}_{\xi_k \leq u} - \mathbb{1}_{\xi_k' \leq u} = \mathbb{1}_{\xi_k \leq u \leq \xi_k'}$ so that

$$\begin{aligned}
|L_N^*(\Xi) - L_N^*(\Xi')| &\leq \frac{1}{N} \left\| \sum_{k=1}^{N} \mathbb{1}_{\xi_k \leq u \leq \xi_k'} \right\|_2 \\
&\leq \frac{1}{N} \left(\int_{[0,1]^d} \left(\sum_{k=1}^{N} \mathbb{1}_{\xi_k \leq u \leq \xi_k'} \right)^2 du \right)^{\frac{1}{2}} \\
&\leq \frac{1}{N} \left(\int_{[0,1]^d} \left(\sum_{k=1}^{N} \mathbb{1}_{[\xi_k,\xi_k']}(u)^2 + 2 \sum_{j,k=1}^{N} \mathbb{1}_{[\xi_k,\xi_k']}(u)\, \mathbb{1}_{[\xi_j,\xi_j']}(u) \right) du \right)^{\frac{1}{2}} .
\end{aligned}$$

Noticing that $\mathbb{1}_{[\xi_k,\xi_k']}(u)\, \mathbb{1}_{[\xi_j,\xi_j']}(u) = \mathbb{1}_{[\max(\xi_k,\xi_j),\min(\xi_k',\xi_j')]}(u)$, one obtains

$$|L_N^*(\Xi) - L_N^*(\Xi')| \leq \frac{1}{N} \left(\int_{[0,1]^d} \left(\sum_{k=1}^{N} \mathbb{1}_{[\xi_k,\xi_k']}(u) + 2 \sum_{j,k=1}^{N} \mathbb{1}_{[\max(\xi_k,\xi_j),\min(\xi_k',\xi_j')]}(u) \right) du \right)^{\frac{1}{2}} .$$

At this stage, one assumes that there exists $u \in [0,1]^d$ for which the previous indicator functions are equal to 1 and deduces that

$$|L_N^*(\Xi) - L_N^*(\Xi^{'})| \leq \frac{1}{N}\left(\sum_{k=1}^{N} |\xi_k - \xi_k^{'}| + 2|\xi_k - \xi_k^{'}|du\right)^{\frac{1}{2}} \leq \frac{\sqrt{3}}{N}\left\|\Xi - \Xi^{'}\right\|_1^{\frac{1}{2}}.$$

Consequently, the L_N^*-discrepancy at the origin is a $\frac{1}{2}$-Hölder function and, thus, a continuous function. □

Now that we know that the function is continuous, we proceed to its minimzation. First, we denote L_1^2 the square of the L_1^*-discrepancy at the origin of a 1 point sequence $\xi = (\xi_1, \ldots, \xi_d)$ and we write it explicitly.

$$\begin{aligned} L_1^2(\xi) &= \int_{[0,1]^d} |\mathbb{1}_{\xi \leq u_1} \ldots \mathbb{1}_{\xi_d \leq u_d} - u_1 \ldots u_d|^2 \, du_1 \ldots du_d \\ &= \prod_{i=1}^{d}(1-\xi_i) - 2\prod_{i=1}^{n}\int_{\xi_i}^{1} u_i du_i + \left(\int_0^1 u_i^2 du_1\right)^2 \\ &= \prod_{i=1}^{d}(1-\xi_i) - 2\prod_{i=1}^{n}\frac{1-\xi_i^2}{2} + \frac{1}{3^d}. \end{aligned}$$

▷ When $d = 1$, the square of the discrepancy is given by

$$L_1^2(\xi) = \xi^2 - \xi + \frac{1}{3}$$

and admits a minimum at $\xi = \frac{1}{2}$. The L^*-discrepancy at this point is equal to $\sqrt{\frac{1}{12}}$.

▷ When $d = 2$, the square of the discrepancy is given by

$$L_1^2(\xi_1, \xi_2) = (1-\xi_1)(1-\xi_2)\left(1 - \frac{(1+\xi_1)(1+\xi_2)}{2}\right) + \frac{1}{9}.$$

Its derivative with respect to ξ_1 is given by

$$\frac{\partial}{\partial \xi_1} L_1^2(\xi_1, \xi_2) = -(1-\xi_2)\left(1 - \xi_1(1+\xi_2)\right)$$

and is equal to zero if $\xi_2 = 1$ or $\xi_1(1+\xi_2) = 1$. In symmetry, the derivative with respect to ξ_2 is equal to 0 for $\xi_1 = 0$ or $\xi_2(1+\xi_1) = 1$. So, on denotes $\xi = \xi_1 = \xi_2$ and concludes with the following condition

$$\xi(1+\xi) = 1 \quad \Leftrightarrow \quad \xi^2 + \xi - 1 = 0$$

that is satisfied for $\xi = \dfrac{-1+\sqrt{5}}{2}$.

Consequently, the L_1^*-discrepancy reaches its lowest value at $\xi = \left(\dfrac{-1+\sqrt{5}}{2}, \dfrac{-1+\sqrt{5}}{2}\right)$ and is equal to $\sqrt{\frac{103-45\sqrt{5}}{36}}$.

▷ When $d > 2$, the derivative of L_1^2 with respect to ξ_i, $i \in \{1, \dots, d\}$ is given by

$$\frac{\partial}{\partial \xi_i} = -\prod_{j \neq i}^{d}(1-\xi_j) + 2\xi_i \left(\prod_{j \neq i}^{d} \frac{1-\xi_j^2}{2} \right) = \prod_{j \neq i}^{d}(1-\xi_j) \left(1 - 2\xi_i \left(\prod_{j \neq i}^{d} \frac{1+\xi_j}{2} \right) \right).$$

This derivative is equal to 0 if $\xi_j = 1$ or $2\,\xi_i \prod_{j \neq i}^{d} \frac{1+\xi_j}{2} = 1$. So, the point minimizing the discrepancy is solution to

$$2\xi_i \left(\prod_{j \neq i}^{d} \frac{1+\xi_j}{2} \right) = 1 \quad \Leftrightarrow \quad 2\xi_i \left(\prod_{j=1}^{d} \frac{1+\xi_j}{2} \right) = \frac{1+\xi_i}{2}, \qquad \forall 1 \leq i \leq d. \tag{4.27}$$

Setting $z_i = \frac{1+\xi_i}{2}$, the previous equation takes the form

$$2(2z_i - 1) \prod_{i=1}^{d} z_i = z_i \quad \Leftrightarrow \quad 2(2z_i - 1)C = z_i$$

where $C = \prod_{i=1}^{d} z_i$. So, (4.27) becomes $z_i(4C-1) = 2C$ which is equivalent to

$$z_i = \frac{2C}{4C-1}, \quad \forall i \in \{1, \dots, d\}.$$

Then,

$$C = \left(\frac{2C}{4C-1} \right)^d \tag{4.28}$$

which yields

$$\xi_i = 2\left(\frac{2C}{4C-1} \right) - 1 = \frac{1}{4C-1}.$$

It is clear that $C \geq \frac{1}{2}$ and decreases towards $\frac{1}{2}$. Now, if we denote $K = 4C - 1$, we have that $\xi_i = \dfrac{1}{K}$ where K satisfies, by (4.28),

$$K^d \left(\frac{K+1}{4} \right) = \left(\frac{K+1}{2} \right)^d \quad \Leftrightarrow \quad K = \frac{1}{2^{d-2}} \left(1 + \frac{1}{K} \right)^{d-1}. \tag{4.29}$$

At this stage, we try to estimate the value of K or find an explicit form of it. We rely on (4.29) and we proceed as follows: If $d = 1$, it is clear that $K = 2$ and $\xi = \frac{1}{2}$. Otherwise, as soon as d becomes larger than 1, the function $K \longmapsto \left(1 + \frac{1}{K}\right)^{d-1}$ is decreasing from $+\infty$ to 1, which means (4.29) admits a single solution K_d.

Assume that $K_d \leq 1$, then $1 + \dfrac{1}{K_d} \geq 2$ so that $K_d \geq \dfrac{1}{2^{d-2}}\, 2^{d-1} = 2$, which is absurd and, consequently, we can assert that $K_d > 1$. Now, we assume that there exists an extracted sub-sequence $K_{d'}$ such that $K_{d'} \geq 1 + \eta$ with $\eta > 0$. Then,

$$K_{d'} < \frac{1}{2^{d-2}} \left(1 + \frac{1}{1+\eta} \right)^{d-1} = 2 \left(\frac{\rho(\eta)}{2} \right)^{d-1}$$

where $\rho(\eta) = 1 + \dfrac{1}{1+\eta} \in]1;2[$, so that $K_{d'} \to 0$ which constitutes a contradiction. Consequently, K_d converges to 1 when d grows to infinity. Furthermore, one checks that, when $d \to +\infty$, K_d can be written as follows

$$K_d = 1 + \frac{2\,\log 2}{d} + o(\frac{1}{d}). \tag{4.30}$$

And the solution to our problem is given by

$$\xi_d \approx \left(1 - \frac{2\,\log 2}{d}\right) + o(\frac{1}{d}). \tag{4.31}$$

In conclusion, when the dimension d increases, the point with the lowest L^*-discrepancy at the origin converges to 1. A Newton algorithm applied to the function $K \longmapsto K - \frac{1}{2^{d-2}}\left(1 + \frac{1}{K}\right)^{d-1}$ was implemented to find the numerical solution to (4.29) and has given results that are in accordance with the theoretical results obtained in (4.30) and (4.31).

4.6.2 Two points-sequence with minimal L^* discrepancy for $d = 2$

In this section, our aim is to find the 2 points-sequence having the lowest L_2^*-discrepancy when the dimension d is equal to 2. We denote $\Xi = (\xi_1, \xi_2)$ this sequence where, for $i \in \{1, 2\}$, ξ_i is written $(\xi_i^{(1)}, \xi_i^{(2)})$ with $\xi_i^{(1)}$ the abscissa of the point ξ_i and $\xi_i^{(2)}$ its ordinate. We derive the corresponding discrepancy and find the points at which it is equal to 0. The square of the L^* discrepancy of a 2 points sequence for $d = 2$ is denoted by L_2^2 and given by

$$\begin{aligned} L_2^2(\xi_1, \xi_2) &= \int_{[0,1]^2} \left| \frac{1}{2} \sum_{k=1}^{2} \mathbb{1}_{\xi_k^{(i)} \leq u^{(i)}, i=1,2} - u^{(1)}u^{(2)} \right|^2 du^{(1)}du^{(2)} \\ &= \frac{1}{4} \sum_{k,l=1}^{2} \left(1 - \xi_k^{(1)} \vee \xi_l^{(1)}\right)\left(1 - \xi_k^{(2)} \vee \xi_l^{(2)}\right) - \frac{1}{4} \sum_{k=1}^{2} \left(1 - \left(\xi_k^{(1)}\right)^2\right)\left(1 - \left(\xi_k^{(2)}\right)^2\right) + \frac{1}{9} \end{aligned} \tag{4.32}$$

We consider several cases depending on the position of the 2 points in the square $[0, 1]^2$.

Case 1: The points are on the first bisector ($\xi_1^{(1)} = \xi_1^{(2)}$ and $\xi_2^{(1)} = \xi_2^{(2)}$)

In this case, we denote $\xi_1 := \xi_1^{(1)} = \xi_1^{(2)}$ and $\xi_2 := \xi_2^{(1)} = \xi_2^{(2)}$. The discrepancy is given by

$$L_2^2(\Xi) = \frac{1}{4}\left[(1 - \xi_1)^2 + (1 - \xi_2)^2 + 2(1 - \xi_1 \vee \xi_2)^2 - \left(1 - (\xi_1)^2\right)^2 \left(1 - (\xi_2)^2\right)^2\right] + \frac{1}{9}. \tag{4.33}$$

Deriving this quantity with respect to each of the components yields

$$\frac{\partial L_2^2}{\partial \xi_1} = (1 - \xi_1)\left[-\frac{1}{2} + \xi_1(1 + \xi_1) - \mathbb{1}_{\xi_1 > \xi_2}\right],$$

$$\frac{\partial L_2^2}{\partial \xi_2} = (1 - \xi_2)\left[-\frac{1}{2} + \xi_2(1 + \xi_2) - \mathbb{1}_{\xi_2 > \xi_1}\right].$$

At this stage, we assume $\xi_1 > \xi_2$ (the second possibility is the same, we simply change the indices) and that ξ_1 and ξ_2 are different than 0 and 1 so that they are inside the square $[0,1]^2$. Hence, the problem is reduced to finding the solution of the following system

$$\begin{cases} \xi_1^2 + \xi_1 - \frac{3}{2} = 0 \\ \xi_2^2 + \xi_2 - \frac{1}{2} = 0. \end{cases}$$

It is clear that the eligible solution is

$$\xi_1 = \frac{-1+\sqrt{7}}{2} \qquad \text{and} \qquad \xi_2 = \frac{-1+\sqrt{3}}{2}$$

and the corresponding discrepancy is approximately equal to 0.147.

Case 2: $\xi_1^{(1)} \neq \xi_1^{(2)}$ and $\xi_2^{(1)} \neq \xi_2^{(2)}$

The square of the discrepancy is given by

$$\begin{aligned} L_2^2(\Xi) =& \frac{1}{9} + \frac{1}{4}\left[\left(1-\xi_1^{(1)}\right)\left(1-\xi_1^{(2)}\right) + 2\left(1-\xi_1^{(1)} \vee \xi_2^{(1)}\right)\left(1-\xi_1^{(2)} \vee \xi_2^{(2)}\right) + \left(1-\xi_2^{(1)}\right)\left(1-\xi_2^{(2)}\right)\right. \\ & \left. - \left(1-\left(\xi_1^{(1)}\right)^2\right)\left(1-\left(\xi_1^{(2)}\right)^2\right) - \left(1-\left(\xi_2^{(1)}\right)^2\right)\left(1-\left(\xi_2^{(2)}\right)^2\right)\right] \end{aligned}$$

The partial derivatives with respect to each of the 4 components are as follows

$$\begin{aligned} \frac{\partial L_2^2}{\partial \xi_1^{(1)}} &= \frac{1}{4}\left[-\left(1-\xi_1^{(2)}\right) - 2\left(1-\xi_1^{(2)} \vee \xi_2^{(2)}\right)\mathbb{1}_{\xi_1^{(1)} > \xi_2^{(1)}} + 2\xi_1^{(1)}\left(1-\left(\xi_1^{(2)}\right)^2\right)\right], \\ \frac{\partial L_2^2}{\partial \xi_1^{(2)}} &= \frac{1}{4}\left[-\left(1-\xi_1^{(1)}\right) - 2\left(1-\xi_1^{(1)} \vee \xi_2^{(1)}\right)\mathbb{1}_{\xi_1^{(2)} > \xi_2^{(2)}} + 2\xi_1^{(2)}\left(1-\left(\xi_1^{(1)}\right)^2\right)\right], \\ \frac{\partial L_2^2}{\partial \xi_2^{(1)}} &= \frac{1}{4}\left[-\left(1-\xi_2^{(2)}\right) - 2\left(1-\xi_1^{(2)} \vee \xi_2^{(2)}\right)\mathbb{1}_{\xi_2^{(1)} > \xi_1^{(1)}} + 2\xi_2^{(1)}\left(1-\left(\xi_2^{(2)}\right)^2\right)\right], \\ \frac{\partial L_2^2}{\partial \xi_2^{(2)}} &= \frac{1}{4}\left[-\left(1-\xi_2^{(1)}\right) - 2\left(1-\xi_1^{(1)} \vee \xi_2^{(1)}\right)\mathbb{1}_{\xi_2^{(2)} > \xi_1^{(2)}} + 2\xi_2^{(2)}\left(1-\left(\xi_2^{(1)}\right)^2\right)\right]. \end{aligned}$$

We consider several sub-cases to study all the possibilities induced by Case 2. In all these situations, we assume that the coordinates are in $]0,1[$.

• **If $\xi_1^{(1)} > \xi_2^{(1)}$ and $\xi_2^{(2)} > \xi_1^{(2)}$:** The problem is finding the solutions of

$$-\left(1-\xi_1^{(2)}\right) - 2\left(1-\xi_2^{(2)}\right) + 2\xi_1^{(1)}\left(1-\left(\xi_1^{(2)}\right)^2\right) = 0 \tag{4.34}$$

$$-\left(1-\xi_1^{(1)}\right) + 2\xi_1^{(2)}\left(1-\left(\xi_1^{(1)}\right)^2\right) = 0 \tag{4.35}$$

$$-\left(1-\xi_2^{(2)}\right) + 2\xi_2^{(1)}\left(1-\left(\xi_2^{(2)}\right)^2\right) = 0 \tag{4.36}$$

$$-\left(1-\xi_2^{(1)}\right) - 2\left(1-\xi_1^{(1)}\right) + 2\xi_2^{(2)}\left(1-\left(\xi_2^{(1)}\right)^2\right) = 0 \tag{4.37}$$

To simplify the notations in the following, we denote

$$a = \xi_1^{(1)} \quad b = \xi_1^{(2)} \quad c = \xi_2^{(1)} \quad d = \xi_2^{(2)}.$$

(4.35) yields $b = \frac{1}{2(1+a)}$ and (4.36) gives $c = \frac{1}{2(1+d)}$. We merge both equality in (4.34) and (4.37) respectively to obtain

$$-3 + 2d + 2a + \frac{1}{2(1+a)^2} = 0. \tag{4.38}$$

and

$$-3 + 2a + 2d + \frac{1}{2(1+d)^2} = 0. \tag{4.39}$$

This yields $a = d$ so $b = c$ as well. Now, (4.38) gives $a = d = \frac{\sqrt{2}}{2}$ and (4.35) gives $b = c = 1 - \frac{\sqrt{2}}{2}$. Consequently, the eligible solution is

$$\xi_1 = \left(\frac{\sqrt{2}}{2}; 1 - \frac{\sqrt{2}}{2}\right) \qquad \text{and} \qquad \xi_2 = \left(1 - \frac{\sqrt{2}}{2}; \frac{\sqrt{2}}{2}\right), \tag{4.40}$$

and the L_2^*-discrepancy is approximately equal to 0.1703.

• **If $\xi_1^{(1)} < \xi_2^{(1)}$ and $\xi_2^{(2)} > \xi_1^{(2)}$:** The goal is to solve the following system

$$-\left(1 - \xi_1^{(2)}\right) + 2\xi_1^{(1)}\left(1 - \left(\xi_1^{(2)}\right)^2\right) = 0 \tag{4.41}$$

$$-\left(1 - \xi_1^{(1)}\right) + 2\xi_1^{(2)}\left(1 - \left(\xi_1^{(1)}\right)^2\right) = 0 \tag{4.42}$$

$$-3\left(1 - \xi_2^{(2)}\right) + 2\xi_2^{(1)}\left(1 - \left(\xi_2^{(2)}\right)^2\right) = 0 \tag{4.43}$$

$$-3\left(1 - \xi_2^{(1)}\right) + 2\xi_2^{(2)}\left(1 - \left(\xi_2^{(1)}\right)^2\right) = 0 \tag{4.44}$$

With the same notations as the previous case, we get, by equation (4.41),

$$a = \frac{1}{2(1+b)} \tag{4.45}$$

and, by equation (4.42),

$$b = \frac{1}{2(1+a)} \tag{4.46}$$

Merging (4.45) with (4.42), one obtains $a = b = \frac{-1+\sqrt{3}}{2}$. The same reasoning with equations (4.43) and (4.44) yields $c = d = \frac{-1+\sqrt{7}}{2}$. Consequently, the 2 points sequence is given by

$$\xi_1 = \left(\frac{-1+\sqrt{7}}{2}, \frac{-1+\sqrt{7}}{2}\right) \qquad \text{and} \qquad \xi_2 = \left(\frac{-1+\sqrt{3}}{2}, \frac{-1+\sqrt{3}}{2}\right). \tag{4.47}$$

and its L_2^*-discrepancy is given by 0.147.

• In all the other situations, we get the same result as in the two detailed situations above.

In conclusion, the 2 points sequence with minimal L_2^*-discrepancy in $[0,1]^2$ is given by (4.47) and its discrepancy is equal to 0.147.

4.6.3 4 points sequence with minimal L^* discrepancy for $d = 2$

In order to find a sequence $\Xi = (\xi_1, \xi_2, \xi_3, \xi_4)$ in $[0,1]^2$ which has the minimal L_4^*-discrepancy, we start by writing the square of this discrepancy as follows

$$L_4^2(\Xi) = \frac{1}{16}\sum_{k,l=1}^{4}\left(1-\xi_k^{(1)}\vee\xi_l^{(1)}\right)\left(1-\xi_k^{(2)}\vee\xi_l^{(2)}\right) - \frac{1}{8}\sum_{k=1}^{4}\left(1-\left(\xi_k^{(1)}\right)^2\right)\left(1-\left(\xi_k^{(2)}\right)^2\right) + \frac{1}{9}. \tag{4.48}$$

The partial derivatives with respect to the 8 components of the discrepancy are given, for every $i \in \{1,2,3,4\}$ by

$$\frac{\partial L_4^2}{\partial \xi_i^{(1)}} = -\frac{2}{4^2}\sum_{k=1}^{4}\left(1-\xi_k^{(2)}\vee\xi_i^{(2)}\right)\mathbb{1}_{\xi_i^{(1)}\geq\xi_k^{(1)}} + \frac{1}{4}\xi_i^{(1)}\left(1-\left(\xi_i^{(2)}\right)^2\right) + \frac{1}{4^2}\left(1-\xi_i^{(2)}\right), \tag{4.49}$$

and

$$\frac{\partial L_4^2}{\partial \xi_i^{(2)}} = -\frac{2}{4^2}\sum_{k=1}^{4}\left(1-\xi_k^{(1)}\vee\xi_i^{(1)}\right)\mathbb{1}_{\xi_i^{(2)}\geq\xi_k^{(2)}} + \frac{1}{4}\xi_i^{(2)}\left(1-\left(\xi_i^{(1)}\right)^2\right) + \frac{1}{4^2}\left(1-\xi_i^{(1)}\right). \tag{4.50}$$

A first intuition is to partition the unit square into 4 sub-squares $(C_i)_{1\leq i\leq 4}$ such that $C_1 = [0, \frac{1}{2}]^2$ and the others are translations of C_1. Then, we predict that, for every, $i \in \{1,2,3,4\}$, $\xi_i \in C_i$. Hence, there is 8 conditions to take into consideration:

$$\begin{array}{llll} \xi_1^{(2)} < \xi_3^{(2)}, & \xi_2^{(2)} < \xi_3^{(2)}, & \xi_1^{(1)} < \xi_2^{(1)}, & \xi_3^{(1)} < \xi_2^{(1)}, \\ \xi_1^{(2)} < \xi_4^{(2)}, & \xi_2^{(2)} < \xi_4^{(2)}, & \xi_1^{(1)} < \xi_4^{(1)}, & \xi_3^{(1)} < \xi_4^{(1)}. \end{array}$$

Under these conditions, the points for which the partial derivatives are equal to 0 satisfy the following system

$$\begin{array}{l}
\left(1-\xi_1^{(2)}\right)\left[-\frac{1}{4}+\xi_1^{(1)}\left(1+\xi_1^{(2)}\right)\right] - \frac{1}{2}\left(1-\xi_3^{(2)}\right)\mathbb{1}_{\xi_1^{(1)}>\xi_3^{(1)}} = 0 \\
\left(1-\xi_2^{(2)}\right)\left[-\frac{1}{4}+\xi_2^{(1)}\left(1+\xi_2^{(2)}\right)\right] - \frac{1}{2}\left(1-\xi_3^{(2)}\right) - \frac{1}{2}\left(1-\xi_2^{(2)}\vee\xi_1^{(2)}\right) - \frac{1}{2}\left(1-\xi_4^{(2)}\right)\mathbb{1}_{\xi_2^{(1)}>\xi_4^{(1)}} = 0 \\
\frac{1}{2} + \mathbb{1}_{\xi_3^{(1)}>\xi_1^{(1)}} - \xi_3^{(1)}\left(1+\xi_3^{(2)}\right) = 0 \\
\left(1-\xi_4^{(2)}\right)\left[-\frac{3}{4}+\xi_4^{(1)}\left(1+\xi_4^{(2)}\right) - \frac{1}{2}\mathbb{1}_{\xi_4^{(1)}>\xi_2^{(1)}}\right] - \frac{1}{2}\left(1-\xi_3^{(2)}\vee\xi_4^{(2)}\right) = 0 \\
\left(1-\xi_1^{(1)}\right)\left[-\frac{1}{4}+\xi_1^{(2)}\left(1+\xi_1^{(1)}\right)\right] - \frac{1}{2}\left(1-\xi_2^{(1)}\right)\mathbb{1}_{\xi_1^{(2)}>\xi_2^{(2)}} = 0 \\
\frac{1}{2} + \mathbb{1}_{\xi_2^{(2)}>\xi_1^{(2)}} - \xi_2^{(2)}\left(1+\xi_2^{(1)}\right) = 0 \\
\left(1-\xi_3^{(1)}\right)\left[-\frac{1}{4}+\xi_3^{(2)}\left(1+\xi_3^{(1)}\right)\right] - \frac{1}{2}\left(1-\xi_2^{(1)}\right) - \frac{1}{2}\left(1-\xi_3^{(1)}\vee\xi_1^{(1)}\right) - \frac{1}{2}\left(1-\xi_4^{(1)}\right)\mathbb{1}_{\xi_3^{(2)}>\xi_4^{(2)}} = 0 \\
\left(1-\xi_4^{(1)}\right)\left[-\frac{3}{4}+\xi_4^{(2)}\left(1+\xi_4^{(1)}\right) - \frac{1}{2}\mathbb{1}_{\xi_4^{(2)}>\xi_3^{(2)}}\right] - \frac{1}{2}\left(1-\xi_2^{(1)}\vee\xi_4^{(1)}\right) = 0
\end{array}$$

Solving this system requires taking a large number of particular cases to try to cover all the possibilities satisfying the 8 conditions taken at the beginning of this study. Among those cases, a few admit no solution while the others are very complicated. Solving this system by direct closed formulae does not yield promising results, that's why one tends to try and find a numerical solution to this problem.

Gradient descent algorithm

A gradient descent algorithm is implemented in order to minimize the L_4^*-discrepancy in the two-dimensional case. The algorithm was tested for different number of points N in the sequence. The derivatives of $L_N^2(\xi_1, \ldots, \xi_N)$ are given, by generalizing the derivatives (4.49) and (4.50) when $N = 4$, as follows

$$\frac{\partial L_N^2}{\partial \xi_i^{(1)}} = -\frac{2}{N^2} \sum_{k=1}^{N} \left(1 - \xi_k^{(2)} \vee \xi_i^{(2)}\right) \mathbb{1}_{\xi_i^{(1)} \geq \xi_k^{(1)}} + \frac{1}{N} \xi_i^{(1)} \left(1 - \left(\xi_i^{(2)}\right)^2\right) + \frac{1}{N^2} \left(1 - \xi_i^{(2)}\right), \quad (4.51)$$

and

$$\frac{\partial L_N^2}{\partial \xi_i^{(2)}} = -\frac{2}{N^2} \sum_{k=1}^{n} \left(1 - \xi_k^{(1)} \vee \xi_i^{(1)}\right) \mathbb{1}_{\xi_i^{(2)} \geq \xi_k^{(2)}} + \frac{1}{N} \xi_i^{(2)} \left(1 - \left(\xi_i^{(1)}\right)^2\right) + \frac{1}{N^2} \left(1 - \xi_i^{(1)}\right). \quad (4.52)$$

We consider a sequence of steps $\gamma_t = 10^{-3} + \frac{10^{-2}}{t} + \left(1 - \frac{1}{t}\right) \frac{1}{t^{\frac{3}{4}}}$ and, after obtaining the sequence, we compute the corresponding discrepancy via the formulas given above. When $N = 1$ or 2, the numerical results are identical to the theoretical results already given. However, when $N = 4$, complexities appear in the numerical procedure (just like in the theoretical procedure). In fact, the algorithm gives us only a local minimum of the discrepancy (depending on the initial sequence starting the algorithm). To find the global minimum, one needs to consider all the different cases, which is illogical, especially that we will never be able to know if we have really reached the global minimum. We present some particular cases.

• If we consider

$$\xi_1^{(1)} < \xi_3^{(1)}, \qquad \xi_2^{(1)} < \xi_4^{(1)}, \qquad \xi_1^{(2)} < \xi_2^{(2)}, \qquad \xi_3^{(2)} < \xi_4^{(2)},$$

the sequence with minimal discrepancy is given by

$$\begin{aligned} \xi_1 &= \left(\tfrac{-1+\sqrt{2}}{2}, \tfrac{-1+\sqrt{2}}{2}\right), & \xi_3 &= (0,7,\ 0,44), \\ \xi_2 &= \left(\tfrac{-1}{2} + \sqrt{2}, \tfrac{-1}{2} + \sqrt{2}\right), & \xi_4 &= (0,44,\ 0,7), \end{aligned}$$

and its discrepancy is approximately equal to $0,2315$.

• If we consider

$$\xi_3^{(1)} < \xi_1^{(1)}, \qquad \xi_4^{(1)} < \xi_2^{(1)}, \qquad \xi_2^{(2)} < \xi_1^{(2)}, \qquad \xi_4^{(2)} < \xi_3^{(2)},$$

We obtain a minimal discrepancy L_4^2 equal to $0,2435$.

• If we rely on the 1 point-sequence with minimal discrepancy and consider, for starting point, its equivalent in each sub-square C_i, i.e. the equivalent of $(1 - \log(2),\ 1 - \log(2))$, which is

$$(0,3,\ 0,3) \qquad (0,8,\ 0,3) \qquad (0,3,\ 0,8) \qquad (0,8,\ 0,8),$$

the minimal discrepancy is approximately equal to $0,2405$.

Remark 4.6.2. *Minimization algorithms without derivatives have been implemented to try to minimize the discrepancy, we can name Generalised Pattern Search, Nelder Mead Simplex, the Coordiante Search and others. However, these algorithms are not effective because even if one gets a result when the algorithm stops, it is not necessarily the global minimum desired.*

In conclusion, taking every possible case to find the sequence with the minimal discrepancy is not really interesting. Thus, constructing such a 4-point sequence is almost impossible. Similarly, it would be more difficult to find sequences with more than 4 points because of the additional conditions one has to take.

Chapter 5

L^s-rate optimality of dilated/contracted L^r-optimal and greedy quantization sequences

Abstract We investigate some L^s-rate optimality properties of dilated/contracted L^r-optimal quantizers and L^r-greedy quantization sequences $(\alpha^n)_{n\geq 1}$ of a random variable X. We establish, for different values of s, L^s-rate optimality results for L^r-optimally dilated/contracted greedy quantization sequences $(\alpha^n_{\theta,\mu})_{n\geq 1}$ defined by $\alpha^n_{\theta,\mu} = \{\mu+\theta(\alpha_i-\mu), \alpha_i \in \alpha^{(n)}\}$. We lead a specific study for L^r-optimal greedy quantization sequences of radial density distributions and show that they are L^s-rate optimal for $s \in (r, r+d)$ under some moment assumption. Based on the results established in [71] for L^r-optimal quantizers, we show, for a larger class of distributions, that the dilatation $(\alpha^n_{\theta,\mu})_{n\geq 1}$ of an L^r-optimal quantizer is L^s-rate optimal for $s < r+d$. We show, for various probability distributions, that there exists a parameter θ^* for which the dilated quantization sequence satisfy the so-called L^s-*empirical measure* Theorem and present an application of this approach to numerical integration.

5.1 Introduction

The aim of this chapter is, on the one hand, to extend some "robustness" results of optimal quantizers to a much wider class of distributions and, on the other hand, to establish similar results for greedy quantization sequences introduced in [45] and developed in [24]. Let $L^r_{\mathbb{R}^d}(\mathbb{P})$ (or simply $L^r(\mathbb{P})$), $r \in (0, +\infty)$, denote the set of d-dimensional random vectors X defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with distribution $P = \mathbb{P}_X$ and such that $\mathbb{E}|X|^r < +\infty$ (for any norm $|\cdot|$ on $\mathbb{R}^d$). Optimal vector quantization consists in finding the best approximation of a multidimensional random vector X by a random variable Y taking at most a finite number n of values. Consider $\Gamma = \{x_1, \ldots, x_n\}$ a d-dimensional grid of size n. The principle is to approximate X by $\pi_\Gamma(X)$ where $\pi_\Gamma : \mathbb{R}^d \to \Gamma$ is a nearest neighbor projection defined by

$$\pi_\Gamma(\xi) = \sum_{i=1}^{n} x_i \mathbb{1}_{W_i(\Gamma)}(\xi)$$

where $(W_i(\Gamma))_{1\leq i\leq n}$ is a so-called *Voronoï partition* of $\mathbb{R}^d$ induced by Γ i.e. a Borel partition satisfying

$$W_i(\Gamma) \subset \{\xi \in \mathbb{R}^d : |\xi - x_i| \leq \min_{j\neq i} |\xi - x_j|\}, \qquad i = 1, \ldots, n. \tag{5.1}$$

Then,

$$\widehat{X}^\Gamma = \pi_\Gamma(X) := \sum_{i=1}^n x_i \mathbb{1}_{W_i(\Gamma)}(X) \tag{5.2}$$

is called the *Voronoï quantization* of X. The L^r-quantization error induced when replacing X by its quantization $\widehat{X}^\Gamma$ is naturally defined by

$$e_r(\Gamma, X) = \|X - \pi_\Gamma(X)\|_r = \|X - \widehat{X}^\Gamma\|_r = \left\| \min_{1\leq i\leq n} |X - x_i| \right\|_r \tag{5.3}$$

where $\|.\|_r$ denotes the $L^r(\mathbb{P})$-norm (or quasi-norm if $0 < r < 1$). Consequently, the optimal quantization problem at *level* n boils down to finding the grid Γ^n of size n that minimizes this error, i.e.

$$e_{r,n}(X) \;=\; \inf_{\Gamma,\, \mathrm{card}(\Gamma)\leq n} e_r(\Gamma, X). \tag{5.4}$$

where $\mathrm{card}(\Gamma)$ denotes the cardinality of Γ. The existence of a solution to this problem and the convergence of $e_{r,n}(X)$ to 0 at an $\mathcal{O}(n^{-\frac{1}{d}})$-rate of convergence when the level (or size) n goes to $+\infty$ have been shown (see [32, 56, 57] for example). The convergence to 0 of such an error induced by a sequence $(\Gamma^n)_{n\geq 1}$ of L^r-optimal quantizers of (the distribution of) X is an easy consequence of the separability of $\mathbb{R}^d$. Its rate of convergence to 0 is a much more challenging problem that has been solved in several steps over between 1950's and the early 2000's and the main results in their final form are summed up in Section 5.2.

However, numerical implementation of multidimensional L^r-optimal quantizers requires to optimize grids of size $n \times d$ which becomes computationally too costly when n or d increase. So, a greedy version of optimal vector quantization (which is easier to handle) has been introduced in [45] as a sub-optimal solution to the quantization problem. It consists in building a *sequence* of points $(a_n)_{n\geq 1}$ in $\mathbb{R}^d$ which is recursively L^r-optimized level by level, in the sense that it minimizes the L^r-quantization error at each iteration in a greedy way. This means that, having the first n points $a^{(n)} = \{a_1, \ldots, a_n\}$ for $n \geq 1$, we add, at the $(n+1)$-th step, the point a_{n+1} solution to

$$a_{n+1} \in \mathrm{argmin}_{\xi\in\mathbb{R}^d}\, e_r(a^{(n)} \cup \{\xi\}, X), \tag{5.5}$$

noting that $a^{(0)} = \varnothing$, so that a_1 *is simply an/the* L^r*-median of the distribution* P *of* X. The sequence $(a_n)_{n\geq 1}$ is called an L^r-optimal greedy quantization sequence for X or its distribution P. It is proved in [45] that the problem (5.5) admits, as soon as X lies in $L_{\mathbb{R}^d}(\mathbb{P})$, a solution $(a_n)_{n\geq 1}$ which may be not unique due to the dependence of greedy quantization on the symmetry of the distribution P. The corresponding L^r-quantization error $e_r(a^{(n)}, X)$ is decreasing w.r.t n and converges to 0 when n goes to $+\infty$. Greedy quantization sequences have an optimal convergence rate to 0 compared to optimal quantizers, in the sense that the grids $\{a_1, \ldots, a_n\}$ are L^r-rate optimal, i.e. the corresponding quantization error converges with an $\mathcal{O}(n^{-\frac{1}{d}})$-rate of convergence. This was established first in [45] for a rather wide family of absolutely continuous distribution using some maximal functions approximating the density f of P. Then, it has been extended in [24] to a much larger class of probability density functions where the authors relied

on an exogenous auxiliary probability distribution ν on $(\mathbb{R}^d, \mathcal{B}or(\mathbb{R}^d))$ satisfying a certain control on balls, the result is recalled in Section 5.2.

A very important field of applications is quantization-based numerical integration where we approximate an expectation $\mathbb{E}h(X)$ of a function h on $\mathbb{R}^d$ by some cubature formulas. The error bounds induced by such numerical schemes always involve the L^s-quantization error induced by the approximation of X by its (optimal or greedy) quantization usually with $s \geq r$. This problem also appears when we use optimal quantization as a space discretization scheme of ARCH models, namely the Euler scheme of a diffusion devised to solve stochastic control, optimal stopping or filtering problems (see [59, 60] for example) where, in order to estimate the upper error bounds induced by such approximation schemes, one needs to evaluate L^s-quantization errors induced by L^r-optimal (or asymptotically optimal) quantizers for $s \geq r$. So, one needs to see whether such quantizers sharing L^r-optimality properties preserve their performances in L^s, this is called the *distortion mismatch* problem and was deeply studied in [33] for sequences of optimal quantizers. As for greedy quantization sequences, it was first investigated in [45] and extended later in [24] as already mentioned.

Another approach to this problem was considered in [71] where the author was interested in the fact that an appropriate dilatation or contraction of a (sequence of) L^r-optimal quantizer(s) $(\Gamma^n)_{n\geq 1}$ remains L^s-rate optimal. This study was also motivated by its application to the algorithms of designing L^s-optimal quantizers for $s \neq 2$. In fact, several stochastic procedures, like Lloyd's algorithm or the Competitive Learning Vector Quantization algorithm (CLVQ), are based on the stationarity property satisfied by optimal quadratic quantizers and designed for $s = 2$. However, when $s > 2$, these procedures become unstable and difficult and their convergence is very dependent on the initialization. So, in order to design L^s-optimal quantizers, $s > 2$, one can use the L^2-dilated quantizers to initialize the algorithms and speed their convergence.

In this chapter, based on the same motivations, we are interested in establishing L^s-rate optimality results of dilatations/contractions of L^r-optimal greedy quantization sequences. Moreover, we extend the original results established for L^r-optimal quantizers in [71] to a larger class of distributions taking advantage of new tools developed in [24] to analyze quantization errors. These tools are based on auxiliary probability distributions with a certain property of control on balls. In other words, if $(\alpha^n)_{n\geq 1}$ is a sequence of L^r-optimal quantizers or an L^r-optimal greedy quantization sequence, then the sequence $(\alpha^n_{\theta,\mu})_{n\geq 1}$ defined, for every $\theta > 0$ and $\mu \in \mathbb{R}^d$, by $\alpha^n_{\theta,\mu} = \{\mu + \theta(a_i - \mu),\ a_i \in \alpha^n\}$, is L^s-rate optimal for $s \neq r$. A lower bound of the L^s-quantization error $e_s(\alpha^n_{\theta,\mu}, P)$ was given in [71] for L^r-optimal quantizers and it also holds for greedy quantization sequences: If $P = f.\lambda_d$, then for every $\theta > 0$, $\mu \in \mathbb{R}^d$ and $n \geq 1$,

$$\liminf_{n\to+\infty} n^{\frac{1}{d}} e_s(\alpha^n_{\theta,\mu}, P) \geq Q^{\mathrm{Inf}}_{r,s}(P, \theta) \tag{5.6}$$

where

$$\begin{aligned} Q^{\mathrm{Inf}}_{r,s}(P,\theta) &= \theta^{1+\frac{d}{s}} \widetilde{J}_{s,d} \left(\int_{\mathbb{R}^d} f^{\frac{d}{d+r}} d\lambda_d \right)^{\frac{1}{d}} \left(\int_{\{f>0\}} f_{\theta,\mu} f^{-\frac{s}{d+r}} d\lambda_d \right)^{\frac{1}{s}} \\ &= \theta \widetilde{J}_{s,d} \left(\int_{\mathbb{R}^d} f^{\frac{d}{d+r}} d\lambda_d \right)^{\frac{1}{d}} \left(\int_{\{f>0\}} f^{-\frac{s}{d+r}} dP_{\theta,\mu} \right)^{\frac{1}{s}} \end{aligned} \tag{5.7}$$

where $\widetilde{J}_{s,d} = \inf_{n\geq 1} n^{\frac{1}{d}} e_{s,n}(U([0,1]^d)) \in (0,+\infty)$ is the constant given in Zador's Theorem (see (5.8)) and $f_{\theta,\mu}$ denotes the function $f_{\theta,\mu}(x) = f(\mu + \theta(x-\mu))$. Likewise, if $X \sim P = f.\lambda_d$, then $P_{\theta,\mu}$ denotes the probability distribution of the random variable $\frac{X-\mu}{\theta}+\mu$ and $dP_{\theta,\mu} = \theta^d f_{\theta,\mu}.d\lambda_d$. Our goal is then to estimate upper bounds of this error. For the L^r-dilated/contracted greedy quantization sequences, we rely on auxiliary probability distributions satisfying a certain control criterion on balls and establish upper estimates depending on the values of s. We obtain Pierce type universal non-asymptotic results of L^s-rate optimality of a greedy quantization sequence $(\alpha^n_{\theta,\mu})_{n\geq 1}$ of a distribution P having finite polynomial moments at any order. On another hand, we lead an interesting study for a particular class of distributions, the radial density probability distributions, showing that the corresponding L^r-greedy quantization sequences are L^s-rate optimal for $s \in (r, d+r)$ under some moment assumption on P and we investigate a particular case, the Hyper-Cauchy distribution, where the distribution P has finite polynomial moments up to a finite order. As for the L^r-dilated/contracted optimal quantizers, two results are already given in [71]: one showing that an asymptotically L^r-optimal sequence of quantizers is L^s-rate optimal and another restricted to a sequence of (exactly) L^r-optimal quantizers and showing that it is L^s-rate optimal for $s \in (0,+\infty)$. In this chapter, we change the approach and use auxiliary probability distributions satisfying a control criterion on balls to extend these results to a larger class of distributions for L^r-optimal quantizers. At this stage, one wonders if the L^r-dilated sequence satisfy the so-called L^s-empirical measure Theorem or if there exists a particular set of parameters (θ^*, μ^*) for which it is satisfied, leading to wonder whether the sequence is L^s-asymptotically optimal. This prompts us to consider several particular probability distributions and establish this study for each distribution. Finally, the application of this study to numerical integration, introduced in [71], is detailed and illustrated, by numerical examples, for optimal and greedy quantization.

This chapter will be organized as follows: We start, in Section 5.2, with some results and tools, mostly from [24], that will be useful in the whole chapter. In Section 5.3, we give upper bounds for dilated/contracted sequences of L^r-greedy quantization sequences of a distribution P having finite polynomial moments at any order, investigate an example of a not so general case and lead a specific study for greedy quantization sequences of radial density distributions. Such error bounds are given for optimal quantizers in Section 5.4. In Section 5.5, we present several studies concerning the convergence of the empirical measure and the L^s-asymptotic optimality of the L^r-dilated/contracted sequence of particular probability distributions. Finally, Section 5.6 is devoted to an application to numerical integration.

5.2 Main tools

In this section, we present some useful results and inequalities which constitute essential tools needed to achieve desired results in the rest of the chapter. Let X be an $\mathbb{R}^d$-valued random variable with distribution P such that $\mathbb{E}|X|^r < +\infty$ for $r > 0$ and a norm $|\cdot|$ on $\mathbb{R}^d$. Let $(\Gamma^n)_{n\geq 0}$ be a sequence of L^r-optimal quantizers of X and $(a_n)_{n\geq 0}$ be a corresponding greedy quantization sequence. We start by giving the result concerning the rate of convergence to 0 of a sequence of L^r-optimal quantizers. The first part of the following theorem is an asymptotic result and the second part is universal non-asymptotic.

Theorem 5.2.1. (*a*) Zador's Theorem (see [75]) *: Let $X \in L^{r+\eta}_{\mathbb{R}^d}(\mathbb{P})$, $\eta > 0$, with distribution P such that $dP(\xi) = \varphi(\xi)d\lambda_d(\xi) + d\nu(\xi)$. Then,*

$$\lim_{n\to+\infty} n^{\frac{1}{d}} e_{r,n}(X) = Q_r(P) = \widetilde{J}_{r,d} \|\varphi\|^{\frac{1}{r}}_{L^{\frac{r}{r+d}}(\lambda_d)} \tag{5.8}$$

where $\widetilde{J}_{r,d} = \inf_{n\geq 1} n^{\frac{1}{d}} e_{r,n}(U([0,1]^d)) \in (0,+\infty)$.
(*b*) Extended Pierce's Lemma (see [44, 57])*: Let $r, \eta > 0$. There exists a constant $\kappa_{d,r,\eta} \in (0,+\infty)$ such that, for any random vector $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^d$,*

$$\forall n \geq 1, \quad e_{r,n}(X) \leq \kappa_{d,r,\eta} \sigma_{r+\eta}(X) n^{-\frac{1}{d}} \tag{5.9}$$

where, for every $r \in (0,+\infty)$, $\sigma_r(X) = \inf_{a\in\mathbb{R}^d} \|X - a\|_r \leq +\infty$.

Note that a sequence of n-quantizers $(\Gamma^n)_{n\geq 1}$ is said to be *asymptotically L^r-optimal* if

$$\lim_n n^{\frac{1}{d}} e_r(\Gamma^n, X) = Q_r(P)$$

and *L^r-rate optimal* if

$$\limsup_{n\to+\infty} n^{\frac{1}{d}} e_r(\Gamma^n, X) < +\infty \quad \text{or equivalently} \quad \forall n \geq 1, \quad e_r(\Gamma^n, X) \leq C_1 n^{-\frac{1}{d}} \tag{5.10}$$

where C_1 is a constant not depending on n.

The L^r-rate optimality of greedy quantization sequences has been recently extended in [24]. The authors relied on auxiliary probability distributions ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ satisfying the following control on balls, with respect to an L^r-median a_1 of P: Assume there exists $\varepsilon_0 \in (0,1]$ such that for every $\varepsilon \in (0, \varepsilon_0)$, there exists a Borel function $g_\varepsilon : \mathbb{R}^d \to [0,+\infty)$ such that, for every $x \in \mathrm{supp}(P)$ and every $t \in [0, \varepsilon|x - a_1|]$,

$$\nu(B(x,t)) \geq g_\varepsilon(x) V_d t^d \tag{5.11}$$

where V_d denotes the volume of the hyper unit ball. Of course, this condition is of interest only if the set $\{g_\varepsilon > 0\}$ is sufficiently large with respect to $\{f > 0\}$ (where f is the density of P).

Theorem 5.2.2. (see [24]) *Let P be such that $\int_{\mathbb{R}^d} |x|^r dP(x) < +\infty$. For any distribution ν and any Borel function $g_\varepsilon : \mathbb{R}^d \to \mathbb{R}_+$, $\varepsilon \in (0, \frac{1}{3})$, satisfying (5.11),*

$$\forall n \geq 2, \quad e_r(a^{(n)}, P) \leq \varphi_r(\varepsilon)^{-\frac{1}{d}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{1}{d}} \left(\int g_\varepsilon^{-\frac{r}{d}} dP\right)^{\frac{1}{r}} (n-1)^{-\frac{1}{d}} \tag{5.12}$$

where $\varphi_r(u) = \left(\frac{1}{3^r} - u^r\right) u^d$.

Considering appropriate auxiliary distributions ν and "companion" functions g_ε satisfying (5.11) yields a Pierce type and a hybrid Zador-Pierce type L^r-rate optimality results as established in [24] (Zador type results are established in [45]).

Now, we give a micro-macro inequality established in [33] (see proof of Theorem 2) to estimate the increments $e_r(\Gamma^n,P)^r - e_r(\Gamma^{n+1},P)^r$, where $(\Gamma^n)_{n\geq 1}$ is a sequence of L^r-optimal quantizers of P. For every $n \geq 1$,

$$e_r(\Gamma^n,P)^r - e_r(\Gamma^{n+1},P)^r \leq \frac{4(2^r-1)e_r(\Gamma^{n+1},P)^r}{n+1} + \frac{4.2^r C_2^r n^{-\frac{r}{d}}}{n+1} \tag{5.13}$$

where C_2 is a finite constant independent of n.

The following Proposition provides a micro-macro inequality established in [24] for any quantizer Γ of X with distribution P.

Proposition 5.2.3. *Assume $\int |x|^r dP(x) < +\infty$. Let $y \in \mathbb{R}^d$ and $\Gamma \subset \mathbb{R}^d$ be a finite quantizer of a random variable X with distribution P such that $\mathrm{card}(\Gamma) \geq 1$. Then, for every probability distribution ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, every $c \in (0, \frac{1}{2})$*

$$e_r(\Gamma,P)^r - e_r(\Gamma\cup\{y\},P)^r \geq \frac{(1-c)^r - c^r}{(c+1)^r} \int \nu\left(B\Big(x, \frac{c}{c+1} d(x,\Gamma)\Big)\right) d(x,\Gamma)^r \, dP(x).$$

From this Proposition, one concludes the following either for L^r-optimal quantizers or for greedy sequences:

$\triangleright$ Since any sequence of L^r-optimal quantizers $(\Gamma^n)_{n\geq 1}$ clearly satisfies $e_r(\Gamma^{n+1},P) \leq e_r(\Gamma^n \cup \{y\},P)$ for every $y \in \mathbb{R}^d$, then

$$\begin{aligned} e_r(\Gamma^n,P)^r - e_r(\Gamma^{n+1},P)^r \geq & e_r(\Gamma^n,P)^r - e_r(\Gamma^n\cup\{y\},P)^r \\ \geq & \frac{(1-c)^r - c^r}{(c+1)^r} \int \nu\left(B\left(x, \frac{c}{c+1} d(x,\Gamma^n)\right)\right) d(x,\Gamma^n)^r \, dP(x). \end{aligned} \tag{5.14}$$

$\triangleright$ Likewise, since the greedy quantization sequence $(a_n)_{n\geq 1}$ satisfies $e_r(a^{(n+1)},P) \leq e_r(a^{(n)} \cup \{y\},P)$ for every $y \in \mathbb{R}^d$, then

$$e_r(a^{(n)},P)^r - e_r(a^{(n+1)},P)^r \geq \frac{(1-c)^r - c^r}{(c+1)^r} \int \nu\left(B\left(x, \frac{c}{c+1} d\left(x,a^{(n)}\right)\right)\right) d\left(x,a^{(n)}\right)^r dP(x). \tag{5.15}$$

5.3 Upper estimates for greedy quantizers

This is the main part of this chapter. Let $r,s>0$ and let $(a_n)_{n\geq 1}$ be an $L^r(\mathbb{R}^d)$-optimal greedy quantization sequence of a random variable X with probability distribution P. We denote $a^{(n)} = \{a_1,\ldots,a_n\}$ the first n terms of this sequence. For every $\mu \in \mathbb{R}^d$ and $\theta > 0$, we denote $a^{(n)}_{\theta,\mu} = \mu + \theta(a^{(n)} - \mu) = \{\mu + \theta(a_i - \mu),\ 1 \leq i \leq n\}$. In this section, we study the L^s-optimality of the sequence $a^{(n)}_{\theta,\mu}$.

For this, we consider auxiliary probability distributions ν satisfying the following control on balls with respect to an L^r-median a_1 of P: for every $\varepsilon \in (0,1)$, there exists a Borel function $g_\varepsilon : \mathbb{R}^d \to (0,+\infty)$ such that, for every $x \in \mathrm{supp}(\mathbb{P})$ and every $t \in [0, \varepsilon|x-a_1|]$,

$$\nu(B(x,t)) \geq g_\varepsilon(x) V_d t^d. \tag{5.16}$$

Note that $a_1 \in a^{(n)}$ for every $n \geq 1$ by construction of the greedy quantization sequence so that $d(x,a^{(n)}) \leq d(x,a_1)$ for every $x \in \mathbb{R}^d$.

5.3.1 Main results

The following result is an avatar of Pierce's Lemma for the L^s-error $e_s(a^{(n)}_{\theta,\mu}, P)$.

Theorem 5.3.1. *Let $s \in [r, d+r)$ and $1-q = \frac{d+r}{d+r-s}$. Let $(a_n)_{n\geq 1}$ be an $L^r(\mathbb{R}^d)$-optimal greedy quantization sequence of an $\mathbb{R}^d$-valued random variable X with distribution $P = f.\lambda_d$ such that $\mathbb{E}|X|^{r+\delta} < +\infty$ for some $\delta > 0$ such that $r + \delta > \frac{sd}{d+r-s}$. Let $\eta \in (0, r+\delta-\frac{sd}{d+r-s})$ and let $p' = \frac{r+\delta-\eta}{d|q|}$, $q' = \frac{r+\delta-\eta}{r+\delta-\eta-d|q|} > 1$ be two conjugate coefficients larger than 1. Assume*

$$\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{(1-q)q'} f d\lambda_d < +\infty. \tag{5.17}$$

Then, for every $n \geq 3$,

$$e_s(a^{(n)}_{\theta,\mu}, P) \leq \theta^{1+\frac{d}{s}} \kappa^{Greedy,Pierce}_{\theta,\mu} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{(1-q)q'} f d\lambda_d\right)^{\frac{1}{q'|q|(d+r)}} \sigma_{r+\delta}(P)(n-2)^{-\frac{1}{d}}. \tag{5.18}$$

where $e_{r+\delta}(a^{(1)}, P) = \sigma_{r+\delta}(P) < +\infty$ denotes the $L^{r+\delta}$-standard deviation of P and

$$\kappa^{Greedy,Pierce}_{\theta,\mu} = 2^{\frac{1}{d}+\frac{r+\delta}{r+d}(1+\frac{1}{|q|p'})} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} \min_{\varepsilon\in(0,\frac{1}{3})} \left[(1+\varepsilon)\varphi_r(\varepsilon)^{-\frac{1}{d}}\right] \left(\int (1\vee|x|)^{\frac{r+\delta}{r+\delta-\eta}} dx\right)^{\frac{1}{d}}.$$

When $s \in (0, r]$, notice that

$$e_s(a^{(n)}_{\theta,\mu}, P) \leq e_r(a^{(n)}_{\theta,\mu}, P)$$

where $e_r(a^{(n)}_{\theta,\mu}, P)$ is upper bounded as in Theorem 5.3.1. However, we are still interested in establishing a specific study for $s \in (0, r)$ and giving an upper bound for the L^s-error in the following theorem.

Theorem 5.3.2. *Let $s < r$ and X be a random variable in $\mathbb{R}^d$ with distribution $P = f.\lambda_d$ such that $\mathbb{E}|X|^{r+\delta} < +\infty$ for some $\delta > 0$. Assume*

$$\int_{\{f>0\}} f^{-\frac{s}{r-s}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d < +\infty.$$

Then, for every $n \geq 3$,

$$e_s(a^{(n)}_{\theta,\mu}, P) \leq \widetilde{\kappa}^{Greedy,Pierce}_{\theta,\mu} \theta^{1+\frac{d}{s}} \left(\int_{\{f>0\}} f^{-\frac{s}{r-s}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d\right)^{\frac{r-s}{sr}} \sigma_{r+\delta}(P)(n-2)^{-\frac{1}{d}} \tag{5.19}$$

where $e_{r+\delta}(a^{(1)}, P) = \sigma_{r+\delta}(P) < +\infty$ and

$$\widetilde{\kappa}^{Greedy,Pierce}_{\theta,\mu} = 2^{1+\frac{1}{d}+\frac{\delta}{r}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} \min_{\varepsilon\in(0,\frac{1}{3})} \left[(1+\varepsilon)\varphi_r(\varepsilon)^{-\frac{1}{d}}\right] \left(\int (1\vee|x|)^{-d(1+\frac{\delta}{r})} dx\right)^{-\frac{1}{d}}.$$

Application to radial densities

In this section, we consider probability distributions with radial densities. In other words, if the random variable X has distribution $P = f.\lambda_d$, we consider the auxiliary distribution

$$\nu = \frac{f^a}{\int f^a d\lambda_d}.\lambda_d := f_a.\lambda_d$$

for $a \in (0,1)$ where the density function f is radial with non-increasing tails w.r.t. $a_1 \in A$ who is peakless w.r.t. a_1. These two terms are defined as follows

Definition 5.3.3. *(a) Let $A \subset \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}_+$ is said to be almost radial non-increasing on A w.r.t. $a \in A$ if there exists a norm $\|.\|_0$ on $\mathbb{R}^d$ and real constant $M \in (0,1]$ such that*

$$\forall x \in A \setminus \{a\}, \quad f_{|B_{\|.\|_0}(a,\|x-a\|_0)} \geq M f(x). \tag{5.20}$$

If (5.20) holds for $M = 1$, then f is called radial non-increasing on A w.r.t. a.
(b) A set A is said to be star-shaped and peakless with respect to a_1 if

$$\mathfrak{p}(A, |\cdot - a_1|) := \inf\left\{\frac{\lambda_d(B(x,t) \cap A)}{\lambda_d(B(x,t))}; x \in A, 0 < t < |x - a_1|\right\} > 0 \tag{5.21}$$

for any norm $|\cdot|$ on $\mathbb{R}^d$.

Remark 5.3.4. *(a) (5.20) reads $f(y) \geq M f(x)$ for all $x, y \in A \setminus \{a\}$ for which $\|y-a\|_0 \leq \|x-a\|_0$.*
(b) If f is radial non-increasing on $\mathbb{R}^d$ w.r.t. $a \in \mathbb{R}^d$ with parameter $\|.\|_0$, then there exists a non-increasing measurable function $g : (0, +\infty) \to \mathbb{R}_+$ satisfying $f(x) = g(\|x - a\|_0)$ for every $x \neq a$.
(c) From a practical point of view, many classes of distributions satisfy (5.20), e.g. the d-dimensional Normal distribution $\mathcal{N}(m, \sigma_d)$ for which one considers $h(y) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\sigma_d)^{\frac{1}{2}}} e^{-\frac{y^2}{2}}$ and density $f(x) = h(\|x - m\|_0)$ where $\|x\|_0 = |\sigma_d^{-\frac{1}{2}} x|$, and the family of distributions defined by $f(x) \propto |x|^c e^{-a|x|^b}$, for every $x \in \mathbb{R}^d, a, b > 0$ and $c > -d$, for which one considers $h(u) = u^c e^{-au^b}$. In the one dimensional case, we can mention the Gamma distribution, the Weibull distributions, the Pareto distributions and the log-Normal distributions.
(d) If $A = \mathbb{R}^d$, then $\mathfrak{p}(A, |\cdot - a|) = 1$ for every $a \in \mathbb{R}^d$.

(e) The most typical unbounded sets satisfying (5.21) are convex cones that is cones $K \subset \mathbb{R}^d$ of vertex 0 with $0 \in K$ ($K \neq \varnothing$) and such that $\lambda x \in K$ for every $x \in K$ and $\lambda \geq 0$. For such convex cones K with $\lambda_d(K) > 0$, we even have that the lower bound

$$\mathfrak{p}(K) := \inf\left\{\frac{\lambda_d(B(x,t) \cap K)}{\lambda_d(B(x,t))}; x \in K,\, t > 0\right\} = \frac{\lambda_d\big(B(0,1) \cap K\big)}{V_d} > 0.$$

Thus if $K = \mathbb{R}^d_+$, then $\mathfrak{p}(K) = 2^{-d}$.

Theorem 5.3.5. *Let $s \in [r, d + r)$ and $1 - q = \frac{d+r}{d+r-s}$. Assume that $P = f.\lambda_d$ has finite polynomial moments of order $\frac{(1-a)(d+\varepsilon)}{a}$ for some $a \in (0,1)$ and $\varepsilon > 0$. Let a_1 denote the L^r-median of P and assume that $supp(P) \subset A$ and $a_1 \in A$ for some A star-shaped and peakless*

with respect to a_1 and that f is almost radial non-increasing with respect to a_1 in the sense of (5.20). Assume

$$\int_{\{f>0\}} f^{\frac{-s(1+a)}{d+r-s}} f_{\theta,\mu}^{\frac{d+r}{d+r-s}} d\lambda_d < +\infty. \tag{5.22}$$

Then, for every $n \geq 3$,

$$e_s(a_{\theta,\mu}^{(n)}, P) \leq \kappa_{\theta,\mu}^{G,Z,P}\, \theta^{1+\frac{d}{s}} \|f\|_{\frac{d}{d+r}}^{\frac{1}{d+r}} \|f\|_a^{\frac{a}{d+r}} \left(\int_{\{f>0\}} f^{\frac{-s(1+a)}{d+r-s}} f_{\theta,\mu}^{\frac{d+r}{d+r-s}} d\lambda_d \right)^{\frac{1}{|q|(d+r)}} (n-2)^{-\frac{1}{d}},$$

where $\kappa_{\theta,\mu}^{G,Z,P} \leq \frac{2^{1+\frac{1}{d}} C_0^2\, r^{\frac{1}{d}}}{d^{\frac{1}{d}} M^{\frac{1}{d}} V_d^{\frac{1}{d}} \mathfrak{p}(A,|\cdot-a_1|)^{\frac{1}{d}}} \min_{\varepsilon\in(0,\frac{1}{3})} \left[\varphi_r(\varepsilon)^{-\frac{1}{d}}\right]$.

Remark 5.3.6. *Note that the condition (5.22) is more restrictive than the condition (5.17) in a sense that the set of values of θ for which (5.22) is satisfied is smaller than the set for which (5.17) is satisfied. This will be made precise and clear in Section 5.5 for particular distributions. However, if P has finite polynomial moments of any order $r > 0$, i.e. the parameter a in Theorem 5.3.5 being as small as possible ($a \to 0^+$), then the condition (5.22) yield the same interval as (5.17).*

5.3.2 Proofs

General results

We first state two rather theoretical results based on the auxiliary distribution ν and its companion function g_ε satisfying (5.16). More operating criterions based on moments of P and/or the radial structure of its densities will appear as consequences of these theorems by specifying the distribution ν (and g_ε).

Theorem 5.3.7. *Let $s \in [r, d+r)$ and $1-q = \frac{d+r}{d+r-s}$. Let $(a_n)_{n\geq 1}$ be an $L^r(\mathbb{R}^d)$-optimal greedy quantization sequence of an $\mathbb{R}^d$-valued random variable X with distribution $P = f.\lambda_d$ such that $\mathbb{E}|X|^{r+\delta} < +\infty$ for some $\delta > 0$. Assume there exists an auxiliary distribution ν and a Borel function g_ε satisfying (5.16) for $\varepsilon \in (0, \frac{1}{3})$ such that*

$$\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|} dP_{\theta,\mu}(x) < +\infty.$$

Then, for every $n \geq 3$,

$$e_s(a_{\theta,\mu}^{(n)}, P) \leq \theta^{1+\frac{d}{d+r}} \kappa_{\theta,\mu}^{greedy} \left(\int g_\varepsilon^{-\frac{r}{d}} dP \right)^{\frac{1}{d+r}} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|} dP_{\theta,\mu}(x) \right)^{\frac{1}{|q|(d+r)}} (n-2)^{-\frac{1}{d}} \tag{5.23}$$

where $\kappa_{\theta,\mu}^{greedy} = 2^{\frac{1}{d}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} \min_{\varepsilon\in(0,\frac{1}{3})} \left[\varphi_r(\varepsilon)^{-\frac{1}{d}}\right]$.

Proof. We start by noticing that, for every $n \geq 1$,

$$e_s(a_{\theta,\mu}^{(n)}, P)^s = \int_{\mathbb{R}^d} d(z, a_{\theta,\mu}^{(n)})^s f(z) d\lambda_d(z) = \int_{\mathbb{R}^d} \min_{x_i \in a^{(n)}, 1\leq i\leq n} \left|z - \mu + \theta(\mu - x_i)\right|^s f(z) d\lambda_d(z).$$

Then, by applying the change of variables $x = \frac{z-\mu}{\theta} + \mu$, one obtains

$$\begin{aligned} e_s(a^{(n)}_{\theta,\mu}, P)^s =& \theta^{s+d} \int_{\mathbb{R}^d} d(x, a^{(n)})^s f(\mu + \theta(x-\mu)) d\lambda_d(x) \\ =& \theta^s \int_{\mathbb{R}^d} d(x, a^{(n)})^s dP_{\theta,\mu}(x) \\ =& \theta^s e_s(a^{(n)}, P_{\theta,\mu})^s. \end{aligned} \tag{5.24}$$

Now, let us study $e_s(a^{(n)}, P_{\theta,\mu})$. Consider $c \in (0, \frac{\varepsilon}{1-\varepsilon}] \cap (0, \frac{1}{2})$ so that $\frac{c}{c+1} \leq \varepsilon$. Hence, for any such c, $\dfrac{c}{c+1}\, d(x, a^{(n)}) \leq \varepsilon |x - a_1|$ since $a_1 \in a^{(n)}$. Consequently, criteria (5.16) is satisfied, so there exists a function g_ε such that

$$\nu\left(B\left(x, \frac{c}{c+1}\, d(x, a^{(n)})\right)\right) \geq V_d \left(\frac{c}{c+1}\right)^d d(x, a^{(n)})^d\, g_\varepsilon(x).$$

Then, noticing that $\frac{(1-c)^r - c^r}{(1+c)^r} \geq \frac{1}{3^r} - \left(\frac{c}{c+1}\right)^r > 0$, since $c \in (0, \frac{1}{2})$, (5.15) yields

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq V_d\, \varphi_r\left(\frac{c}{c+1}\right) \int g_\varepsilon(x) d(x, a^{(n)})^{d+r} dP(x) \tag{5.25}$$

where $\varphi_r(u) = \left(\dfrac{1}{3^r} - u^r\right) u^d,\ u \in (0, \frac{1}{3})$. Consequently,

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq V_d\, \varphi_r\left(\frac{c}{c+1}\right) \theta^{-d} \int_{\mathbb{R}^d} g_\varepsilon(x) d(x, a^{(n)})^{d+r} f(x) f^{-1}_{\theta,\mu}(x) dP_{\theta,\mu}(x).$$

Now, applying the reverse Hölder inequality with conjugate exponents $p = \frac{s}{d+r} \in (0,1)$ and $q = \frac{-s}{d+r-s} < 0$ yields

$$\begin{aligned} e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq& V_d\, \varphi_r\left(\frac{c}{c+1}\right) \theta^{-d} \left(\int_{\{f>0\}} (g_\varepsilon(x) f(x) f^{-1}_{\theta,\mu}(x))^q dP_{\theta,\mu}(x)\right)^{\frac{1}{q}} \\ & \times \left(\int_{\mathbb{R}^d} d(x, a^{(n)})^s dP_{\theta,\mu}(x)\right)^{\frac{1}{p}} \\ \geq& V_d\, \varphi_r\left(\frac{c}{c+1}\right) \theta^{-d} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|}(x) dP_{\theta,\mu}(x)\right)^{\frac{1}{q}} e_s(a^{(n)}, P_{\theta,\mu})^{d+r}. \end{aligned} \tag{5.26}$$

Consequently, denoting $C_1 = V_d\, \varphi_r\left(\dfrac{c}{c+1}\right) \theta^{-d} \left(\displaystyle\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|}(x) dP_{\theta,\mu}(x)\right)^{\frac{1}{q}}$, one obtains

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq C_1 e_s(a^{(n)}, P_{\theta,\mu})^{d+r}. \tag{5.27}$$

At this stage, we know that $e_r(a^{(k)}, P)$ is decreasing w.r.t k and it is clear that it is the same for $e_s(a^{(k)}, P_{\theta,\mu})$, since

$$e_s(a^{(k)}, P_{\theta,\mu}) = \mathbb{E}\left[\min_{1\leq i\leq k} |a_i - \frac{X-\mu}{\theta} - \mu|^s\right]^{\frac{1}{s}} \geq \mathbb{E}\left[\min_{1\leq i\leq k+1} |a_i - \frac{X-\mu}{\theta} - \mu|^s\right]^{\frac{1}{s}} = e_s(a^{(k+1)}, P_{\theta,\mu}),$$

so, one has

$$n\, e_s(a^{(2n-1)}, P_{\theta,\mu})^{d+r} \leq \sum_{k=n}^{2n-1} e_s(a^{(k)}, P_{\theta,\mu})^{d+r} \leq \frac{1}{C_1} \sum_{k=n}^{2n-1} e_r(a^{(k)}, P)^r - e_r(a^{(k+1)}, P)^r \leq \frac{1}{C_1} e_r(a^{(n)}, P)^r.$$

and, since $2\left\lceil \frac{n}{2} \right\rceil - 1 \leq n$,

$$\frac{n}{2} e_s(a^{(n)}, P_{\theta,\mu})^{d+r} \leq \left\lceil \frac{n}{2} \right\rceil e_s(a^{(n)}, P_{\theta,\mu})^{d+r} \leq \left\lceil \frac{n}{2} \right\rceil e_s\left(a^{2\left\lceil \frac{n}{2} \right\rceil - 1}, P_{\theta,\mu}\right)^{d+r} \leq \frac{1}{C_1} e_r\left(a^{\left\lceil \frac{n}{2} \right\rceil}, P\right)^r.$$

Consequently, using the result of Theorem 5.2.2

$$\begin{aligned} e_s(a^{(n)}, P_{\theta,\mu}) \leq & \left(\frac{2}{C_1}\right)^{\frac{1}{d+r}} n^{-\frac{1}{d+r}} e_r\left(a^{\left\lceil \frac{n}{2} \right\rceil}, P\right)^{\frac{r}{d+r}} \\ & \leq 2^{\frac{1}{d}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} \varphi_r\left(\frac{c}{c+1}\right)^{-\frac{1}{d}} \theta^{\frac{d}{d+r}} \left(\int_{\mathbb{R}^d} g_\varepsilon^{-\frac{r}{d}} dP\right)^{\frac{1}{d+r}} \\ & \quad \times \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|} (x) dP_{\theta,\mu}(x)\right)^{\frac{1}{|q|(d+r)}} (n-2)^{-\frac{1}{d}}. \end{aligned}$$

We are led to study $\varphi_r\left(\frac{c}{c+1}\right)^{-\frac{1}{d}}$ subject to the constraint $c \in (0, \frac{\varepsilon}{1-\varepsilon}] \cap (0, \frac{1}{2})$. φ_r is increasing in the neighborhood of 0 and $\varphi_r(0)$, so, one has, for every $\varepsilon \in (0, \frac{1}{3})$ small enough, $\varphi_r\left(\frac{c}{c+1}\right) \leq \varphi_r(\varepsilon)$, for $c \in (0, \frac{\varepsilon}{1-\varepsilon}]$. This leads to specify c as $c = \frac{\varepsilon}{1-\varepsilon}$, so that $\frac{c}{c+1} = \varepsilon$ which means that one can use

$$\varphi_r\left(\frac{c}{c+1}\right)^{-\frac{1}{d+r}} \leq \min_{\varepsilon \in (0, \frac{1}{3})} \left[\varphi_r(\varepsilon)^{-\frac{1}{d+r}}\right] \tag{5.28}$$

which yields

$$\begin{aligned} e_s(a^{(n)}, P_{\theta,\mu}) \leq & 2^{\frac{1}{d}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} \min_{\varepsilon \in (0, \frac{1}{3})} \left[\varphi_r(\varepsilon)^{-\frac{1}{d}}\right] \theta^{\frac{d}{d+r}} \left(\int_{\mathbb{R}^d} g_\varepsilon^{-\frac{r}{d}} dP\right)^{\frac{1}{d+r}} \\ & \times \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|} (x) dP_{\theta,\mu}(x)\right)^{\frac{1}{|q|(d+r)}} (n-2)^{-\frac{1}{d}}. \end{aligned} \tag{5.29}$$

Finally, one concludes by merging this with (5.24). □

Theorem 5.3.8. *Let $s < r$ and X a random variable in $\mathbb{R}^d$ with distribution $P = f.\lambda_d$ and such that $\mathbb{E}|X|^{r+\delta} < +\infty$ for some $\delta > 0$. Assume there exists an auxiliary distribution ν and a Borel function g_ε satisfying (5.16) for every $\varepsilon \in (0, \frac{1}{3})$ such that*

$$\int_{\mathbb{R}^d} g_\varepsilon^{-\frac{r}{d}} dP < +\infty \quad \text{and} \quad \int_{\{f>0\}} f^{-\frac{s}{r-s}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d < +\infty.$$

Then, for every $n \geq 3$,

$$e_s(a^{(n)}_{\theta,\mu}, P) \leq \theta^{1+\frac{d}{s}} \kappa^{Greedy}_{\theta,\mu} \left(\int_{\mathbb{R}^d} g_\varepsilon^{-\frac{r}{d}} dP \right)^{\frac{1}{r}} \left(\int_{\{f>0\}} f^{-\frac{s}{r-s}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d \right)^{\frac{r-s}{sr}} (n-2)^{-\frac{1}{d}} \tag{5.30}$$

where $\kappa^{Greedy}_{\theta,\mu} = 2^{1+\frac{1}{d}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} \min_{\varepsilon \in (0,\frac{1}{3})} \left[\varphi_r(\varepsilon)^{-\frac{1}{d}}\right]$.

Proof. We start from Equation (5.27) in the proof of Theorem 5.3.7 recalled below

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq C_1 e_s(a^{(n)}, P_{\theta,\mu})^{d+r}$$

where $C_1 = \varphi_r\left(\frac{c}{c+1}\right) \theta^{-d+\frac{d}{q}} \left(\int_{\{f>0\}} g_\varepsilon^q(x) f^q(x) f_{\theta,\mu}^{1-q}(x) d\lambda_d(x) \right)^{\frac{1}{q}}$ and $q = -\frac{s}{d+r-s} < 0$ so that $1-q = \frac{d+r}{d+r-s}$. At this stage, follow the lines of the proof of Theorem 5.3.7 to get, for $n \geq 3$,

$$\begin{aligned} e_s(a^{(n)}, P_{\theta,\mu}) &\leq \left(\frac{2}{C_1}\right)^{\frac{1}{d+r}} (n-1)^{-\frac{1}{d+r}} e_r\left(a^{\lceil \frac{n}{2} \rceil}, P\right)^{\frac{r}{d+r}} \\ &\leq \kappa^{\text{Greedy}}_{\theta,\mu} \theta^{\frac{d}{s}} \left(\int_{\mathbb{R}^d} g_\varepsilon^{-\frac{r}{d}} dP \right)^{\frac{1}{d+r}} \left(\int_{\{f>0\}} g_\varepsilon^q f^q f_{\theta,\mu}^{1-q} d\lambda_d \right)^{\frac{1}{|q|(d+r)}} (n-2)^{-\frac{1}{d}}. \end{aligned}$$

where $\kappa^{\text{Greedy}}_{\theta,\mu} = 2^{1+\frac{1}{d}} V_d^{-\frac{1}{d}} \left(\frac{r}{d}\right)^{\frac{r}{d(d+r)}} \min_{\varepsilon \in (0,\frac{1}{3})} \left[\varphi_r(\varepsilon)^{-\frac{1}{d}}\right]$.
Now, since $s < r$, one can apply Hölder inequality with the conjugate exponents $p' = \frac{r(d+r-s)}{r(d+r-s)-ds} > 1$ and $q' = \frac{r}{d|q|} = \frac{r(d+r-s)}{ds} > 1$ which yields

$$\int_{\{f>0\}} g_\varepsilon^q f^q f_{\theta,\mu}^{1-q} d\lambda_d = \int_{\{f>0\}} g_\varepsilon^q f^{q-1} f_{\theta,\mu}^{1-q} dP \leq \left(\int_{\mathbb{R}^d} g_\varepsilon^{-\frac{r}{d}} dP \right)^{\frac{1}{q'}} \left(\int_{\{f>0\}} f^{\frac{r}{s-r}+1} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d \right)^{\frac{1}{p'}}$$

so

$$\left(\int_{\{f>0\}} g_\varepsilon^q f^q f_{\theta,\mu}^{1-q} d\lambda_d \right)^{-\frac{1}{q(d+r)}} \leq \left(\int_{\mathbb{R}^d} g_\varepsilon^{-\frac{r}{d}} dP \right)^{\frac{d}{r(d+r)}} \left(\int_{\{f>0\}} f^{\frac{s}{s-r}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d \right)^{\frac{r-s}{rs}}$$

and

$$e_s(a^{(n)}, P_{\theta,\mu}) \leq \kappa^{\text{Greedy}}_{\theta,\mu} \theta^{\frac{d}{s}} \left(\int_{\mathbb{R}^d} g_\varepsilon^{-\frac{r}{d}} dP \right)^{\frac{1}{r}} \left(\int_{\{f>0\}} f^{-\frac{s}{r-s}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d \right)^{\frac{r-s}{sr}} (n-2)^{-\frac{1}{d}}.$$

and one deduces the result just as in the proof of Theorem 5.3.7. □

Proofs of main results

Proof of Theorem 5.3.1. We consider $\nu(dx) = \gamma_{r,\delta}(x)\lambda_d(dx)$ where

$$\gamma_{r,\delta}(x) = \frac{K_{\delta,r}}{(1 \vee |x - a_1|)^{d\frac{r+\delta}{r+\delta-\eta}}} \qquad \text{with} \qquad K_{\delta,r} = \left(\int \frac{dx}{(1 \vee |x|)^{\frac{r+\delta}{r+\delta-\eta}}} \right)^{-1} < +\infty$$

is a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$. For every $x \in \mathbb{R}^d$ such that $\varepsilon|x-a_1| \geq t$ and every $y \in B(x,t)$, one has $|y-a_1| \leq |y-x| + |x-a_1| \leq (1+\varepsilon)|x-a_1|$ so that

$$\nu(B(x,t)) \geq \frac{K_{\delta,r} V_d \, t^d}{\left(1 \vee (1+\varepsilon)|x-a_1|\right)^{d\frac{r+\delta}{r+\delta-\eta}}}.$$

Hence, (5.16) is satisfied with

$$g_\varepsilon(x) = \frac{K_{\delta,r}}{\left(1 \vee (1+\varepsilon)|x-a_1|\right)^{\frac{r+\delta}{r+\delta-\eta}}}$$

so we apply Theorem 5.3.7 where one has to handle the term

$$\left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|}(x) dP_{\theta,\mu}(x)\right)^{\frac{1}{|q|(d+r)}} = \theta^{\frac{d}{|q|(d+r)}} \left(\int_{\{f>0\}} g_\varepsilon^q \left(\frac{f_{\theta,\mu}}{f}\right)^{1-q} dP(x)\right)^{\frac{1}{|q|(d+r)}}$$

where $q = \frac{-s}{d+r-s} < 0$ so that $1-q = \frac{d+r}{d+r-s}$. To do this, we apply Hölder inequality with the conjugate coefficients $p' = \frac{r+\delta-\eta}{d|q|} > 1$ (due to the moment assumption on P) and $q' = \frac{r+\delta-\eta}{r+\delta-\eta-d|q|} > 1$. This yields

$$\begin{aligned}\left(\int_{\{f>0\}} g_\varepsilon^q \left(\frac{f_{\theta,\mu}}{f}\right)^{1-q} dP\right)^{\frac{1}{|q|(d+r)}} &\leq \left(\int_{\mathbb{R}^d} g_\varepsilon^{qp'} dP\right)^{\frac{1}{p'|q|(d+r)}} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{(1-q)q'} dP\right)^{\frac{1}{q'|q|(d+r)}} \\ &\leq \left(\int_{\mathbb{R}^d} g_\varepsilon^{qp'} dP\right)^{\frac{1}{p'|q|(d+r)}} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{(1-q)q'} f d\lambda_d\right)^{\frac{1}{q'|q|(d+r)}}\end{aligned}$$

so that

$$\begin{aligned}\left(\int_{\{f>0\}} (g_\varepsilon(x) f(x) f_{\theta,\mu}^{-1}(x))^q dP_{\theta,\mu}(x)\right)^{\frac{1}{|q|(d+r)}} &\leq \theta^{\frac{d}{|q|(d+r)}} \left(\int_{\mathbb{R}^d} g_\varepsilon^{qp'} dP\right)^{\frac{1}{p'|q|(d+r)}} \\ &\quad \times \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{(1-q)q'} f d\lambda_d\right)^{\frac{1}{q'|q|(d+r)}}. \quad (5.31)\end{aligned}$$

Consequently,

$$\begin{aligned}e_s(a_{\theta,\mu}^{(n)}, P) \leq \theta^{1+\frac{d}{s}} \kappa_{\theta,\mu}^{\text{Greedy}} &\left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{(1-q)q'} f d\lambda_d\right)^{\frac{1}{q'|q|(d+r)}} \left(\int_{\mathbb{R}^d} g_\varepsilon^{-\frac{r}{d}} dP\right)^{\frac{1}{d+r}} \\ &\times \left(\int_{\mathbb{R}^d} g_\varepsilon^{qp'} dP\right)^{\frac{1}{p'|q|(d+r)}} (n-2)^{-\frac{1}{d}}\end{aligned}$$

By our choice of g_ε,

$$\left(\int_{\mathbb{R}^d} g_\varepsilon^{-\frac{r}{d}} dP\right)^{\frac{1}{d+r}} \leq \left(\int_{\mathbb{R}^d} \left(1 \vee (1+\varepsilon)\|x-a_1\|\right)^{\frac{r(r+\delta)}{d(r+\delta-\eta)}} dP\right)^{\frac{1}{d+r}}$$

and

$$\left(\int_{\mathbb{R}^d} g_\varepsilon^{qp'} dP\right)^{\frac{1}{p'|q|(d+r)}} \leq \left(\int_{\mathbb{R}^d} \left(1 \vee (1+\varepsilon)\|x-a_1\|\right)^{r+\delta} dP\right)^{\frac{1}{p'|q|(d+r)}}.$$

At this stage, notice that $\frac{r(r+\delta)}{d(r+\delta-\eta)} < r+\delta$ since $r+\delta-\eta > \frac{sd}{d+r-s} > \frac{r}{d}$. So,

$$\int_{\mathbb{R}^d} \left(1 \vee (1+\varepsilon)\|x-a_1\|\right)^{\frac{r(r+\delta)}{d(r+\delta-\eta)}} dP < \int_{\mathbb{R}^d} \left(1 \vee (1+\varepsilon)\|x-a_1\|\right)^{r+\delta} dP$$

since the function $x \mapsto a^x$ is increasing w.r.t x for $a > 1$. Moreover, owing to $L^{r+\delta}$-Minkowski inequality,

$$\left(\int_{\mathbb{R}^d} \left(1 \vee (1+\varepsilon)\|x-a_1\|\right)^{r+\delta} dP\right)^{\frac{1}{d+r}\left(1+\frac{1}{|q|p'}\right)} \leq \left(1+(1+\varepsilon)\sigma_{r+\delta}(P)\right)^{\frac{r+\delta}{r+d}\left(1+\frac{1}{|q|p'}\right)}$$

where $\sigma_{r+\delta}(P) = \inf_a \|X-a\|_{r+\delta}$ is the $L^{r+\delta}$-standard deviation of P. Consequently,

$$e_s(a^{(n)}_{\theta,\mu}, P) \leq \theta^{1+\frac{d}{s}} \frac{\kappa^{\text{Greedy}}_{\theta,\mu}}{K^{\frac{1}{d}}_{\delta,r}} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{(1-q)q'} f d\lambda_d\right)^{\frac{1}{q'|q|(d+r)}} \left(1+(1+\varepsilon)\sigma_{r+\delta}(P)\right)^{\frac{(r+\delta)(1+|q|p')}{|q|p'(r+d)}} (n-2)^{-\frac{1}{d}}.$$

Now, we introduce an equivariance argument. For $\lambda > 0$, let $X_\lambda := \lambda(X-a_1)+a_1$ and $(\alpha_{\lambda,n})_{n\geq 1} := (\lambda(\alpha_n - a_1)+a_1)_{n\geq 1}$. It is clear that $e_r(\alpha^{(n)}, X) = \frac{1}{\lambda} e_r(\alpha^{(n)}_\lambda, X_\lambda)$. Plugging this in the previous inequality yields

$$\begin{aligned} e_s(a^{(n)}_{\theta,\mu}, P) \leq \theta^{1+\frac{d}{s}} \kappa^{\text{Greedy}}_{\theta,\mu} K^{-\frac{1}{d}}_{\delta,r} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{(1-q)q'} f d\lambda_d\right)^{\frac{1}{q'|q|(d+r)}} \\ \times \frac{1}{\lambda}\left(1+(1+\varepsilon)\lambda\sigma_{r+\delta}(P)\right)^{\frac{(r+\delta)(1+|q|p')}{|q|p'(r+d)}} (n-2)^{-\frac{1}{d}}. \end{aligned}$$

Finally, one deduces the result by setting $\lambda = \dfrac{1}{(1+\varepsilon)\sigma_{r+\delta}}$. □

Proof of Theorem 5.3.2. We consider the function g_ε defined by

$$g_\varepsilon(x) = \frac{K_{\delta,r}}{\left(1 \vee (1+\varepsilon)|x-a_1|\right)^{d(1+\frac{\delta}{r})}}$$

where $K_{\delta,r} = \left(\int \dfrac{dx}{(1 \vee |x|)^{d(1+\frac{\delta}{r})}}\right)^{-1} < +\infty$. One has

$$\left(\int_{\mathbb{R}^d} g_\varepsilon^{-\frac{r}{d}}(x) dP\right)^{\frac{1}{r}} \leq K^{-\frac{1}{d}}_{\delta,r} \left(\int \left(1 \vee (1+\varepsilon)|x-a_1|\right)^{r+\delta} dP\right)^{\frac{1}{r}}$$

so that, applying the $L^{r+\delta}$-Minkowski inequality, one obtains

$$\left(\int g_\varepsilon(x)^{-\frac{r}{d}} dP(x)\right)^{\frac{1}{r}} \leq K^{-\frac{1}{d}}_{\delta,r} \left(1+(1+\varepsilon)\sigma_{r+\delta}\right)^{1+\frac{\delta}{r}}.$$

Then, applying Theorem 5.3.8 yields , for every $n \geq 3$,

$$e_s(a^{(n)}_{\theta,\mu}, P) \leq \theta^{1+\frac{d}{s}} \kappa^{\text{Greedy}}_{\theta,\mu} K^{-\frac{1}{d}}_{\delta,r} \left(1+(1+\varepsilon)\sigma_{r+\delta}\right)^{1+\frac{\delta}{r}} \left(\int_{\{f>0\}} f^{-\frac{s}{r-s}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d\right)^{\frac{r-s}{sr}} (n-2)^{-\frac{1}{d}} \tag{5.32}$$

Finally, using the equivariance argument introduced in the proof of Theorem 5.3.1, one deduces, in the same spirit, the result by considering $\lambda = \frac{1}{(1+\varepsilon)\sigma_{r+\delta}(P)}$. □

For the proof of Theorem 5.3.5, we use the following technical lemma (established in [24]).

Lemme 5.3.9. *Let $\nu = f.\lambda_d$ be a probability measure on $\mathbb{R}^d$ where f is almost radial non-increasing on $A \in \mathcal{B}(\mathbb{R}^d)$ w.r.t. $a_1 \in A$, A being star-shaped relative to a_1 and satisfying (5.21). Then, for every $x \in A$ and $t \in (0, |x - a_1|)$,*

$$\nu(B(x,t)) \geq M\mathfrak{p}(A, |\cdot - a_1|)(2C_0^2)^{-d} V_d f(x) t^d$$

where $C_0 \in [1, +\infty)$ is such that, for every $x \in \mathbb{R}^d$, $\frac{1}{C_0}\|x\|_0 \leq |x| \leq C_0\|x\|_0$.

Proof of theorem 5.3.5. We consider $\nu = f_a d\lambda_d$ for $a \in (0,1)$ where

$$f_a = K_a\, f^a \quad \text{with} \quad K_a = \left(\int f^a d\lambda_d\right)^{-1}.$$

Note that $\int f^a d\lambda_d < +\infty$. In fact, if we denote $f^a = f^a(1+|x|)^b(1+|x|)^{-b}$ where $b = (1-a)(d+\varepsilon)$, $\varepsilon > 0$, then, applying Hölder's inequality with the conjugate coefficients $\frac{1}{a}$ and $\frac{1}{1-a}$ yields

$$\int f^a(x) d\lambda_d(x) \leq \left(\int f(x)\,(1+|x|)^{\frac{1-a}{a}(d+\varepsilon)} d\lambda_d(x)\right)^a \left(\int (1+|x|)^{-(d+\varepsilon)} d\lambda_d(x)\right)^{1-a}$$

where the first factor is finite due to the moment assumption made on P and the second factor is finite for $\varepsilon > 0$.

Let $c \in (0, \frac{1}{2})$. Since $\frac{c}{c+1} < 1$ and $a_1 \in a^{(n)}$ then, for every $x \in \mathbb{R}^d$, $\frac{c}{c+1} d(x, a^{(n)}) \leq d(x, a^{(n)}) \leq |x - a_1|$. Moreover, notice that f_a is radial non-increasing with parameter M^a. So, merging (5.15) with Lemma 5.3.9, one obtains

$$e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r \geq \varphi_r\left(\frac{c}{c+1}\right) M^a \mathfrak{p}(A, |\cdot - a_1|)(2C_0^2)^{-d} V_d \int f_a(x) d(x, a^{(n)})^{d+r} dP(x).$$

Now, denoting $C = \varphi_r\left(\frac{c}{c+1}\right) M^a \mathfrak{p}(A, |\cdot - a_1|)(2C_0^2)^{-d} V_d$ and having in mind that $dP = f.d\lambda_d$ and $dP_{\theta,\mu} = \theta^d f_{\theta,\mu}.d\lambda_d$, yields

$$\begin{aligned} e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r &\geq C\theta^{-d} \int_{\{f>0\}} f_a(x) f(x) f_{\theta,\mu}^{-1}(x) d(x, a^{(n)})^{d+r} dP_{\theta,\mu}(x) \\ &\geq C\theta^{-d} K_a \int_{\{f>0\}} f(x)^{1+a} f_{\theta,\mu}^{-1}(x) d(x, a^{(n)})^{d+r} dP_{\theta,\mu}(x). \end{aligned}$$

Applying the reverse Hölder inequality with the conjugate exponents $p = \frac{s}{d+r} \in (0,1)$ and $q = \frac{-s}{d+r-s} < 0$ yields

$$\begin{aligned} e_r(a^{(n)}, P)^r - e_r(a^{(n+1)}, P)^r &\geq C\theta^{-d} K_a \left(\int_{\{f>0\}} f(x)^{-|q|(1+a)} f_{\theta,\mu}^{|q|}(x) dP_{\theta,\mu}(x)\right)^{\frac{1}{q}} \left(\int_{\mathbb{R}^d} d(x, a^{(n)})^s dP_{\theta,\mu}(x)\right)^{\frac{d+r}{s}} \\ &\geq C\theta^{-d+\frac{d}{q}} K_a \left(\int_{\{f>0\}} f(x)^{-|q|(1+a)} f_{\theta,\mu}^{1-q}(x) d\lambda_d(x)\right)^{\frac{1}{q}} e_s(a^{(n)}, P_{\theta,\mu})^{d+r}. \end{aligned}$$

At this stage, we denote $C_1 = C\theta^{-d+\frac{d}{q}} K_a \left(\int_{\{f>0\}} f(x)^{-|q|(1+a)} f_{\theta,\mu}^{1-q}(x) d\lambda_d(x) \right)^{\frac{1}{q}}$, follow the same steps as in the proof of Theorem 5.3.7 and use the result of Theorem 2.2.8 in [24] to obtain

$$
\begin{aligned}
e_s(a^{(n)}, P_{\theta,\mu}) &\leq \left(\frac{2}{C_1} \right)^{\frac{1}{d+r}} (n-2)^{-\frac{1}{d+r}} e_r\left(a^{\lceil \frac{n}{2} \rceil}, P \right)^{\frac{r}{d+r}} \\
&\leq \frac{2^{1+\frac{1}{d}} C_0^2 \, r^{\frac{1}{d}}}{d^{\frac{1}{d}} M^{\frac{1}{d}} V_d^{\frac{1}{d}} \mathfrak{p}(A, |\cdot - a_1|)^{\frac{1}{d}}} \min_{\varepsilon \in (0, \frac{1}{3})} [\varphi_r(\varepsilon)^{-\frac{1}{d}}] \theta^{\frac{d}{s}} \left(\int_{\{f>0\}} f(x)^{-|q|(1+a)} f_{\theta,\mu}^{\frac{d+r}{d+r-s}}(x) d\lambda_d(x) \right)^{\frac{1}{|q|(d+r)}} \\
&\quad \times \|f\|_{\frac{d}{d+r}}^{\frac{1}{d+r}} \|f\|_a^{\frac{a}{d+r}} (n-2)^{-\frac{1}{d}}.
\end{aligned}
$$

The result is deduced using the same arguments as in the end of the proof of Theorem 5.3.7. □

5.3.3 Example of distributions with finite polynomial moments up to a finite order

Theorem 5.3.1 treats the case of a distribution P that has finite polynomial moments at any order. However, this condition is not always satisfied. The goal of this example is to see what happens if the distribution P has finite moments up to a finite order $r+\delta$ i.e when there exists a finite number M such that $\mathbb{E}|X|^{r+\delta} < +\infty$ for $r+\delta < M$. For this, let us consider the hyper-Cauchy distribution $P = f.\lambda_d$ where

$$
f(x) = \frac{Cm}{(1+|x|^2)^m}
$$

for a finite constant $C > 0$ and $m > \frac{d}{2}$, this ensures the integrability of f w.r.t. the Lebesgue measure λ_d. This probability distribution has finite moments of order $r+\delta < 2m-d$, i.e. $\mathbb{E}|X|^{r+\delta} < +\infty$ if $r+\delta < 2m-d$.
In order to obtain Pierce type results, one proceeds as in the proof of Theorem 5.3.1. Criterion (5.16) is verified with

$$
g_\varepsilon(x) = \frac{K_{\delta,r}}{\left(1 \vee (1+\varepsilon)|x-a_1|\right)^{d\frac{r+\delta}{r+\delta-\eta}}}
$$

and the reasoning is the same until inequality (5.31). At this stage, since P does not have finite moments of any order, one wonders if the above inequality makes sense, i.e. if the integrals in the right side are finite. First, it is clear that

$$
\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f} \right)^{(1-q)q'} f d\lambda_d = \int \left(\frac{1+|x|^2}{1+\theta^2|x|^2} \right)^{m(1-q)q'} \frac{Cm}{(1+|x|^2)^m} d\lambda_d < +\infty \tag{5.33}
$$

where $1-q = \frac{d+r}{d+r-s}$ and q', p' are two conjugate coefficients larger than 1, since $\frac{1+|x|^2}{1+\theta^2|x|^2}$ is bounded for $\theta > 0$ and $\frac{Cm}{(1+|x|^2)^m} \in L^1(\lambda_d)$ as mentioned previously. Secondly, one notices that, since $\frac{r+\delta}{r+\delta-\eta} > 1$, then $|q|p' d\frac{r+\delta}{r+\delta-\eta} = |q|dp' + \eta'$ for some $\eta' > 0$. Hence, one can write

$$
\int_{\mathbb{R}^d} g_\varepsilon^{qp'} dP < +\infty \quad \Leftrightarrow \quad \int_{\mathbb{R}^d} \frac{|x|^{|q|dp'+\eta'}}{(1+|x|^2)^m} d\lambda_d(x) < +\infty \quad \Leftrightarrow \quad \int_0^{+\infty} \frac{|y|^{|q|dp'+\eta'+d-1}}{(1+|y|^2)^m} dy < +\infty.
$$

This is equivalent to

$$2m-(d|q|p'+\eta'+d-1)>1 \quad \Leftrightarrow \quad p'<\frac{1}{d|q|}(2m-d-\eta')<\frac{2m-d}{d|q|}.$$

At this stage, we note that one can choose p' as close to 1 as possible, since its conjugate q' can be chosen as large as possible without affecting (5.33). Hence, the above condition boils down to $d|q|<2m-d \quad \Leftrightarrow \quad \frac{s}{d+r-s}<\frac{2m-d}{d}$. Consequently, in order for this study to have sense, one must have

$$s<\Big(1-\frac{d}{2m}\Big)(d+r)$$

which is more restrictive than the condition $s<d+r$ in the case of distributions with finite moments of any order.

5.4 Upper estimates for L^r-optimal quantizers

Let $r,s>0$ and $(\Gamma^n)_{n\geq 1}$ a sequence of $L^r(\mathbb{R}^d)$-optimal quantizers of a random vector X with probability distribution P. For every $\mu\in\mathbb{R}^d$ and $\theta>0$, we denote $\Gamma^n_{\theta,\mu}=\mu+\theta(\Gamma^n-\mu)=\{\mu+\theta(x_i-\mu),\ x_i\in\Gamma^n,\ 1\leq i\leq n\}$.

In [71], the L^s-optimality of the sequence $(\Gamma^n_{\theta,\mu})_{n\geq 1}$ was studied. The author provided some conditions for the L^s-rate optimality of this sequence depending on whether Γ^n is an asymptotically L^r-optimal quantizer (study done for $s<r$) or exactly L^r-optimal (for $s<r+d$). This study was based on the integrability of the b-maximal functions associated to an L^r-optimal sequence of quantizers $(\Gamma^n)_{n\geq 1}$ defined by

$$\forall\xi\in\mathbb{R}^d,\quad \Psi_b(\xi)=\sup_{n\in\mathbb{N}}\frac{\lambda_d\left(B(\xi,b\,\mathrm{dist}(\xi,\Gamma^n))\right)}{P\left(B(\xi,b\,\mathrm{dist}(\xi,\Gamma^n))\right)}. \tag{5.34}$$

Throughout this section, we focus on the case where Γ^n is exactly L^r-optimal and $0<s<r+d$ and extend the results established in [71] to a larger class of distributions using tools that appeared meanwhile in [24]. Instead of maximal functions, our study relies on micro-macro inequalities using auxiliary probability distributions ν satisfying the following control on balls with respect to an $a_1\in\Gamma^n$: for every $\varepsilon\in(0,1)$, there exists a Borel function $g_\varepsilon:\mathbb{R}^d\to(0,+\infty)$ such that, for every $x\in\mathrm{supp}(\mathbb{P})$ and every $t\in[0,\varepsilon|x-a_1|]$,

$$\nu(B(x,t))\geq g_\varepsilon(x)V_d t^d. \tag{5.35}$$

where V_d denotes the volume of the hyper unit ball.

5.4.1 Main results

The case where $r<s$ and $(\Gamma^n)_{n\geq 0}$ is a sequence of L^r-asymptotically optimal quantizers of P has been studied in [71] without the use of maximal functions but requiring the couple (θ,μ) to be P-admissible, i.e. such that

$$\{f>0\}\subset\mu(1-\theta)+\theta\{f>0\}.$$

Note that if $\mathrm{supp}(P)=\mathbb{R}^d$, then every couple (θ,μ) is P-admissible. This condition is not needed to establish upper error bounds in this chapter but will be considered in the studies for $s<r$ in Section 5.5.

Theorem 5.4.1. *Let $s \in [r, d+r)$ and $1-q = \frac{d+r}{d+r-s}$. Let X be an $\mathbb{R}^d$-valued random vector with distribution $P = f.\lambda_d$ such that $\mathbb{E}|X|^{r+\delta} < +\infty$ for some $\delta > 0$ such that $r + \delta > \frac{sd}{d+r-s}$. Let $\eta \in (0, r+\delta - \frac{sd}{d+r-s})$, $p' = \frac{r+\delta-\eta}{d|q|}$ and $q' = \frac{r+\delta-\eta}{r+\delta-\eta-d|q|}$ and let $(\Gamma^n)_{n\geq 1}$ be a sequence of $L^r(\mathbb{R}^d)$-optimal quantizers of X. Assume*

$$\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{(1-q)q'} f d\lambda_d < +\infty.$$

Then, for every $n \geq 1$,

$$e_s(\Gamma^n_{\theta,\mu}, P) \leq \widetilde{\kappa}^{Optimal}_{\theta,\mu} \theta^{1+\frac{d}{s}} \sigma_{r+\delta}(P) \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{(1-q)q'} f d\lambda_d\right)^{\frac{1}{|q|q'(d+r)}} n^{-\frac{1}{d}}$$

where $\sigma_{r+\delta}(P) = \inf_a \|X-a\|_{r+\delta}$ is the $L^{r+\delta}$-standard deviation of P and

$$\widetilde{\kappa}^{Optimal}_{\theta,\mu} = 2^{\frac{2qp'-1}{qp'(d+r)}} \left(\frac{(2^r-1)C_1^r + 2^r C_2^r}{V_d}\right)^{\frac{1}{d+r}} \min_{\varepsilon \in (0,\frac{1}{3})} \left[(1+\varepsilon)\varphi_r(\varepsilon)^{-\frac{1}{d+r}}\right] \left(\int_{\mathbb{R}^d} (1 \vee |x|)^{-d\frac{r+\delta}{r+\delta-\eta}} dx\right)^{\frac{1}{d+r}}$$

with C_1 and C_2 are finite constants not depending on n, θ and μ and $\varphi_r : u \to \left(\frac{1}{3^r} - u^r\right) u^d$, $u \in (0, \frac{1}{3})$.

Remark 5.4.2. *One checks that φ_r attains its maximum at $\frac{1}{3}\left(\frac{d}{d+r}\right)^{\frac{1}{r}}$ on $(0, \frac{1}{3})$.*

Note that, like for greedy quantizeration sequences, the case $s < r$ can be easily treated by remarking that $e_s(\Gamma^n_{\theta,\mu}, P) \leq e_r(\Gamma^n_{\theta,\mu}, P)$ which is upper bounded in Theorem 5.4.1.

5.4.2 Proof

We start with a general theoretical result based on the auxiliary distribution ν and its companion function g_ε satisfying (5.35).

Theorem 5.4.3. *Let $s \in (0, d+r)$ and $1-q = \frac{d+r}{d+r-s}$. Let X be an $\mathbb{R}^d$-valued random vector with distribution $P = f.\lambda_d$ such that $\mathbb{E}|X|^{r+\delta} < +\infty$ for some $\delta > 0$ and let $(\Gamma^n)_{n\geq 1}$ be a sequence of $L^r(\mathbb{R}^d)$-optimal quantizers of X such that $\Gamma^n = \{x_1, \ldots, x_n\}$. Assume there exist a distribution ν and a function g_ε satisfying (5.35), for $\varepsilon \in (0, \frac{1}{3})$, such that*

$$\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|} dP_{\theta,\mu} < +\infty.$$

Then, for every $n \geq 1$,

$$e_s(\Gamma^n_{\theta,\mu}, P) \leq \kappa^{Optimal}_{\theta,\mu} \theta^{1+\frac{d}{d+r}} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|}(x) dP_{\theta,\mu}(x)\right)^{\frac{1}{|q|(d+r)}} n^{-\frac{1}{d}}$$

where $\kappa^{Optimal}_{\theta,\mu} = \left(4(2^r-1)C_1^r + 4.2^r C_2^r\right)^{\frac{1}{d+r}} V_d^{-\frac{1}{d+r}} \min_{\varepsilon\in(0,\frac{1}{3})} \left[\varphi_r(\varepsilon)^{-\frac{1}{d+r}}\right]$ with C_1 and C_2 finite constants not depending on n, θ and μ and $\varphi_r : u \to \left(\frac{1}{3^r} - u^r\right) u^d$, $u \in (0, \frac{1}{3})$.

Proof. First, as in the proof of Theorem 5.3.7, we have for every $n \geq 1$,

$$e_s(\Gamma^n_{\theta,\mu}, P)^s = \theta^s e_s(\Gamma^n, P_{\theta,\mu})^s. \tag{5.36}$$

Then, assume that $c \in (0, \frac{\varepsilon}{1-\varepsilon}] \cap (0, \frac{1}{2})$ so that $\frac{c}{c+1} \leq \varepsilon$. Moreover, $d(x, \Gamma^n) \leq |x - a_1|$ for an $a_1 \in \Gamma^n$. So, $\frac{c}{c+1} d(x, \Gamma^n) \leq \varepsilon |x - a_1|$ and, hence, ν satisfies (5.35) w.r.t. a_1. Consequently, there exists a Borel function $g_\varepsilon : \mathbb{R}^d \to (0, +\infty)$ such that

$$\nu\left(B\left(x, \frac{c}{c+1}\, d(x, \Gamma^n)\right)\right) \geq V_d \left(\frac{c}{c+1}\right)^d d(x, \Gamma^n)^d\, g_\varepsilon(x).$$

Then, noticing that $\frac{(1-c)^r - c^r}{(1+c)^r} \geq \frac{1}{3^r} - \left(\frac{c}{c+1}\right)^r > 0$ in (5.14), since $c \in (0, \frac{1}{2})$, yields

$$e_r(\Gamma^n, P)^r - e_r(\Gamma^{n+1}, P)^r \geq V_d\, \varphi_r\left(\frac{c}{c+1}\right) \int g_\varepsilon(x) d(x, \Gamma^n)^{d+r} dP(x)$$

where $\varphi_r(u) = \left(\frac{1}{3^r} - u^r\right) u^d$, $u \in (0, \frac{1}{3})$. This inequality is the version of (5.25) for optimal quantizers so we follow the same steps as in the proof of Theorem 5.3.7 until we obtain

$$e_r(\Gamma^n, P)^r - e_r(\Gamma^{n+1}, P)^r \geq C e_s(\Gamma^n, P_{\theta,\mu})^{d+r}$$

where $C = V_d\, \varphi_r\left(\frac{c}{c+1}\right) \theta^{-d} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|} dP_{\theta,\mu}(x)\right)^{\frac{1}{q}}$. At this stage, since $(\Gamma^n)_{n\geq 1}$ is a sequence of L^r-optimal quantizers, we use (5.13) to obtain the following upper bound

$$\begin{aligned}
e_s(\Gamma^n, P_{\theta,\mu}) \leq & C^{-\frac{1}{d+r}} \left(\frac{4(2^r-1) e_r(\Gamma^{n+1}, P)^r}{n+1} + \frac{4.2^r C_2^r n^{-\frac{r}{d}}}{n+1}\right)^{\frac{1}{d+r}} \\
\leq & \left(\frac{4(2^r-1) e_r(\Gamma^{n+1}, P)^r}{n+1} + \frac{4.2^r C_2^r n^{-\frac{r}{d}}}{n+1}\right)^{\frac{1}{d+r}} V_d^{-\frac{1}{d+r}}\, \varphi_r\left(\frac{c}{c+1}\right)^{-\frac{1}{d+r}} \theta^{\frac{d}{d+r}} \\
& \times \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|} dP_{\theta,\mu}(x)\right)^{\frac{1}{|q|(d+r)}} \\
\leq & \left(4(2^r-1) C_1^r + 4.2^r C_2^r\right)^{\frac{1}{d+r}} n^{-\frac{1}{d}} V_d^{-\frac{1}{d+r}}\, \varphi_r\left(\frac{c}{c+1}\right)^{-\frac{1}{d+r}} \theta^{\frac{d}{d+r}} \\
& \times \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|} dP_{\theta,\mu}(x)\right)^{\frac{1}{|q|(d+r)}}
\end{aligned} \tag{5.37}$$

where we used, in the last inequality, the definition of an L^r-optimal quantizer given by (5.10). Now, we use (5.28) to obtain

$$\begin{aligned}
e_s(\Gamma^n, P_{\theta,\mu}) \leq & \left(4(2^r-1) C_1^r + 4.2^r C_2^r\right)^{\frac{1}{d+r}} n^{-\frac{1}{d}} V_d^{-\frac{1}{d+r}}\, \theta^{\frac{d}{d+r}} \min_{\varepsilon \in (0, \frac{1}{3})} \left[\varphi_r(\varepsilon)^{-\frac{1}{d+r}}\right] \\
& \times \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f\, g_\varepsilon}\right)^{|q|} dP_{\theta,\mu}(x)\right)^{\frac{1}{|q|(d+r)}}.
\end{aligned}$$

Finally, one deduces the result by injecting this last inequality in (5.36). □

By specifying the function g_ε in Theorem 5.4.3, we obtain a universal non asymptotic bound for the error $e_s(\Gamma^n_{\theta,\mu}, P)$ given in Theorem 5.4.1 which proof is the following.

Proof of Theorem 5.4.1. We consider $\nu(dx) = \gamma_{r,\delta}(x)\lambda_d(dx)$ where

$$\gamma_{r,\delta}(x) = \frac{K_{\delta,r}}{(1 \vee |x - a_1|)^{d\frac{r+\delta}{r+\delta-\eta}}} \qquad \text{with} \qquad K_{\delta,r} = \left(\int \frac{dx}{(1 \vee |x|)^{d\frac{r+\delta}{r+\delta-\eta}}} \right)^{-1} < +\infty$$

is a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$ and $|\cdot|$ denotes any norm on $\mathbb{R}^d$. Similarly as in the proof of Theorem 5.3.1, (5.35) is verified with

$$g_\varepsilon(x) = \frac{K_{\delta,r}}{(1 \vee (1+\varepsilon)|x - a_1|)^{d\frac{r+\delta}{r+\delta-\eta}}}.$$

So, we apply Theorem 5.4.3 and use (5.31) to obtain

$$e_s(\Gamma^n_{\theta,\mu}, P) \le \kappa^{\text{Optimal}}_{\theta,\mu} \theta^{1+\frac{d}{s}} K_{\delta,r}^{-\frac{1}{d+r}} \left(\int_{\mathbb{R}^d} g_\varepsilon^{qp'} dP \right)^{\frac{1}{|q|p'(d+r)}} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f} \right)^{(1-q)q'} f d\lambda_d \right)^{\frac{1}{|q|q'(d+r)}} n^{-\frac{1}{d}}$$

where $q = \frac{-s}{d+r-s}$ so that $1 - q = \frac{d+r}{d+r-s}$ and p' and q' are two conjugate coefficients larger than 1. By our choice of g_ε and the $L^{r+\delta}$ Minkowski inequality,

$$\left(\int_{\mathbb{R}^d} g_\varepsilon^{qp'} dP \right)^{\frac{1}{|q|p'(d+r)}} \le K_{\delta,r}^{-\frac{1}{d+r}} \left(1 + (1+\varepsilon)\sigma_{r+\delta}(P) \right)^{\frac{1}{|q|p'(d+r)}}.$$

where $\sigma_{r+\delta}(P) = \inf_a \|X - a\|_{r+\delta}$ is the $L^{r+\delta}$-standard deviation of P. Consequently, one has

$$e_s(\Gamma^n_{\theta,\mu}, P) \le \kappa^{\text{Optimal}}_{\theta,\mu} \theta^{1+\frac{d}{s}} K_{\delta,r}^{-\frac{1}{d+r}} \left(1 + (1+\varepsilon)\sigma_{r+\delta}(P) \right)^{\frac{1}{|q|p'(d+r)}} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f} \right)^{(1-q)q'} f d\lambda_d \right)^{\frac{1}{|q|q'(d+r)}} n^{-\frac{1}{d}}$$

Now, we introduce an equivariance argument. For $\lambda > 0$, let $X_\lambda := \lambda(X - a_1) + a_1$ and $(\alpha_{\lambda,n})_{n\ge1} := (\lambda(\alpha_n - a_1) + a_1)_{n\ge1}$. It is clear that $e_r(\alpha^{(n)}, X) = \frac{1}{\lambda} e_r(\alpha^{(n)}_\lambda, X_\lambda)$. Plugging this in the previous inequality yields

$$e_s(\Gamma^n_{\theta,\mu}, P) \le \frac{\kappa^{\text{Optimal}}_{\theta,\mu}}{K_{\delta,r}^{\frac{1}{d+r}}} \theta^{1+\frac{d}{s}} \frac{1}{\lambda} \left(1 + (1+\varepsilon)\lambda\sigma_{r+\delta}(P) \right)^{\frac{1}{|q|p'(d+r)}} \left(\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f} \right)^{(1-q)q'} f d\lambda_d \right)^{\frac{1}{|q|q'(d+r)}} n^{-\frac{1}{d}}$$

Finally, one deduces the result by setting $\lambda = \frac{1}{(1+\varepsilon)\sigma_{r+\delta}(P)}$. □

5.5 More examples and a dilatation optimization

Let X be a random variable with distribution $P = f.\lambda_d$. The upper bounds established in Sections 5.3 and 5.4, induce that the quantizers $\Gamma^n_{\theta,\mu}$ and $a^{(n)}_{\theta,\mu}$ are $L^s(P)$-rate optimal under one of the following necessary and sufficient conditions depending on the value of s, as follows

$\triangleright$ If $s < r$ and (θ, μ) is P-admissible, then $a^{(n)}_{\theta,\mu}$ is $L^s(P)$-rate optimal iff P has finite moments of order $r + \delta$ for $\delta > 0$ and

$$\int f^{-\frac{s}{r-s}} f_{\theta,\mu}^{\frac{r}{r-s}} d\lambda_d < +\infty. \tag{5.38}$$

Note that it is the same condition for $\Gamma^n_{\theta,\mu}$ but this case is fully treated in [71].

$\triangleright$ If $s < r + d$, then the L^r-dilated greedy sequence $a^{(n)}_{\theta,\mu}$ and the L^r-dilated optimal sequence $\Gamma^n_{\theta,\mu}$ are $L^s(P)$-rate optimal iff P has finite moments of order $r + \delta$ for $\delta > 0$ and

$$\int_{\{f>0\}} \left(\frac{f_{\theta,\mu}}{f}\right)^{\frac{(d+r)(r+\delta-\eta)}{(d+r-s)(r+\delta-\eta)-ds}} f d\lambda_d < +\infty \tag{5.39}$$

where $\eta \in (0, r + \delta - \frac{sd}{d+r-s})$. In particular, when f is a radial non-increasing density, the L^r-dilated greedy sequence $a^{(n)}_{\theta,\mu}$ is $L^s(P)$-rate optimal iff P has finite moments of order $\frac{1-a}{a}(d + \varepsilon)$, $\varepsilon > 0$, and

$$\int f(x)^{\frac{-s(1+a)}{d+r-s}} f_{\theta,\mu}^{\frac{d+r}{d+r-s}}(x) d\lambda_d(x) < +\infty \tag{5.40}$$

where $a \in (0, 1)$.

This leads to determining the values of (θ, μ) for which these conditions are satisfied and hence obtain an interval $I_P(\theta, \mu)$ of the parameters for which the L^r-dilated sequence is L^s-optimal. Let us denote, for the sake of simplicity, $\alpha^{(n)}_{\theta,\mu}$ both sequences $(\Gamma^n_{\theta,\mu})_{n\geq 1}$ and $(a^{(n)}_{\theta,\mu})_{n\geq 1}$. Generally, μ is chosen to be equal to $\mathbb{E}[X]$ in order to ensure that the distribution $P_{\theta,\mu}$ lies in the same family of distributions of P, and the values of θ for which the above conditions are satisfied depend entirely on the density f of P. So, the problem is to determine the interval $I_P(\theta)$ depending on the distribution P. This way, based on L^r-optimal or greedy sequences $\alpha^{(n)}$, we obtain sequences $\alpha^{(n)}_{\theta,\mu}$ that are L^s-rate optimal, but not optimal nor even L^s-asymptotically optimal. We will carry out the study for specified families of distributions, like the multivariate Normal distribution $\mathcal{N}(m, \Sigma)$, the hyper-exponential, hyper-Gamma and hyper-Cauchy distributions. For each case, we determine the interval $I_P(\theta)$ and show that the dilated/contracted sequence does not satisfy the L^s-empirical measure Theorem for every $\theta \in I_P(\theta)$. However, the computations established allow us to determine, for some probability distributions, a particular value $\theta^* \in I_P(\theta)$ for which the sequence $\alpha^{(n)}_{\theta^*,\mu}$ satisfies the theorem. Let us first recall this Theorem.

Theorem 5.5.1 (Empirical measure Theorem)**.** *Let P be a L^r-Zador distribution, absolutely continuous w.r.t the Lebesgue measure on $\mathbb{R}^d$ with density f. Let Γ^n be an asymptotically optimal n-quantizer of P. Then, denoting $C_{f,r} = \int_{\mathbb{R}^d} f^{\frac{d}{d+r}} d\lambda_d$, one has*

$$\frac{1}{n} \sum_{x_i \in \Gamma^n} \delta_{x_i} \underset{n\to+\infty}{\Rightarrow} P_r = \frac{1}{C_{f,r}} \int f^{\frac{d}{r+d}} d\lambda_d, \tag{5.41}$$

or, in other words, for every $a, b \in \mathbb{R}^d$,

$$\frac{1}{n} \operatorname{card}\{x_i \in \Gamma^n \cap [a, b]\} \to \frac{1}{C_{f,r}} \int_{[a,b]} f^{\frac{d}{r+d}} d\lambda_d.$$

Moreover, for some distributions, the particular value θ^* mentioned above allows the lower bound (5.6) induced by $\alpha^{(n)}_{\theta^*,\mu}$ to attain the sharp constant in Zador's Theorem. This leads to wonder whether this sequence is L^s-asymptotically optimal.

Before proceeding with the particular studies, let us precise that, if $\theta > 1$, the sequence $\alpha^{(n)}_{\theta,\mu}$ is called a dilatation of $\alpha^{(n)}$ with scaling parameter θ and translating number μ. Likewise, if $\theta < 1$, the sequence $\alpha^{(n)}_{\theta,\mu}$ is called a contraction of $\alpha^{(n)}$ with scaling parameter θ and translating number μ.

5.5.1 The multivariate Gaussian distribution

Let $P = \mathcal{N}(m, \Sigma)$. We consider $\mu = m$ so that the distribution $P_{\theta,\mu}$ lies in the same family of distributions as P. Since $\mathrm{supp}(P) = \mathbb{R}^d$, then every couple (θ, μ) is P-admissible.
$\triangleright$ If $s < r$, the sequence $\alpha^n_{\theta,m}$ is L^s-rate optimal iff $\theta \in I_P(\theta) = \left(\sqrt{\frac{s}{r}}, +\infty\right)$. These computations are carried out in [71] for optimal quantizers and are the same for greedy quantizers.
$\triangleright$ If $r \leq s < d + r$, we lead two studies, relying first on condition (5.39) and then on condition (5.40)for radial densities and see what link we can make between both of them. Let us start with the general case, i.e. condition (5.39). For $q = \frac{-s}{d+r-s}$ and every $q' > 1$, one has

$$\int_{\{f>0\}} \left(\frac{f_{\theta,m}}{f}\right)^{(1-q)q'} f d\lambda = \left((2\pi)^d |\Sigma|\right)^{-\frac{1}{2}} \int e^{-\frac{1}{2}\left((1-q)q'\theta^2 + (q-1)q' + 1\right)(x-m)^2|\Sigma|^{-2}} dx.$$

So, the sequence $\alpha^{(n)}_{\theta,m}$ is L^s-rate optimal iff

$$(1-q)q'\theta^2 + (q-1)q' + 1 > 0 \quad \Leftrightarrow \quad \theta^2 > 1 - \frac{1}{q'(1-q)}$$

and this for every $q' > 1$. So, one can consider q' as close to 1 as possible and deduce that (5.39) is satisfied iff

$$I_P(\theta) = \left(\sqrt{\frac{s}{d+r}}, +\infty\right).$$

Now, since the Normal distribution is a radial density distribution, it is interesting to see what the condition (5.40) yields. For every $a \in (0,1)$, one has

$$\int_{\{f>0\}} f^{q(1-a)} f^{1-q}_{\theta,m} d\lambda_d = \left((2\pi)^d |\Sigma|\right)^{-\frac{1}{2}} \int e^{-\frac{1}{2}\left(q(1+a) + \theta^2(1-q)\right)(x-m)^2|\Sigma|^{-2}} d\lambda_d.$$

So, $\alpha^{(n)}_{\theta,m}$ is L^s-rate optimal iff

$$(1-q)\theta^2 + q(1+a) > 0 \quad \Leftrightarrow \quad \theta^2 > \frac{s}{d+r}(1+a)$$

and this for every $a \in (0,1)$. At this stage, note that the Normal distribution has finite r-th moment for every $r > 0$ so the moment assumption made in Theorem 5.3.5 allows us to choose a as small as possible in a way that, even if $\frac{(1-a)(d+\varepsilon)}{a}$ goes to infinity, we can still apply the theorem. Hence, one chooses $a \to 0^+$ and the condition made on θ reads $\theta^2 > \frac{s}{d+r}$ and the interval $I_P(\theta)$ becomes

$$I_P(\theta) = \left(\sqrt{\frac{s}{d+r}}, +\infty\right)$$

coinciding with the interval deduced from condition (5.39) as explained in Remark 5.3.6.

Remark 5.5.2. *One should note that choosing a scalar θ^* is optimal in the case of radial density probability distributions but, in the general case, it would be more precise if θ^* is a matrix.*

Empirical measure Theorem This study relies on the fact that the L^r-quantizers themselves satisfy the L^r-empirical measure Theorem so it is conducted only for L^r-dilated optimal quantizers $\Gamma^n_{\theta,m}$ since greedy quantizers do not satisfy this theorem. In order to conclude whether the sequence $\Gamma^n_{\theta,m}$ satisfies the empirical measure Theorem, we start by determining the "*limit measure*" of the empirical measure, i.e. determine the limit of $\frac{1}{n}\operatorname{card}\{x_i \in \Gamma^n_{\theta,m} \cap [a,b]\}$. For every $n \geq 1$, it is clear that

$$\{x_i \in \Gamma^n_{\theta,m} \cap [a,b]\} = \Big\{x_i \in \Gamma^n \cap \Big[\frac{a}{\theta}, \frac{b}{\theta}\Big]\Big\}.$$

So, since Γ^n satisfies the L^r-empirical measure Theorem,

$$\frac{1}{n}\text{card}\{x_i \in \Gamma^n_{\theta,m} \cap [a,b]\} \to \frac{1}{C_{f,r}} \int_{\left[\frac{a}{\theta}, \frac{b}{\theta}\right]} f^{\frac{d}{r+d}} d\lambda_d = \frac{1}{C_{f,r}} \theta^{-d} \int_{[a,b]} f\Big(\frac{x-m}{\theta} + m\Big)^{\frac{d}{r+d}} d\lambda_d$$

where $C_{f,r} = \int_{\mathbb{R}^d} f^{\frac{d}{d+r}} d\lambda_d$. For every $\theta \in I_P(\theta)$, one has

$$\int_{[a,b]} f\Big(\frac{x-m}{\theta} + m\Big)^{\frac{d}{r+d}} d\lambda_d = \left((2\pi)^d |\Sigma|\right)^{\frac{-d}{2(d+r)}} \int_{[a,b]} e^{-\frac{1}{2}\frac{d}{d+r}\theta^{-2}(x-m)^2|\Sigma|^{-2}} d\lambda_d$$

and

$$\int_{\mathbb{R}^d} f^{\frac{d}{d+r}} d\lambda_d = \left((2\pi)^d |\Sigma|\right)^{\frac{r}{2(d+r)}} \left(\frac{d+r}{d}\right)^{\frac{d}{2}}.$$

So, the limit of the empirical measure is given by

$$\begin{aligned}\frac{1}{n}\text{card}\{x_i \in \Gamma^n_{\theta,m} \cap [a,b]\} &\to \Big(\frac{d+r}{d\theta^2}\Big)^{\frac{d}{2}} \left((2\pi)^d |\Sigma|\right)^{-\frac{1}{2}+\frac{1}{2}\frac{d}{(d+r)\theta^2}} \int_{[a,b]} f^{\frac{d}{(d+r)\theta^2}} d\lambda_d \\ &= \frac{1}{\int_{\mathbb{R}^d} f^{\frac{d}{(d+r)\theta^2}} d\lambda_d} \int_{[a,b]} f^{\frac{d}{(d+r)\theta^2}} d\lambda_d.\end{aligned}$$

With this limit, one clearly does not find the limit needed to satisfy the empirical measure Theorem for every $\theta \in I_P(\theta)$. Instead, one can notice that it is possible for a particular value θ^* given by

$$\frac{d}{d+r}\theta^{*-2} = \frac{d}{d+s} \qquad \Leftrightarrow \qquad \theta^* = \sqrt{\frac{d+s}{d+r}}.$$

This leads to the following Proposition.

Proposition 5.5.3. *Let $r, s > 0$ and $P = \mathcal{N}(m, \Sigma)$ be a multivariate Normal distribution. Assume Γ^n is an asymptotically L^r-optimal quantizer of P. Consider*

$$\theta^* = \sqrt{\frac{d+s}{d+r}},$$

then the sequence $\Gamma^n_{\theta^*,m}$ satisfies the L^s-empirical measure Theorem, i.e.

$$\frac{1}{n}\operatorname{card}\{x_i \in \Gamma^n_{\theta^*,m} \cap [a,b]\} \to \frac{1}{C_{f,s}} \int_{[a,b]} f^{\frac{d}{s+d}} d\lambda_d.$$

This has been shown in [71] in addition to the fact that this particular θ^* minimizes the upper bound of the L^s-quantization error $e_s(\Gamma^{(n)}_{\theta,\mu}, P)$ induced by the L^r-dilated optimal quantizer of the Normal distribution. Moreover, the author has showed that, even if the lower bound (5.6) coincide with the sharp limiting constant in Zador's Theorem for this value of θ^*, the sequence $\Gamma^{(n)}_{\theta^*,m}$ is still not L^s-asymptotically optimal.

5.5.2 Hyper-exponential distributions

Let $X \sim P = f.\lambda_d$ where $f(x) = e^{-\lambda|x|^\alpha}$ for $\alpha, \lambda > 0$ and $|.|$ denotes a norm on $\mathbb{R}^d$. We consider $\mu = 0$ so that the distribution $P_{\theta,\mu}$ lies in the same family of distributions as P. Note that if one considers the density function $f(x) = e^{-\lambda|x-m|^\alpha}$ for $m \in \mathbb{R}$, the study will be the same since the quantities considered are invariant by translation. In other words, if Γ is an optimal quantizer of X, then $\Gamma - m(1,\dots,1)$ is an optimal quantizer for $X - m$. Moreover, it is clear that every couple (θ, μ) is P-admissible.
$\triangleright$ If $s < r$, one has

$$\int f^{-\frac{s}{r-s}}(x) f_{\theta,0}^{\frac{r}{r-s}}(x)dx = \int e^{-\frac{s\lambda|x|^\alpha}{s-r}} e^{-\frac{r\lambda|\theta x|^\alpha}{r-s}} = \int e^{-\lambda\left(\frac{s}{s-r}+\frac{r}{r-s}\theta^\alpha\right)|x|^\alpha}$$

So $\alpha^n_{\theta,0}$ is L^s-optimal iff (5.38) is satisfied which is clearly equivalent to $\theta^\alpha > \frac{s}{r}$. Hence, the interval $I_P(\theta)$ is equal to

$$I_P(\theta) = \left(\left(\frac{s}{r}\right)^{\frac{1}{\alpha}}, +\infty\right).$$

$\triangleright$ For $s \in (r, d+r)$, the idea is as follows. Just as for the Normal distribution, the hyper-Exponential distribution has finite moments of order r for every $r > 0$ so the moment assumption made in Theorem 5.3.5 allows us to choose a as small as possible and the condition (5.40) coincides with condition (5.39) as explained in Remark 5.3.6. Consequently, we will lead the study relying on (5.39). One has, for $q = \frac{-s}{d+r-s}$ and every $q' > 1$, that

$$\int \left(\frac{f_{\theta,0}}{f}\right)^{(1-q)q'} f d\lambda_d = \int e^{-\lambda\left((1-q)q'\theta^\alpha + (q-1)q' + 1\right)|x|^\alpha} d\lambda_d.$$

So $\alpha^n_{\theta,0}$ is L^s-optimal iff (5.39) is satisfied which is clearly equivalent to

$$(1-q)q'\theta^\alpha + (q-1)q' + 1 > 0 \quad \Leftrightarrow \quad \theta^\alpha > 1 - \frac{1}{q'(1-q)}$$

and this for every $q' > 1$. Hence, one can choose q' as small as possible, for example $q' \to 1^+$, yielding

$$I_P(\theta) = \left(\left(\frac{s}{d+r}\right)^{\frac{1}{\alpha}}, +\infty\right).$$

Empirical measure Theorem As explained in the previous example, this study is conducted for L^r-dilated optimal quantizers. As previously, we start by determining the limit of the empirical measure

$$\frac{1}{n}\text{card}\{x_i \in \Gamma^n_\theta \cap [a,b]\} \to \frac{1}{C_{f,r}} \int_{\left[\frac{a}{\theta}, \frac{b}{\theta}\right]} f^{\frac{d}{r+d}} d\lambda_d = \frac{1}{C_{f,r}} \theta^{-d} \int_{[a,b]} f\left(\frac{x}{\theta}\right)^{\frac{d}{r+d}} d\lambda_d$$

where $C_{f,r} = \int_{\mathbb{R}^d} f^{\frac{d}{d+r}} d\lambda_d$. For every $\theta \in I_P(\theta)$,

$$\int_{[a,b]} f(\theta^{-1}x)^{\frac{d}{r+d}} d\lambda_d = \int_{[a,b]} e^{-\lambda \frac{d}{d+r}\theta^{-\alpha}|x|^\alpha} = \int_{[a,b]} f(x)^{\frac{d}{(d+r)\theta^\alpha}} d\lambda_d.$$

Moreover, one uses the fact that

$$\int_{\mathbb{R}^d} f(|x|)dx = V_d \int_0^{+\infty} f(r) r^{d-1} dr \qquad \text{and} \qquad \int_0^{+\infty} x^n e^{-ax^b} dx = \frac{\Gamma\left(\frac{n+1}{b}\right)}{b a^{(n+1)/b}}, \tag{5.42}$$

where $V_d = V(B_d)$ is the volume of the hyper-unit ball on $\mathbb{R}^d$ and Γ is the Gamma function, to obtain

$$\int_{\mathbb{R}^d} f^{\frac{d}{d+r}} d\lambda_d = \int_{\mathbb{R}^d} e^{-\lambda \frac{d}{d+r}|x|^\alpha} d\lambda_d = V_d \frac{\Gamma(\frac{d}{\alpha})}{\alpha} \left(\lambda \frac{d}{d+r}\right)^{-\frac{d}{\alpha}}.$$

By the same arguments, one deduces that

$$\int_{\mathbb{R}^d} f(x)^{\frac{d}{(d+r)\theta^\alpha}} d\lambda = \frac{1}{\theta^d C_{f,r}}$$

so that the limiting measure is

$$\frac{1}{n}\text{card}\{x_i \in \Gamma_{\theta,m}^n \cap [a,b]\} \to \frac{1}{\int_{\mathbb{R}^d} f^{\frac{d}{(d+r)\theta^\alpha}} d\lambda_d} \int_{[a,b]} f^{\frac{d}{(d+r)\theta^\alpha}} d\lambda_d.$$

Consequently, we deduce that the sequence $\Gamma_{\theta,0}^{(n)}$ does not satisfy the empirical measure Theorem for every $\theta \in I_P(\theta)$ except for a particular value θ^* given by

$$\frac{d}{d+r}\theta^{*-\alpha} = \frac{d}{d+s} \qquad \Leftrightarrow \qquad \theta^* = \left(\frac{d+s}{d+r}\right)^{\frac{1}{\alpha}}$$

hence leading to the following Proposition

Proposition 5.5.4. *Let $r, s > 0$ and $P = f.\lambda_d$ where $f(x) = e^{-\lambda|x|^\alpha}$ for $\alpha, \lambda > 0$. Assume Γ^n is an asymptotically L^r-optimal quantizer of P. Consider*

$$\theta^* = \left(\frac{d+s}{d+r}\right)^{\frac{1}{\alpha}},$$

then the sequence $\Gamma_{\theta^*,0}^n$ satisfies the L^s-empirical measure Theorem, i.e.

$$\frac{1}{n}\,\text{card}\left\{x_i \in \Gamma_{\theta^*,0}^n \cap [a,b]\right\} \to \frac{1}{C_{f,s}} \int_{[a,b]} f^{\frac{d}{s+d}} d\lambda_d.$$

Note that θ^* does not depend on the parameter λ of the distribution, only on α. In the next proposition, we show that the sequence $\alpha_{\theta^*,0}^n$ satisfies the lower bound (5.6).

Proposition 5.5.5. *Let $r, s > 0$ and $P = f.\lambda_d$ where $f(x) = e^{-\lambda|x|^\alpha}$ for $\alpha, \lambda > 0$. Then, the asymptotic lower bound of the L^s-error of the sequence $\alpha_{\theta^*,0}^n$ with $\theta^* = \left(\frac{d+s}{d+r}\right)^{\frac{1}{\alpha}}$ satisfies*

$$Q_{r,s}^{\text{Inf}}(P, \theta^*) = Q_s(P)$$

where $Q_{r,s}^{\text{Inf}}(P, \theta^*) = (\theta^*)^{s+d} \widetilde{J}_{s,d} \left(\int f^{\frac{d}{d+r}} d\lambda_d\right)^{\frac{s}{d}} \int f^{-\frac{s}{d+r}}(x) f_{\theta^*,0}(x) dx$.

Proof. Elementary computations based on (5.42) show that

$$\int f^{-\frac{s}{d+r}}(x) f_{\theta^*,0}(x)dx = V_d \frac{\Gamma(\frac{d}{\alpha})}{\alpha\lambda^{-\frac{d}{\alpha}}} \left(\frac{d}{r+d}\right)^{-\frac{d}{\alpha}} \qquad \text{and} \qquad \int f^{\frac{d}{d+r}} d\lambda_d = V_d \frac{\Gamma(\frac{d}{\alpha})}{\alpha\lambda^{-\frac{d}{\alpha}}} \left(\frac{d}{r+d}\right)^{-\frac{d}{\alpha}}$$

so that

$$(\theta^*)^{s+d} \left(\int f^{\frac{d}{d+r}} d\lambda_d\right)^{\frac{s}{d}} \int f^{-\frac{s}{d+r}}(x) f_{\theta^*,0}(x)dx = \left(V_d \frac{\Gamma(\frac{d}{\alpha})}{\alpha} \lambda^{-\frac{d}{\alpha}}\right)^{1+\frac{d}{s}} \left(\frac{s+d}{d}\right)^{\frac{s+d}{\alpha}} = \left(\int f^{\frac{d}{d+s}} d\lambda_d\right)^{\frac{d+s}{d}}$$

and hence the result. □

It is interesting to see whether $\Gamma^{(n)}_{\theta^*,0}$ for $\theta^* = \left(\frac{s+d}{r+d}\right)^{\frac{1}{\alpha}}$ is L^s-asymptotically optimal. For this, we compute the upper bound of the L^s-quantization error $e_s(\Gamma^n_{\theta^*,0}, P)$ given in in Corollary 5.4.1 and see if it reaches the sharp constant in Zador's Theorem for the different values of s. Note that if $\alpha^{(n)}$ is a greedy quantization sequence, one cannot make any interesting conclusions since it is clear that the sharp Zador constant cannot be attained by our upper bounds.
Let $r, s > 0$ and Γ^n an L^r-optimal quantizer of P. Elementary computations based on (5.42) show that the upper bounds of the quantization error of P induced by $\Gamma^n_{\theta^*,0}$, for $\theta^* = \left(\frac{s+d}{r+d}\right)^{\frac{1}{\alpha}}$, are given by

$$Q^{\sup,\theta^*}_{r,s} = \begin{cases} \widetilde{J}^{\frac{1}{r}}_{r,d} \left(\int f^{\frac{d}{d+s}} d\lambda_d\right)^{\frac{d+s}{ds}} & \text{if } s < r, \\ \widetilde{\kappa}^{\text{Optimal}}_{\theta^*,m} \left(\frac{V_d \Gamma(\frac{d}{\alpha})}{\alpha\lambda^{\frac{d}{\alpha}}}\right)^{\frac{1}{s}} \left(\frac{s+d}{d}\right)^{\frac{d}{s\alpha}} \left(\frac{s+d}{r+d}\right)^{\frac{1}{\alpha}} & \text{if } r < s < d+r. \end{cases}$$

One can easily notice that, for the different values of s, $Q_s(P) \leq Q^{\sup,\theta^*}_{r,s}$. Consequently, no conclusions can be made on the L^s-asymptotically optimality of the sequence $(\Gamma^n_{\theta^*,0})_{n\geq 0}$. However, if we have $\widetilde{J}^{\frac{1}{s}}_{s,d}$ instead of $\widetilde{J}^{\frac{1}{r}}_{r,d}$, then one can reach Zador's sharp constant for $r < s$ and gets closer to it for $s \in (r, d+r)$.

5.5.3 Hyper-Gamma distributions

Let $X \sim P = f.\lambda_d$ where $f(x) = |x|^\beta e^{-\lambda|x|^\alpha}$ for $\alpha, \lambda > 0$ and $\beta > -d$ and $|\cdot|$ denotes any norm on $\mathbb{R}^d$. We consider $\mu = 0$ so that $P_{\theta,\mu}$ lies in the same family of distributions as P. In this case, every couple (θ, μ) is P-admissible since $\text{supp}(P) = \mathbb{R}^d$.
▷ If $s < r$, one has

$$\int f^{-\frac{s}{r-s}}(x) f^{\frac{r}{r-s}}_{\theta,0}(x)dx = \theta^{\frac{r\beta}{r-s}} \int |x|^\beta e^{-\lambda\left(\frac{s}{s-r} + \frac{r}{r-s}\theta^\alpha\right)|x|^\alpha}$$

So $\alpha^n_{\theta,0}$ is L^s-optimal iff (5.38) is satisfied which is clearly equivalent to $\theta^\alpha > \frac{s}{r}$. Consequently,

$$I_P(\theta) = \left(\left(\frac{s}{r}\right)^{\frac{1}{\alpha}}, +\infty\right).$$

▷ If $s < d+r$, the conditions (5.39) and (5.40) yield the same result as explained in Remark 5.3.6. For $q = \frac{-s}{d+r-s}$ and every $q' > 1$, one has

$$\int \left(\frac{f_{\theta,0}}{f}\right)^{(1-q)q'} f(x)d\lambda_d = \int |x|^\beta e^{-\lambda\left((1-q)q'\theta^\alpha + (q-1)q' + 1\right)|x|^\alpha} d\lambda_d.$$

So $\alpha^n_{\theta,0}$ is L^s-optimal iff

$$(1-q)q'\theta^\alpha + (q-1)q' + 1 > 0 \quad \Leftrightarrow \quad \theta^\alpha > 1 - \frac{1}{q'(1-q)}$$

and this for every $q' > 1$. Hence, one can choose q' as small as possible, for example $q' \to 1^+$, yielding

$$I_P(\theta) = \left(\left(\frac{s}{d+r}\right)^{\frac{1}{\alpha}}, +\infty \right).$$

Empirical measure Theorem As explained in the previous examples, this study is conducted for L^r-dilated optimal quantizers. First, we compute the limit

$$\frac{1}{n}\mathrm{card}\{x_i \in \Gamma^n_{\theta,0} \cap [a,b]\} \to \frac{1}{C_{f,r}} \int_{\left[\frac{a}{\theta}, \frac{b}{\theta}\right]} f^{\frac{d}{r+d}} d\lambda_d = \frac{1}{C_{f,r}} \theta^{-d} \int_{[a,b]} f\Big(\frac{x}{\theta}\Big)^{\frac{d}{r+d}} d\lambda_d$$

where $C_{f,r} = \int_{\mathbb{R}^d} f^{\frac{d}{d+r}} d\lambda_d$. For every $\theta \in I_P(\theta)$,

$$\int_{[a,b]} f(\theta^{-1}x)^{\frac{d}{r+d}} d\lambda_d = \theta^{-\frac{d\beta}{d+r}} \int_{[a,b]} |x|^{\frac{d\beta}{d+r}} e^{-\lambda \frac{d}{d+r} \frac{1}{\theta^\alpha} |x|^\alpha} d\lambda_d.$$

Moreover, using (5.42) yields

$$\int_{\mathbb{R}^d} f^{\frac{d}{d+r}} d\lambda_d = \int_{\mathbb{R}^d} |x|^{\frac{d\beta}{d+r}} e^{-\lambda \frac{d}{d+r}|x|^\alpha} d\lambda_d = V_d \frac{\Gamma(\frac{d+\frac{d\beta}{d+r}}{\alpha})}{\alpha} \left(\lambda \frac{d}{d+r}\right)^{-\frac{1}{\alpha}\left(d+\frac{d\beta}{d+R}\right)}.$$

Likewise, one obtains

$$\int_{\mathbb{R}^d} |x|^{\frac{\beta d(\theta^\alpha - 1)}{\theta^\alpha (d+r)}} f(x)^{\frac{d}{(d+r)\theta^\alpha}} d\lambda = C_{f,r} \theta^{d + \frac{d\beta}{d+r}}.$$

Consequently, the limiting measure is

$$\frac{1}{n}\mathrm{card}\{x_i \in \Gamma^n_{\theta,m} \cap [a,b]\} \to \frac{1}{\int_{\mathbb{R}^d} |x|^{\frac{\beta d(\theta^\alpha - 1)}{\theta^\alpha (d+r)}} f^{\frac{d}{(d+r)\theta^\alpha}} d\lambda_d} \int_{[a,b]} |x|^{\frac{\beta d(\theta^\alpha - 1)}{\theta^\alpha (d+r)}} f^{\frac{d}{(d+r)\theta^\alpha}} d\lambda_d.$$

Hence, in order for the sequence $\Gamma^{(n)}_{\theta,0}$ to satisfy the empirical measure Theorem, there are two conditions to fulfill

$$\frac{d}{(d+r)\theta^\alpha} = \frac{d}{d+s} \qquad \text{and} \qquad \frac{\beta d(\theta^\alpha - 1)}{\theta^\alpha (d+r)} = 0.$$

This is true for

$$\beta^* = \frac{d+r}{d(d+s)} \qquad \text{and} \qquad \theta^* = \left(\frac{d+s}{d+r}\right)^{\frac{1}{\alpha}}.$$

So, one can deduce with the following proposition.

Proposition 5.5.6. *Let $r, s > 0$ and $P = f.\lambda_d$ where $f(x) = |x|^\beta e^{-\lambda|x|^\alpha}$ for $\alpha, \lambda > 0$ and $\beta > -d$ and $|\cdot|$ is any norm on $\mathbb{R}^d$. Assume Γ^n is an asymptotically L^r-optimal quantizer of P. Consider*

$$\beta = \frac{d+r}{d(d+s)} \qquad \textit{and} \qquad \theta^* = \left(\frac{d+s}{d+r}\right)^{\frac{1}{\alpha}},$$

Normal Distribution		Exponential distribution		$P = f.\lambda_d$ with $f(x) = x^2 e^{-x^2}$	
n	Regression coefficient	n	Regression coefficient	n	Regression coefficient
255	0.9818	373	0.981	255	0.9399
511	0.9855	745	0.988	511	0.9405
1 023	0.9945	1 489	0.990	1 023	0.9406

Table 5.1: Regression coefficients of the optimally L^2-dilated greedy sequence on the L^3-optimal greedy sequence for $\mathcal{N}(0,1)$, $\mathcal{E}(1)$ and $P = f.\lambda_d$ with $f(x) = x^2 e^{-x^2}$.

then the sequence $\Gamma^n_{\theta^*,0}$ *satisfies the* L^s*-empirical measure Theorem, i.e.*

$$\frac{1}{n}\,\mathrm{card}\left\{x_i \in \Gamma^n_{\theta^*,0} \cap [a,b]\right\} \to \frac{1}{C_{f,s}} \int_{[a,b]} f^{\frac{d}{s+d}} d\lambda_d.$$

Note that one obtains the same results for the distribution with density $|x-m|^\beta e^{-\lambda|x-m|^\alpha}$ since it is invariant by translation.

Elementary computations, similar to those established previously, show that one cannot make any conclusions on the L^s-optimality of the L^r-dilated sequence considering the values of β and θ^* deduced in the previous proposition. In other words, one cannot know whether the lower and upper bound of the L^s-quantization error induced by $\alpha^n_{\theta^*,0}$ are equal or comparable to the sharp limiting constant $Q_s(P)$ in Zador's Theorem.

5.5.4 Numerical observations

We just showed that, for a particular value θ^*, the sequence $\alpha^{(n)}_{\theta^*,\mu}$ satisfies the L^s-empirical measure Theorem and that the lower bound of the L^s-quantization error induced by this sequence attains the sharp constant in Zador's Theorem, the upper bound only getting close. This pushes to conjecture that the optimally L^r-dilated sequence $(\alpha^n_{\theta^*,\mu})$ is asymptotically L^s-optimal. Numerical experiments were established in [71] to prove this conjecture numerically for optimal quantizers. In this section, we implement similar experiments to come to this type of conclusion for optimally L^r-dilated greedy quantization sequences. We denote $a^{r,(n)}$ the L^r-greedy quantization sequence.

Normal distribution We start with the Normal distribution $\mathcal{N}(0,1)$ and compute the corresponding L^3-optimal greedy quantization sequence $a^{3,(n)}$ by a standard Newton Raphson algorithm on one hand, and the optimally L^2-dilated greedy quantization sequence $a^{2,(n)}_{\theta^*,\mu}$ with $\theta^* = \sqrt{\frac{s+d}{r+d}} = \sqrt{\frac{4}{3}}$ and $\mu = 0$, on the other hand. We make a linear regression of the two resulting sequences for different values of the size n and expose, in table 5.1, the corresponding regression coefficients.

Exponential distribution We consider the exponential distribution $\mathcal{E}(1)$ with parameter $\lambda = 1$. In other words, it is the distribution studied in Example 5.5.2 for $d = 1$ and $\alpha = 1$. We compute the L^3-optimal greedy quantization sequence $a^{3,(n)}$ by a Newton Raphson algorithm and the optimally L^2-dilated greedy quantization sequence $a^{2,(n)}_{\theta^*,\mu}$ with $\theta^* = \left(\frac{s+d}{r+d}\right)^{\frac{1}{\alpha}} = \frac{4}{3}$ and $\mu = 0$.

The L^2-optimal greedy quantization sequence is obtained by a standard Lloyd's algorithm. We expose, in table 5.1, the regression coefficients obtained by regressing the L^2-dilated sequences on the L^3 greedy sequences.

Hyper-Gamma distribution Let $d = 1$. We consider the Hyper-Gamma probability distribution with parameters $\lambda = 1$ and $\alpha = \beta = 2$ so the density is given by

$$f(x) = x^2 e^{-x^2}.$$

In example 5.5.3, we showed that the hyper-Gamma distribution satisfy the L^s-empirical measure for a particular parameter β and a particular $\theta^* \in I_P(\theta)$. However, we conduct here the experiment for different values and see if one always have the same convergence of the regression coefficients to 1. We compute the L^3-optimal greedy quantization sequence $a^{3,(n)}$ by a Newton Raphson algorithm and the L^2-optimal greedy quantization sequence $a^{2,(n)}$ by a Lloyd's algorithm. The optimally L^2-dilated greedy sequence is given by $\alpha^{2,(n)}_{\theta^*,\mu}$ with $\theta^* = \left(\frac{s+d}{r+d}\right)^{\frac{1}{\alpha}} = \sqrt{\frac{4}{3}}$ and $\mu = 0$. Table 5.1 shows the regression coefficients obtained by regressing the L^2-dilated sequences on the L^3 greedy sequences where we observe a slower convergence, even a divergence of the coefficients to 1, hence deducing that this sequence cannot be L^s-asymptotically optimal.

Conjecture For the Normal and exponential distributions, the regression coefficient converges to 1 for specific values of n. This leads us to conjecture that there exists a sub-sequence of the greedy quantization sequence for which the regression coefficient converges to 1, i.e. for which the sequence is asymptotically L^s-optimal.
In fact, this "*subsequence*" topic has already been investigated in [24] where it has been shown (numerically) that there exist sub-optimal greedy quantization sequences, in the sense that the graphs representing the weights of the Voronoï cells converge towards the density curve of the distribution for certain sizes n of the sequence. For example, the greedy quantization sequence of $\mathcal{N}(0,1)$, and more generally of symmetrical distributions around 0, is sub-optimal and the optimal sub-sequence is of the form $a^{(n)} = a^{(2^k-1)}$ for $k \in \mathbb{N}^*$.
Hence, it is natural to conjecture that the optimally L^r-dilated sub-sequences of the same size are asymptotically L^s-optimal.

5.6 Application to numerical integration

Optimal quantizers and greedy quantization sequences are used in numerical probability where one relies on cubature formulas to approximate the exact value of $\mathbb{E}f(X)$, for a continuous bounded function f and a random variable X with distribution P, by

$$\mathbb{E}f(X) \approx \mathbb{E}f(\widehat{X}^{\alpha^{(n)}}) = \sum_{i=1}^{n} p_i^n f(\alpha_i^n) \tag{5.43}$$

where $\alpha^{(n)}$ designates the optimal or greedy quantization sequence of the random variable X and $p_i^n = P(X \in W_i(\alpha^{(n)}))$ represents the weight of the i^{th} Voronoï cell corresponding to $\alpha^{(n)}$ for every $i \in \{1, \dots, n\}$. A new iterative formula for the approximation of $\mathbb{E}f(X)$ using greedy quantization sequences is given in [24], based on the recursive character of greedy quantization.

Upper error bounds of these approximations have been investigated repeatedly in the literature, in [24, 56, 57] for example.

In this section, we present what advantages the dilated quantization sequences bring to the numerical integration field. This application was first introduced in [71] by A. Sagna for optimal quantizers. Here, we briefly recall his idea and emphasize that it also works with dilated greedy quantization sequences as well.

Let $X \in L^\beta, \beta \in (2, +\infty)$ and let f be a locally Lipschitz function, in the sense that, there exists a bounded constant $C > 0$ such that

$$|f(x) - f(y)| \leq C|x - y|(1 + |x|^{\beta-1} + |y|^{\beta-1}). \tag{5.44}$$

For every quantizer $\alpha^{(n)}$ (not necessarily stationary), one has, by applying Hölder's inequality with the conjugate exponents r and $r' = \frac{r}{r-1}$, that

$$\begin{aligned}|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{\alpha^{(n)}})| \leq \mathbb{E}|f(X) - f(\widehat{X}^{\alpha^{(n)}})| &\leq C\,\mathbb{E}\Big(|X - \widehat{X}^{\alpha^{(n)}}|\,(1 + |X|^{\beta-1} + |\widehat{X}^{\alpha^{(n)}}|^{\beta-1})\Big) \\ &\leq C\,\|X - \widehat{X}^{\alpha^{(n)}}\|_r \left(1 + \|X\|^{\beta-1}_{(\beta-1)r'} + \|\widehat{X}^{\alpha^{(n)}}\|^{\beta-1}_{(\beta-1)r'}\right).\end{aligned} \tag{5.45}$$

In order for this upper bound to make sense, one should have

$$(\beta - 1)r' = \frac{(\beta - 1)r}{r - 1} \leq \beta \qquad \Longleftrightarrow \qquad r \geq \beta > 2. \tag{5.46}$$

In practice, since most algorithms to optimize quantization (of n-tuples of greedy sequences) are much easier to implement in the quadratic case, it is more convenient to use such quadratic optimal or greedy quantizers in this type of applications to approximate expectations of the form $\mathbb{E}f(X)$. However, if we use L^2-quantizers $\alpha^{(n)}$ in our case, we obtain upper bounds involving an L^r-quantization error for $r > 2$ (see (5.46)) which is not really optimal since the quantizer used is not L^r-optimal for $r > 2$. So, an idea is to use L^2-dilated quantizers $\alpha^{(n)}_{\theta,\mu}$ which is itself L^r-rate optimal for given values of θ and μ depending on the probability distribution P. For example, if $X \sim \mathcal{N}(m, I_d)$, then one chooses $\mu = m$ and $\theta = \sqrt{\frac{r+d}{2+d}}$.

Hence, one approximates $\mathbb{E}f(X)$ by $\mathbb{E}f(\widehat{X}^{\alpha^{(n)}_{\theta,\mu}})$ rather than $\mathbb{E}f(\widehat{X}^{\alpha^{(n)}})$ via

$$\mathbb{E}f(\widehat{X}^{\alpha^{(n)}_{\theta,\mu}}) = \sum_{i=1}^{n} p_i^{\theta,\mu} f(\alpha_i^{\theta,\mu})$$

with $p_i^{\theta,\mu}$ being the weight of the i^{th} Voronoï cell corresponding to the quantization sequence $\alpha^{(n)}_{\theta^*,\mu}$ given by

$$P(X \in W_i(\alpha^{(n)}_{\theta^*,\mu})) = \int_{W_i(\alpha^{(n)}_{\theta^*,\mu})} f(x)d\lambda_d(x) = \theta^d \int_{W_i(\alpha^{(n)})} f_{\theta^*,\mu}(z)d\lambda_d(z) = P(\widehat{X}^{\alpha^{(n)}_{\theta^*,\mu}} \in W_i(\alpha^{(n)})) \tag{5.47}$$

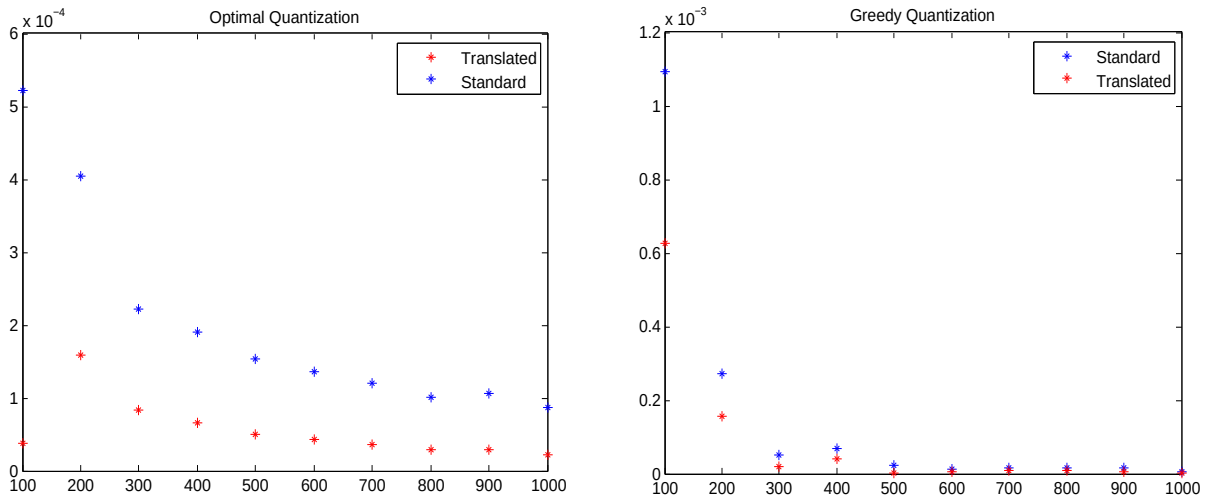

Figure 5.1: Errors of the approximation of $\mathbb{E}f(X)$, where $f(x) = x^4 + \sin(x)$, by quadrature formulas based on L^2 quantizers (blue) and dilated L^2 quantizers (red) for different sizes n.

where we applied the change of variables $z = \mu + \frac{x-\mu}{\theta}$. Then, since $\|X - \widehat{X}^{\alpha^{(n)}_{\theta,\mu}}\|_r$ converges faster to 0 than $\|X - \widehat{X}^{\alpha^{(n)}}\|_r$ for $r > 2$ if we consider an L^2-quantizer $\alpha^{(n)}$, one may expect to observe that

$$\left|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{\alpha^{(n)}_{\theta,\mu}})\right| \leq \left|\mathbb{E}f(X) - \mathbb{E}f(\widehat{X}^{\alpha^{(n)}})\right|.$$

To illustrate this numerically, we consider a one-dimensional example and approximate $\mathbb{E}f(X)$, where X is a random variable with Normal distribution $\mathcal{N}(0,1)$ and f is defined on $\mathbb{R}$ by $f(x) = x^4 + \sin(x)$ and satisfies (5.44) with $\beta = 5$. To satisfy (5.46), we choose $r = 5$ and implement the approximation by quadrature formulas based, on the one hand, on L^2-optimal and greedy sequences $\alpha^{(n)}$ and, on the other hand, on the L^2-dilated optimal and greedy quantizer $\alpha^{(n)}_{\theta^*,0}$, with $\theta^* = \sqrt{\frac{r+d}{2+d}} = \sqrt{2}$, which is L^r-rate optimal. The exact value of $\mathbb{E}f(X)$ is 3. In Figure 5.1, we illustrate the errors induced by these approximations and we observe that, for a same size n of the quantization sequence, the L^2-dilated quantizers $\alpha^{(n)}_{\theta^*,0}$ give more precise results than the standard sequences $\alpha^{(n)}$ themselves.

Chapter 6

Quantization-based approximation of reflected BSDEs with extended upper bounds for recursive quantization

Abstract We establish upper bounds for the L^p-quantization error, $p \in (1, 2+d)$, induced by the recursive Markovian quantization of a d-dimensional diffusion discretized via the Euler scheme. We introduce a *hybrid* recursive quantization scheme, easier to implement in the high-dimensional framework, and establish upper bounds of the corresponding L^p-quantization error. To take advantage of these extensions, we propose a time discretization scheme and a recursive quantization-based discretization scheme associated to a Reflected Backward Stochastic Differential Equation and estimate L^p-error bounds induced by the space approximation. We explain how to numerically compute the solution of the reflected BSDE relying on recursive quantization and compare it to others types of quantization.

6.1 Introduction

We are interested in the discretization and the computation of the solution of the following reflected backward stochastic differential equation RBSDE with maturity T

$$Y_t = g(X_T) + \int_t^T f(s, X_s, Y_s, Z_s)ds + K_T - K_t - \int_t^T Z_s.dW_s\,, \qquad t \in [0,T], \tag{6.1}$$

$$Y_t \geq h(t, X_t) \quad \text{and} \quad \int_0^T \big(Y_s - h(s, X_s)\big)dK_s = 0. \tag{6.2}$$

$(X_t)_{t\geq 0}$ is a Brownian diffusion process taking values in $\mathbb{R}^d$ and solution to the SDE

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s\,, \qquad X_0 = x_0 \in \mathbb{R}^d, \tag{6.3}$$

where the drift coefficient $b : [0,T]\times\mathbb{R}^d \to \mathbb{R}^d$ and the matrix diffusion coefficient $\sigma : [0,T]\times\mathbb{R}^d \to \mathcal{M}(d,q)$ are Lipschitz continuous in (t,x) so that $b(.,0)$ and $\sigma(.,0)$ are bounded on $[0,T]$ and satisfy the linear growth condition

$$\big\|\sigma(.,x)\big\| + \big\|b(.,x)\big\| \leq L_{b,\sigma}(1 + \|x\|)$$

with $L_{b,\sigma} = \max\left([b]_{\text{Lip}}, [\sigma]_{\text{Lip}}, \|b(.,0)\|_{\sup}, \|\sigma(.,0)\|_{\sup}\right)$ and $\|\cdot\|$ denoting any norm on $\mathbb{R}^d$. $(W_t)_{t\geq 0}$ is a q-dimensional Brownian motion defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ equipped with its augmented natural filtration $(\mathcal{F}_t)_{t\geq 0}$ where $\mathcal{F}_t = \sigma(W_s, s \leq t, \mathcal{N}_{\mathbb{P}})$, $\mathcal{N}_{\mathbb{P}}$ denotes the class of all $\mathbb{P}$-negligible sets of $\mathcal{A}$. The solution of this equation is defined as a $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+$-valued triplet (Y_t, Z_t, K_t) of $\mathcal{F}_t$-progressively measurable square integrable processes. K_t is continuous, non-decreasing, such that $K_0 = 0$ and grows exclusively on $\{t : Y_t = h(t, X_t)\}$. The driver $f(t, x, y, z) : [0, T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ is $[f]_{\text{Lip}}$-Lipschitz continuous with respect to (t, x, y, z), $g(X_T)$ is the terminal condition where $g : \mathbb{R}^d \to \mathbb{R}$ is $[g]_{\text{Lip}}$-Lipschitz continuous and $h : [0, T] \times \mathbb{R}^d \to \mathbb{R}$ is $[h]_{\text{Lip}}$-Lipschitz continuous such that $g \geq h$ for every t and x. Under these assumptions on b, σ, h, g and f, the RBSDE (6.1) and the SDE (6.3) admit both a unique solution. The existence of a process (Y_t, Z_t, K_t), solution of (6.1), was established in [25] where the authors also showed that this solution satisfies the following property

$$\left\| \sup_{t\in[0,T]} |Y_t| \right\|_{2p} \vee \|K_T\|_{2p} \vee \left\| \int_0^T |Z_t|^2 dt \right\|_p < \gamma_0 \tag{6.4}$$

for a finite constant γ_0 (see also [3]). In general, these solutions admit no closed form. Approximation schemes are needed to approximate them. In the literature, many authors studied different types of RBSDEs, for example, in [3, 18, 25, 48, 49] and many approximation schemes were investigated: Feynman-Kac type representation formula were given in [48] for the solutions of RBSDEs, a four step algorithm was developed in [50] to solve FBSDEs, a random time scheme in [2] and many more. In this chapter, we start by a time discretization scheme of the forward process $(X_t)_{t\in[0,T]}$, the Euler scheme, with the uniform mesh $t_k = k\Delta$, $k \in \{0, \ldots, n\}$, with $\Delta = \frac{T}{n}$. The discrete time Euler scheme $(\bar{X}_{t_k})_{0\leq k\leq n}$ associated to the process $(X_t)_{t\in[0,T]}$ is recursively defined by

$$\bar{X}^n_{t_{k+1}} = \bar{X}^n_{t_k} + \Delta b(t_k, \bar{X}^n_{t_k}) + \sigma(t_k, \bar{X}^n_{t_k})\Delta W_{t_{k+1}}, \qquad \bar{X}^n_{t_0} = X_0 = x_0 \in \mathbb{R}^d, \tag{6.5}$$

where $\Delta W_{t_{k+1}} = W_{t_{k+1}} - W_{t_k}$, for every $k \in \{0. \ldots, n-1\}$. This leads to consider the time discretization scheme $(\bar{Y}^n_t, \bar{\zeta}^n_t)$ associated to (Y_t, Z_t) given by the following backward recursion

$$\bar{Y}^n_T = g(\bar{X}^n_T) \tag{6.6}$$

$$\widetilde{Y}^n_{t_k} = \mathbb{E}(\bar{Y}^n_{t_{k+1}}|\mathcal{F}_{t_k}) + \Delta f\big(t_k, \bar{X}^n_{t_k}, \mathbb{E}(\bar{Y}^n_{t_{k+1}}|\mathcal{F}_{t_k}), \bar{\zeta}^n_{t_k}\big), \quad k = 0, \ldots, n-1, \tag{6.7}$$

$$\bar{\zeta}^n_{t_k} = \frac{1}{\Delta}\mathbb{E}\big(\bar{Y}^n_{t_{k+1}}(W_{t_{k+1}} - W_{t_k})\,|\,\mathcal{F}_{t_k}\big), \quad k = 0, \ldots, n-1, \tag{6.8}$$

$$\bar{Y}^n_{t_k} = \widetilde{Y}^n_{t_k} \vee h(t_k, \bar{X}^n_{t_k}), \quad k = 0, \ldots, n-1. \tag{6.9}$$

It is important to notice that, in this scheme, the conditional expectation is applied directly to $\bar{Y}^n_{t_{k+1}}$ inside the driver function f depending itself on the process Z_t (or $\bar{\zeta}^n_{t_k}$). This is slightly different of what have been already introduced and investigated in the literature. In fact, such schemes were considered for BSDE (without reflection) in [65] and for doubly reflected BSDE in [37], whereas in most chapters in the literature, the expectation is usually applied to the driver f from the outside. In some of these chapters devoted to time(-space) discretization of RBSDE, the driver does not depend on the process Z_t, (see [3, 4, 9, 48] for example).

After the time discretization, the solution of the scheme (6.6) – (6.7) – (6.8) – (6.9) still admits no closed form since it involves the computation of conditional expectations which cannot be obtained analytically. Therefore, we are led to devise a space discretization scheme to approximate it. In the literature, we can find various approaches: one can cite, among others, regression methods with Monte Carlo simulations (see [9]), the multi-step schemes methods (see [7]), a hybrid approach combining Picard iterates with a decomposition in Wiener chaos (see [13]), a connection with the semi-linear PDE associated to the BSDE (see [34]) and Monte Carlo simulations with Malliavin calculus (see [9, 17, 35]). Another approach is the optimal quantization introduced for RBSDEs in [5] and then developed in a

series of chapters ([3, 4, 37, 65] for example). In this chapter, we will rely on the recursive quantization of the time-discretized Euler scheme $(\bar{X}^n_{t_k})_{0\le k\le n}$. This method, originally introduced in [59] and then studied deeply in [51] and [63] for one-dimensional diffusions, consists in building a Markov chain having values into a *grid (or quantizer)* Γ_k of the discrete Euler scheme $\bar{X}_{t_k}$ at time t_k. The grids Γ_k can be optimized in a recursive way as a kind of *embedded* procedure.

In order to explain the principle of this recursive Markovian quantization, let us first recall briefly what optimal quantization is. Assume that $\mathbb{R}^d$ is equipped with a norm $\|\cdot\|$ (usually the canonical Euclidean norm for our purpose). Let $X \in L^p_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ and let $N \ge 1$ be a *quantization level*. The aim of L^p-optimal quantization is to find the best approximation of X in $L^p(\mathbb{P})$ by a random vector Y defined on $(\Omega, \mathcal{A}, \mathbb{P})$ taking at most N values. As a first step, we may consider the grid (or quantization grid) $\Gamma^N = Y(\Omega) = \{x_1, \ldots, x_N\}$ (with possibly repeated elements). One easily checks that, Γ^N being fixed, the best possible choice is given by a (Borel) nearest neighbor projection of X on Γ^N. It is called a Voronoï quantization of X defined by

$$\widehat{X}^{\Gamma^N} = \mathrm{Proj}_{\Gamma^N}(X) := \sum_{i=1}^{N} x_i \mathbb{1}_{C_i(\Gamma^N)}(X) \tag{6.10}$$

where $\big(C_i(\Gamma^N)\big)_{1\le i\le N}$ is a Borel partition of $\mathbb{R}^d$ satisfying

$$C_i(\Gamma^N) \subset \{\xi \in \mathbb{R}^d : \|\xi - x_i\| \le \min_{j\ne i} \|\xi - x_j\|\}, \qquad i = 1, \ldots, N. \tag{6.11}$$

The N-tuple $\big(C_i(\Gamma^N)\big)_{1\le i\le N}$ is called the *Voronoï partition* induced by Γ^N. The induced L^p-quantization error associated to the grid Γ^N is defined by

$$e_p(\Gamma^N, X) = \|X - \widehat{X}^{\Gamma^N}\|_p \tag{6.12}$$

where $\|.\|_p$ denotes the $L^p(\mathbb{P})$-norm. The optimal quantization problem boils down to finding the grid Γ^N that minimizes this error i.e. solving the problem

$$e_{p,N}(X) := \inf_{\Gamma, |\Gamma|\le N} e_p(\Gamma, X).$$

where $|\Gamma|$ denotes the cardinality of the grid Γ. A solution to this problem exists, as established in [32, 56, 57] for example, and is called an L^p-optimal quantization grid of (the distribution of) X. The corresponding quantization error converges to 0 as N goes to $+\infty$ and its rate of convergence is given by two well known results exposed in the following theorem.

Theorem 6.1.1. *(a)* Zador's Theorem (see [75])*: Let $X \in L^{p+\eta}_{\mathbb{R}^d}(\mathbb{P})$, $\eta > 0$, with distribution P such that $dP(\xi) = \varphi(\xi)d\lambda_d(\xi) + d\nu(\xi)$ where λ_d denotes the Lebesgue measure on $(\mathbb{R}^d, \mathcal{B}or(\mathbb{R}^d))$. Then,*

$$\lim_{N\to+\infty} N^{\frac{1}{d}} e_{p,N}(X) = \tilde{J}_{p,d} \|\varphi\|^{\frac{1}{p}}_{L^{\frac{p}{p+d}}(\lambda_d)} \tag{6.13}$$

where $\tilde{J}_{p,d} = \inf_{N\ge1} N^{\frac{1}{d}} e_{p,N}(\mathcal{U}([0,1]^d)) \in (0, +\infty)$.

(b) Extended Pierce's Lemma (see [44, 57])*: Let $p, \eta > 0$. There exists a constant $\kappa_{d,p,\eta} \in (0, +\infty)$ such that, for any random vector $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^d$,*

$$\forall N \ge 1, \quad e_{p,N}(X) \le \kappa_{d,p,\eta} \sigma_{p+\eta}(X) N^{-\frac{1}{d}} \tag{6.14}$$

where, for every $p \in (0, +\infty)$, $\sigma_p(X) = \inf_{a\in\mathbb{R}^d} \|X - a\|_p$ is the L^p-(pseudo-)standard deviation of X.

An important property, shared by quadratic optimal quantizers, is the stationarity property: an L^2-optimal quantizer Γ^N is said to be stationary if

$$\mathbb{E}(X|\widehat{X}^{\Gamma^N}) = \widehat{X}^{\Gamma^N}. \tag{6.15}$$

Let us now explain what recursive quantization is. If we define the Euler operator with step Δ by

$$\mathcal{E}_k(x, \varepsilon_{k+1}) = x + \Delta b(t_k, x) + \sqrt{\Delta}\sigma(t_k, x)\varepsilon_{k+1}$$

where $(\varepsilon_k)_{0\leq k\leq n}$ is an i.i.d. sequence of random variables with distribution $\mathcal{N}(0, I_q)$, then the recursive quantization $(\widehat{\bar{X}}_{t_k})_{0\leq k\leq n}$ of $(\bar{X}^n_{t_k})_{0\leq k\leq n}$ is defined by $\widehat{\bar{X}}_{t_0} = \bar{X}^n_{t_0} = x_0$ and

$$\begin{cases} \widetilde{X}_{t_k} &= \mathcal{E}_{k-1}(\widehat{X}^{\Gamma_{k-1}}_{t_{k-1}}, \varepsilon_k), \\ \widehat{X}^{\Gamma_k}_{t_k} &= \mathrm{Proj}_{\Gamma_k}(\widetilde{X}_{t_k}), \qquad \forall k = 1, \dots, n \end{cases} \tag{6.16}$$

where $(\Gamma_k)_{0\leq k\leq n}$ is a sequence of optimal quantizers of $(\widetilde{X}_{t_k})_{0\leq k\leq n}$ of size N_k, $k = 0, \dots, n$. The optimal quantizers $(\Gamma_k)_{1\leq k\leq n}$ can be either quadratic or L^p-optimal quantizers, we will detail the difference between these two frameworks later in the chapter. The main advantage of this method is that it preserves the Markov property of the Euler scheme with respect to the filtration $(\mathcal{F}_{t_k})_{0\leq k\leq n}$, the process $\widehat{X}_{t_k}$ is $\mathcal{F}_{t_k}$-measurable for every $k \in \{0, \dots, n\}$. In fact, the transition matrices $(p^k_{ij})_{1\leq i,j\leq N_k}$ where $p^k_{ij} = \mathbb{P}\big(\widehat{X}_{t_{k+1}} \in C_j(\Gamma_{k+1}) \,|\, \widehat{X}_{t_k} \in C_i(\Gamma_k)\big)$ and the initial distribution characterize the distribution of the Markov chain $(\widehat{X}_{t_k})_{k\geq 0}$, which was not the case with the optimal quantization in [63] for example. This Markov property will bring much help to carry on computations of the weights p^k_i of the Voronoï cells and the transition weights p^k_{ij}, as well as with the quantized scheme of the RBSDE itself.

Going back to our problem, we consider, in this chapter, the recursive quantization scheme associated to (6.6)-(6.7)-(6.8)-(6.9) based on the recursive quantization $(\widehat{X}_{t_k})_{0\leq k\leq n}$ of the Euler scheme $(\bar{X}^n_{t_k})_{0\leq k\leq n}$. It is defined recursively in a backward way as follows:

$$\widehat{Y}^n_T = g(\widehat{X}_T) \tag{6.17}$$

$$\widehat{\zeta}^n_{t_k} = \frac{1}{\Delta}\mathbb{E}\left(\widehat{Y}^n_{t_{k+1}}(W_{t_{k+1}} - W_{t_k}) \,|\, \mathcal{F}_{t_k}\right), \quad k = 0, \dots, n-1, \tag{6.18}$$

$$\widehat{Y}^n_{t_k} = \max\left(h_k(\widehat{X}_{t_k}), \mathbb{E}\left(\widehat{Y}^n_{t_{k+1}} \,|\, \mathcal{F}_{t_k}\right) + \Delta f\big(t_k, \widehat{X}_{t_k}, \mathbb{E}\big(\widehat{Y}^n_{t_{k+1}} \,|\, \mathcal{F}_{t_k}\big), \widehat{\zeta}^n_{t_k}\big)\right), \quad k = 0, \dots, n-1, \tag{6.19}$$

where $(\widehat{X}_{t_k})_{0\leq k\leq n}$ is the recursively quantized Euler scheme associated to $(\bar{X}^n_{t_k})_{0\leq k\leq n}$ given by (6.16). As a preliminary step, we are interested in estimating the L^p-quantization error $\|\widehat{X}_{t_k} - \bar{X}^n_{t_k}\|_p$, not only for $p = 2$ like in [63] but for any $p \in (1, 2+d)$. The fact that we are limited to $p < 2+d$ will become clear later in the chapter, as well as the type of optimal quantizers Γ_k of $\widetilde{X}_{t_k}$ needed to obtain satisfactory upper bounds for the L^p-quantization error. Note that in the quadratic case $p = 2$, the proof was based on a Pythagoras property which cannot be applied in a general framework. Furthermore, we introduce a kind of *hybrid* recursive quantization where the white noise $(\varepsilon_k)_{0\leq k\leq n}$ is replaced by its (already computed) quantized version $(\widehat{\varepsilon}_k)_{0\leq k\leq n}$.

In a second part, we will proceed with the time and space discretization of the RBSDE (6.1), as explained briefly before, and give more details about these schemes. We establish a priori estimates for the time discretization error $\|Y_{t_k} - \bar{Y}^n_{t_k}\|_2$ in a quadratic case. Although time discretization have already been studied in the literature (see [3, 9, 48, 65, 76]), our approach is still different mostly because of the combination of the reflection in the backward SDE and the conditional expectation applied directly to $\bar{Y}^n_{t_k}$ and $\widehat{Y}^n_{t_k}$ inside the driver f depending itself on the process Z_t (or its approximations). Likewise, estimates for the space discretization error $\|\bar{Y}^n_{t_k} - \widehat{Y}^n_{t_k}\|_p$ in L^p for $p \in (1, 2+d)$ will be established. To illustrate these theoretical results, we detail the numerical techniques available to compute the recursive

quantization $\widehat{X}^n_{t_k}$ of $\bar{X}^n_{t_k}$, for every $k \in \{1, \dots, \dots n\}$, their distributions and the corresponding transition weight matrices. Moreover, we will explain how to compute numerically the solution of the discretized scheme (6.17)-(6.18)-(6.19) associated to the RBSDE (6.1). These computations will be useful to carry on numerical tests and experiments illustrating the above error bounds. One of the most important applications of these quantization-based discretizations is the pricing of American options for which the driver f is equal to 0, among other examples (with a non-zero driver) that will be presented at the end of this chapter. This link between BSDEs and the pricing of financial options have been first introduced in [26].

Throughout this chapter, we will replace, for convenience, the indices t_k by k for $k \in \{0, \dots, n\}$, i.e. we will use, for example, $\widehat{X}_k$ instead of $\widehat{X}_{t_k}$. Also, we will replace $f(t_k, x, y, z)$ by $\mathcal{E}_k(x, y, z)$, $b(t_k, .)$ by $b_k(.)$ and $\sigma(t_k, .) = \sigma_k(.)$. And, we will omit the n in $\bar{Y}^n_{k+1}, \bar{X}^n_{k+1}$, etc.

This chapter is organized as follows: In section 6.2, we provide some short background on recursive quantization and establish the new L^p-error bounds for $p \in (1, 2+d)$, of the recursive quantization error as well as those of the *hybrid* recursive quantization error. Section 6.3 is devoted to the time discretization of the RBSDE and to the estimation of the corresponding error. The space disretization of the RBSDE will be treated in Section 6.4. In Section 6.5, we will present the numerical techniques to compute the recursive quantizers and the solution of the RBSDE. Finally, Section 6.6 is devoted to several numerical examples.

6.2 Recursive Quantization: background, L^p-error bounds and hybrid schemes.

In this section, we study the discretization of the forward process $(X_t)_{t\geq 0}$. It is a Brownian diffusion process taking values in $\mathbb{R}^d$, solution to the SDE (6.3) given in the introduction and recalled below

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s\,, \qquad X_0 = x_0 \in \mathbb{R}^d.$$

First, we start by the time discretization and we present the Euler scheme $(\bar{X}_{t_k})_{0\leq k\leq n}$, with uniform mesh $t_k = k\Delta$ for $k \in \{0, \dots, n\}$ and $\Delta = \frac{T}{n}$, associated to the process $(X_t)_{t\in[0,T]}$ which is recursively given by

$$\bar{X}_{t_{k+1}} = \bar{X}_{t_k} + \Delta b_k(\bar{X}_{t_k}) + \sigma_k(\bar{X}_{t_k})(W_{t_{k+1}} - W_{t_k}), \qquad \bar{X}_0 = X_0 = x_0, \tag{6.20}$$

where $W_{t_{k+1}} - W_{t_k} = \sqrt{\Delta}\varepsilon_{k+1}$, for every $k \in \{0, \dots, n-1\}$ and $(\varepsilon_k)_{0\leq k\leq n}$ is a sequence of i.i.d. random variables with distribution $\mathcal{N}(0, I_q)$. Its continuous counterpart, the *genuine Euler scheme*, is given by

$$d\bar{X}_t = b(\underline{t}, \bar{X}_{\underline{t}})dt + \sigma(\underline{t}, \bar{X}_{\underline{t}})dW_t \tag{6.21}$$

where $\underline{t} = t_k$ when $t \in [t_k, t_{k+1})$. This process satisfies for every $p \in (0, +\infty)$ and every $n \geq 1$, (see [10])

$$\Big\| \sup_{t\in[0,T]} X_t \Big\|_p + \sup_{n\geq 1} \Big\| \sup_{t\in[0,T]} \bar{X}_t \Big\|_p \leq C_{b,T,\sigma}(1 + |x_0|) \qquad \text{and} \qquad \Big\| \sup_{t\in[0,T]} |X_t - \bar{X}_t| \Big\|_p \leq C_{b,T,\sigma}\sqrt{\Delta}(1 + |x_0|)$$

where $C_{b,T,\sigma}$ is a positive constant depending on p, T, b and σ.

After the time discretization, one must proceed with space discterization schemes. As introduced, we consider in this chapter the approximation of the Euler scheme $(\bar{X}_{t_k})_{0\leq k\leq n}$ by recursive quantization.

6.2.1 Background

Our aim is to design, for $k \in \{0, \dots, n\}$, optimal quantizers Γ_k of size N_k of a function of the discrete Euler scheme $(\bar{X}_k)_{0 \le k \le n}$. So, the problem is to find the grid Γ_k that minimizes the L^p-distortion function $G_k^p(\Gamma) = \mathbb{E}\Big[\mathrm{dist}\big(\mathcal{E}_{k-1}(\bar{X}_{k-1}, \varepsilon_k), \Gamma\big)^p\Big]$ corresponding to $\mathcal{E}_{k-1}(\bar{X}_{k-1}, \varepsilon_k)$ where

$$\mathcal{E}_{k-1}(x, \varepsilon_k) = x + \Delta b_k(x) + \sqrt{\Delta}\sigma_k(x)\varepsilon_k$$

and $(\varepsilon_k)_k$ is an i.i.d. sequence of $\mathcal{N}(0, I_q)$-distributed random vectors independent from X_0.

Since $X_0 = \bar{X}_0 = x_0$ is fixed, its quantizer is given by $\Gamma_0 = \{x_0\}$. Then, we compute $\widetilde{X}_1 = F_0(\widehat{X}_0^{\Gamma_0}, \varepsilon_1)$ and we build an optimal quantization grid Γ_1 of size N_1 that minimizes $G_1^p(\widetilde{X}_1, \Gamma)$ on the set of grids Γ of size N_1 (see Section 6.5). Doing so, we are able to define the quantization of $\bar{X}_1$ by $\widehat{X}_1^{\Gamma_1} = \mathrm{Proj}_{\Gamma_1}(\widetilde{X}_1)$. Repeating this procedure, we define a(n optimized) recursive quantization of $(\bar{X}_k)_{0 \le k \le n}$ by the following recursion: $\widehat{X}_0 = \bar{X}_0 = x_0$ and

$$\left\{ \begin{array}{ll} \widetilde{X}_k & = \mathcal{E}_{k-1}(\widehat{X}_{k-1}^{\Gamma_{k-1}}, \varepsilon_k), \\ \widehat{X}_k^{\Gamma_k} & = \mathrm{Proj}_{\Gamma_k}(\widetilde{X}_k), \qquad \forall k = 1, \dots, n. \end{array} \right. \tag{6.22}$$

In practice, we ask the grids Γ_k to share some optimality properties, typically to be L^p-optimal or in higher dimension to be a product grid with optimal marginals, etc. For that purpose, the following identities play a crucial role: the L^p-distortion function associated to $\Gamma_k = (x_1^k, \dots, x_{N_k}^k)$ is approximated by

$$G_k^p(x_1^k, \dots, x_{N_k}^k) = \mathbb{E}[\mathrm{dist}(\widetilde{X}_k, \{x_1^k, \dots, x_{N_k}^k\})^p] = \sum_{i=1}^{N_k} \mathbb{E}[\mathrm{dist}(\mathcal{E}_{k-1}(x_i^{k-1}, \varepsilon_k), x_i^k)^p] \mathbb{P}\big(\widehat{X}_k^{\Gamma_k} \in C_i(\Gamma_k)\big) \tag{6.23}$$

where $\mathbb{P}\big(\widehat{X}_k^{\Gamma_k} \in C_i(\Gamma_k)\big)$ is the weight of the Voronoï cell of centroïd $x_i^k \in \Gamma_k$. Note that one can write the distortion function as a function of the grid Γ_k but writing it as a function of an N_k-tuple is needed to be able to talk of its differentiability. In fact, if the N_k-tuple $(x_1^k, \dots, x_{N_k}^k)$ has pairwise distinct components and the boundaries of the Voronoï diagram $\big(\partial C_i(\Gamma_k)\big)_{1 \le i \le N_k}$ are negligible w.r.t. the distribution of $\widetilde{X}_k$, then the gradient of the differentiable L^p-distortion function is given by

$$\nabla G_k^p(x_1^k, \dots, x_{N_k}^k) = p\Big(\mathbb{E}\big[\mathbb{1}_{\widetilde{X}_k \in C_i(\Gamma_k)}(x_i^k - \widetilde{X}_k)^{p-1}\big]\Big)_{1 \le i \le N_k}.$$

Note that since the grid Γ_k has pairwise distinct components for every $k \in \{0, \dots, n\}$, the distribution of $\widetilde{X}_k$ exists as soon as $\sigma\sigma^*$ is invertible. From now on, we denote $\widehat{X}_k$ instead $\widehat{X}_k^{\Gamma_k}$ for simplicity.

6.2.2 L^p-error bounds for recursive quantization

Our aim is to establish L^p-upper bounds for the recursive quantization error $\|\bar{X}_{t_k} - \widehat{X}_{t_k}\|_p$ for $p \in (1, 2+d)$ and $k \in \{0, \dots, n\}$. As explained, the recursive quantization schemes of $\bar{X}_{t_k}$ are based on optimal quantization sequences of $\widetilde{X}_{t_k}$ which can be either quadratic or L^p-quantization sequences, $p \neq 2$. The more interesting case is when we rely on L^2-optimal quantization because, from an algorithmic point of view, one has direct access to optimal quadratic quantizers since they are stationary and the algorithms used to produce optimal quantizers are either directly based on the stationarity property or easier to manage in a quadratic framework. Nevertheless, establishing an upper bound for the error $\|\bar{X}_{t_k} - \widehat{X}_{t_k}\|_p$ where $\widehat{X}_{t_k}$ is itself an L^p-optimal quantizer of $\widetilde{X}_{t_k}$ still seems a natural track to consider.

L^2-optimal quantization

We consider the case where, for every $k \in \{1, \dots, n\}$, $\widehat{X}_{t_k}$ is a quadratic optimal quantization of $\widetilde{X}_{t_k}$, hence it is stationary in the sense of (6.15) (see [57]). In the following, we assume that $\Delta \in [0, \Delta_{\max})$, $\Delta_{\max} > 0$. Note that for the Euler scheme, one can have $\Delta_{\max} = \frac{T}{n_0}$ if we consider schemes with step $\Delta = \frac{T}{n}$ and a number of steps $n > n_0$ for some $n_0 > 0$.

Theorem 6.2.1. *Let $p \in (1, 2+d)$, $(\bar{X}_k)_{0\le k\le n}$ defined by (6.20) and $(\widehat{X}_k)_{0\le k\le n}$ the corresponding recursive quantization sequence defined by (6.22). Assume that, for every $k \in \{0, \dots, n\}$, $\widehat{X}_k$ is a stationary quadratic optimal quantization of $\widetilde{X}_k$ of size N_k in the sense of (6.15), with $\widehat{X}_0 = \bar{X}_0 = x_0 \in \mathbb{R}^d$. For every $k \in \{1, \dots, n\}$ and every $\delta \in (0, 1]$,*

$$\|\bar{X}_k - \widehat{X}_k\|_p \le \left(\widetilde{K}_{d,2,2+\delta,p} \vee \kappa_{d,2,\delta}\right) \sum_{l=1}^{k} [\mathcal{E}_k]_{\mathrm{Lip}}^{k-l} C_{2+\delta,b,\sigma,T}(l)^{\frac{1}{2+\delta}} N_l^{-\frac{1}{d}}$$

where $\kappa_{d,2,\delta}$ is the constant from Pierce's Lemma 6.1.1(b),

$$\widetilde{K}_{d,2,2+\delta,p} \le 2^{\frac{p(2+\delta)}{(2+d)^2-dp}} V_d^{-\frac{1}{2+d}} \kappa_{X,2}^{\frac{1}{2+d}} \min_{\varepsilon\in(0,\frac{1}{3})} \left[(1+\varepsilon)\varphi_2(\varepsilon)^{-\frac{1}{d+2}}\right] \left(\int_{\mathbb{R}^d} \left(1 \vee \|x\|\right)^{-\frac{(d+2-p)(2+\delta)}{p}} dx\right)^{\frac{1}{2+d}}$$

with $\kappa_{X,r}$ a finite positive constant independent from N, V_d the volume of the hyper-unit ball and $\varphi_2(u) = \left(\frac{1}{3^2} - u^2\right)u^d$,

$$[\mathcal{E}_k]_{\mathrm{Lip}} = \begin{cases} e^{\Delta\left(s[b]_{\mathrm{Lip}} + c_s^{(1)} + c_{s,\Delta_{\max},\varepsilon_{k+1}}^{(3)} [\sigma]_{\mathrm{Lip}}^s\right)/p} & \text{if } p \in (1,2) \\ e^{\Delta\left(p[b]_{\mathrm{Lip}} + c_p^{(1)} + c_{p,\Delta_{\max},\varepsilon_{k+1}}^{(3)} [\sigma]_{\mathrm{Lip}}^p\right)/p} & \text{if } p \in [2, 2+d) \end{cases}$$

with $s = p + 1 > 2$, $c_p^{(1)} = 2^{(p-3)_+} \frac{(p-1)(p-2)}{2}$ and $c_{p,\Delta_{\max},\varepsilon_{k+1}}^{(3)} = 2^{(p-3)_+}(p-1)\mathbb{E}|\varepsilon_{k+1}|^p\left(1 + \frac{p}{2}\Delta_{\max}^{\frac{p}{2}-1}\right)$ and

$$C_{2+\delta,b,\sigma,T}(l) = e^{t_k(C_1+C_2)}|x_0|^{2+\delta} + \frac{C_3}{C_1+C_2}\left(e^{t_{k-1}(C_1+C_2)} - 1\right)$$

where C_1, C_2 and C_3 are defined in Lemma 6.2.4.

Before sharing the proof, we need to present some a priori useful results, mainly the distortion mismatch problem and two lemmas. We reconsider the notations where we replace the indices t_k by k to alleviate notations.

(L^r, L^s)-PROBLEM OR DISTORTION MISMATCH PROBLEM

Let $r, s \in (0, +\infty)$, the (L^r, L^s)-problem, also called distortion mismatch problem, consists in determining whether the optimal rate of L^r-optimal quantizers holds for L^s-quantizers for $s \neq r$, i.e. whether an L^r-optimal quantizer Γ_N of size N of a random vector X has an L^s-optimal convergence rate for $s \neq r$. For $s < r$, it is clear that an L^r-optimal quantizer is L^s-rate optimal due to the monotony of $r \to \|.\|_r$. When s becomes higher than r, we do not have such direct results. This problem was first introduced and treated in [32, 33] for radial density distributions on $\mathbb{R}^d$ and then generalized in [65] for all random vectors satisfying a certain moment condition. In the following theorem, we sum up this result and give a universal non-asymptotic Pierce type optimality result (in the sense of (6.14)).

Theorem 6.2.2 (Extended Pierce's Lemma)**.** *(a) Let $r > 0$ and X be an $\mathbb{R}^d$-valued random vector such that $\mathbb{E}|X|^{r'} < +\infty$ for some $r' > r$. Assume that its distribution $\mathbb{P}_X$ has a non-zero absolutely continuous component and let $(\Gamma_N)_{N\ge1}$ be a sequence of L^r-optimal quantizers of X. Then, for every $s \in \left(0, \frac{(d+r)r'}{d+r'}\right)$,*

$$e_s(\widehat{X}^{\Gamma_N}, X) \le \widetilde{K}_{d,r,r',s}\, \sigma_{r'}(X)\, N^{-\frac{1}{d}} \tag{6.24}$$

where $\sigma_{r'}(X) = \inf_{a\in\mathbb{R}^d} \|X-a\|_{r'}$ *is the* $L^{r'}$*-standard deviation of* X *and*

$$\widetilde{K}_{d,r,r',s} \le 2^{\frac{sr'}{(r+d)^2-ds}} V_d^{-\frac{1}{r+d}} \kappa_{X,r}^{\frac{1}{r+d}} \min_{\varepsilon\in(0,\frac{1}{3})} \Psi_r(\varepsilon) \left(\int \left(1\vee\|x\|\right)^{-\frac{(d+r-s)r'}{s}} dx\right)^{\frac{1}{r+d}}$$

with $\kappa_{X,r}$ *a finite positive constant independent from* N, V_d *the volume of the hyper-unit ball and* $\Psi_r(u) = (1+u)\left(\frac{1}{3^r}-u^r\right)^{-\frac{1}{d+r}} u^{-\frac{d}{d+r}}$.

(*b*) *In particular if* X *has finite polynomial moments at any order, then* (6.24) *is satisfied for every* $s\in(r,d+r)$ *and* $r'>\frac{sd}{d+r-s}$.

The following lemma is a technical one used repeatedly in the proofs in this chapter. Its proof will be postponed to the appendix.

Lemme 6.2.3. *Let* $r\in[2,+\infty)$ *and* $h_0>0$. *Let* $Z\in L^r_{\mathbb{R}^q}(\mathbb{P})$ *with* $\mathbb{E}\,Z=0$ *and let* $a\in\mathbb{R}^d$, $A\in\mathcal{M}(d,q,\mathbb{R})$. *Then for every* $h\in(0,h_0)$,

$$\mathbb{E}\left|a+\sqrt{h}AZ\right|^r \le |a|^r(1+c_r^{(1)}h) + c_{r,h_0}^{(2)}\, h\|A\|^r\mathbb{E}\,|Z|^r \tag{6.25}$$

where $c_r^{(1)} = 2^{(r-3)_+}\frac{(r-1)(r-2)}{2}$, $c_{r,h_0}^{(2)} = 2^{(r-3)_+}(r-1)\left(1+\frac{r}{2}h_0^{\frac{r}{2}-1}\right)$ *and* $\|A\|$ *is the operator norm.*

The following lemma is important for the proof of Theorem 6.2.1.

Lemme 6.2.4. *Consider* $(\bar{X}_k)_{0\le k\le n}$ *defined by* (6.20) *and* $(\widehat{X}_k)_{0\le k\le n}$ *its recursive quantization sequence defined by* (6.22). *Assume that, for every* $k\in\{0,\dots,n\}$, $\widehat{X}_k$ *is a stationary quadratic optimal quantization of* $\widetilde{X}_k$ *of size* N_k *in the sense of* (6.15), *with* $\widehat{X}_0=\bar{X}_0=x_0\in\mathbb{R}^d$. *For every* $r\ge 2$ *and every* $k\in\{1,\dots,n\}$,

$$\mathbb{E}|\widetilde{X}_k|^r \le e^{t_k(C_1+C_2)}|x_0|^r + \frac{C_3}{C_1+C_2}\left(e^{t_{k-1}(C_1+C_2)}-1\right). \tag{6.26}$$

where
$C_1 = rL_{b,\sigma} + (r-1)2^{r-2} + c_r^{(1)}$, $C_2 = 2^{r-1}L_{b,\sigma}^r\mathbb{E}|Z|^r\Delta_{\max}^r c_{r,\Delta_{\max}}^{(2)} := L_{b,\sigma}^r 2^{r-1}\Delta_{\max}^r c_{r,\Delta_{\max},Z}^{(3)}$ *and* $C_3 = C_2 + 2^{r-2}L_{b,\sigma}^r(1+r\Delta_{\max}^{r-1})(1+c_r^{(1)}\Delta_{\max})$ *with* $c_r^{(1)}$ *and* $c_{r,\Delta_{\max}}^{(2)}$ *defined in Lemma 6.2.3.*

Proof. The starting point is to use inequality (6.25) with $a=x+\Delta b(t,x)$ and $A=\sigma(t,x)$. On the one hand, we notice that

$$|a| \le |x| + \Delta L_{b,\sigma}(1+|x|) \le |x|(1+\Delta L_{b,\sigma}) + \Delta L_{b,\sigma}.$$

Then, using the fact that, for every $\varepsilon>0$,

$$\begin{aligned}
(\alpha+\beta)^r &\le \alpha^r + r\beta(\alpha+\beta)^{r-1}\\
&\le \alpha^r + r2^{r-2}\left((\varepsilon\alpha)^{r-1}\frac{\beta}{\varepsilon^{r-1}} + \beta^r\right)\\
&\le \alpha^r + r2^{r-2}\left(\beta^r + \frac{\varepsilon^r\alpha^r(r-1)}{r} + \frac{\beta^r}{r\varepsilon^{r(r-1)}}\right) \quad \text{(Young's inequality with } \tfrac{r}{r-1} \text{ and } r)\\
&\le \alpha^r\left(1+(r-1)2^{r-2}\varepsilon^r\right) + 2^{r-2}\beta^r\left(r+\frac{1}{\varepsilon^{r(r-1)}}\right),
\end{aligned} \tag{6.27}$$

one has, by considering $\alpha=|x|(1+\Delta L_{b,\sigma})$ and $\beta=\Delta L_{b,\sigma}$, that

$$|a|^r \le |x|^r(1+\Delta L_{b,\sigma})^r\left(1+(r-1)2^{r-2}\varepsilon^r\right) + 2^{r-2}\Delta^r L_{b,\sigma}^r\left(r+\frac{1}{\varepsilon^{r(r-1)}}\right).$$

On the other hand,

$$\|A\| \le \Delta L_{b,\sigma}(1+|x|) \quad \text{so that} \quad \|A\|^r \le 2^{r-1}\Delta^r L_{b,\sigma}^r(1+|x|^r).$$

Consequently, Lemma 6.2.3 yields

$$\begin{aligned}\mathbb{E}|a+A\sqrt{\Delta}Z|^r \leq & |x|^r(1+\Delta L_{b,\sigma})^r\left(1+(r-1)2^{r-2}\varepsilon^r\right)\left(1+c_r^{(1)}\Delta\right)+L_{b,\sigma}^r 2^{r-1}\mathbb{E}|Z|^r\Delta^{r+1}c_{r,\Delta_{\max}}^{(2)}|x|^r\\ &+\left(1+c_r^{(1)}\Delta\right)2^{r-2}L_{b,\sigma}^r\Delta^r\left(r+\frac{1}{\varepsilon^{r(r-1)}}\right)+L_{b,\sigma}^r 2^{r-1}\mathbb{E}|Z|^r\Delta^{r+1}c_{r,\Delta_{\max}}^{(2)}.\end{aligned}$$

At this stage, we are interested in considering a particular value of ε to avoid any explosion at infinity in the rest of the proof. The best choice (up to a multiplicative constant) is

$$\varepsilon=\Delta^{\frac{1}{r}}.$$

Now, we recall that $\Delta\in[0,\Delta_{\max})$, $\Delta_{\max}>0$ and denote

$$\begin{aligned}C_1:=&C_1(r)=rL_{b,\sigma}+(r-1)2^{r-2}+c_r^{(1)}\\ C_2:=&C_2(r,L_{b,\sigma},Z,\Delta_{\max})=2^{r-1}L_{b,\sigma}^r\mathbb{E}|Z|^r\Delta_{\max}^r c_{r,\Delta_{\max}}^{(2)}:=L_{b,\sigma}^r 2^{r-1}\Delta_{\max}^r c_{r,\Delta_{\max},Z}^{(3)}\\ C_3:=&C_3(r,Z,L_{b,\sigma},\Delta_{\max})=C_2+2^{r-2}L_{b,\sigma}^r(1+r\Delta_{\max}^{r-1})(1+c_r^{(1)}\Delta_{\max})\end{aligned}$$

Having $1+x\leq e^x$ yields

$$\mathbb{E}|a+A\sqrt{\Delta}Z|^r\leq|x|^re^{C_1\Delta}+\Delta\big(C_2|x|^r+C_3\big)\leq|x|^re^{C_1\Delta}\big(1+\Delta C_2e^{-\Delta C_1}\big)+\Delta C_3\leq e^{\Delta(C_1+C_2)}|x|^r+\Delta C_3.$$

Thus, since $\mathbb{E}|\widetilde{X}_k|^r=\mathbb{E}|\mathcal{E}_{k-1}(\widehat{X}_{k-1},\varepsilon_k)|^r$, one can write

$$\mathbb{E}|\widetilde{X}_k|^r\leq e^{\Delta(C_1+C_2)}\mathbb{E}|\widehat{X}_{k-1}|^r+\Delta C_3.$$

Using the fact that $\widehat{X}_{k-1}$ is a stationary quadratic optimal quantization of $\widetilde{X}_{k-1}$ and Jensen inequality yield

$$\mathbb{E}|\widehat{X}_{k-1}|^r=\mathbb{E}|\mathbb{E}(\widetilde{X}_{k-1}|\widehat{X}_{k-1})|^r\leq\mathbb{E}\big[\mathbb{E}\big(|\widetilde{X}_{k-1}|^r|\widehat{X}_{k-1}\big)\big]\leq\mathbb{E}|\widetilde{X}_{k-1}|^r.$$

Therefore,

$$\mathbb{E}|\widetilde{X}_k|^r\leq e^{\Delta(C_1+C_2)}\mathbb{E}|\widetilde{X}_{k-1}|^r+\Delta C_3.$$

Finally, it follows by induction that

$$\begin{aligned}\mathbb{E}|\widetilde{X}_k|^r\leq&e^{k\Delta(C_1+C_2)}\mathbb{E}|\widetilde{X}_0|^r+\Delta C_3\sum_{j=0}^{k-1}e^{j\Delta(C_1+C_2)}\\ \leq&e^{k\Delta(C_1+C_2)}|x_0|^r+\Delta C_3\frac{e^{(k-1)\Delta(C_1+C_2)}-1}{e^{\Delta(C_1+C_2)}-1}\\ \leq&e^{k\Delta(C_1+C_2)}|x_0|^r+\frac{C_3}{C_1+C_2}\left(e^{(k-1)\Delta(C_1+C_2)}-1\right).\end{aligned}$$

The result is obtained by noting that $k\Delta=k\frac{T}{n}=t_k$. □

Proof of Theorem 6.2.1. The first step of the proof is to show that the function $\mathcal{E}_k(.,\varepsilon_{k+1})$ is L^p-lipschitz continuous with Lipschitz coefficient $[\mathcal{E}_k]_{\text{Lip}}$ for every $k\in\{0,\dots,n-1\}$. We consider two cases depending on the values of p.

• If $p\in[2,2+d)$: For every $x,x'\in\mathbb{R}^d$,

$$\mathbb{E}\big|\mathcal{E}_k(x,\varepsilon_{k+1})-\mathcal{E}_k(x',\varepsilon_{k+1})\big|^p=\mathbb{E}\big|x-x'+\Delta\big(b_k(x)-b_k(x')\big)+\sqrt{\Delta}\varepsilon_{k+1}\big(\sigma_k(x)-\sigma_k(x')\big)\big|^p.$$

Since $p\geq2$, one applies Lemma 6.2.3 with $a=x-x'+\Delta\big(b_k(x)-b_k(x')\big)$ and $A=\sigma_k(x)-\sigma_k(x')$. We have

$$|a|^p\leq\big(|x-x'|+\Delta[b]_{\text{Lip}}|x-x'|\big)^p\leq|x-x'|^p\big(1+\Delta[b]_{\text{Lip}}\big)^p\leq|x-x'|^p\,e^{p\Delta[b]_{\text{Lip}}}$$

and

$$\|A\|^p \leq [\sigma]^p_{\mathrm{Lip}} |x - x'|^p.$$

At this stage, reusing the constants $c_p^{(1)} = 2^{(p-3)_+} \frac{(p-1)(p-2)}{2}$ and $c^{(3)}_{p,\Delta_{\max},\varepsilon_{k+1}} = 2^{(p-3)_+}(p-1)\mathbb{E}|\varepsilon_{k+1}|^p\big(1+\frac{p}{2}\Delta_{\max}^{\frac{p}{2}-1}\big)$ defined in Lemmas 6.2.3 and 6.2.4 yields

$$\begin{aligned}
\mathbb{E}|\mathcal{E}_k(x,\varepsilon_{k+1}) - \mathcal{E}_k(x',\varepsilon_{k+1})|^p &\leq \left(e^{\Delta(p[b]_{\mathrm{Lip}}+c_p^{(1)})} + \Delta[\sigma]^p_{\mathrm{Lip}} c^{(3)}_{p,\Delta_{\max},\varepsilon_{k+1}}\right)|x-x'|^p \\
&\leq |x-x'|^p e^{\Delta(p[b]_{\mathrm{Lip}}+c_p^{(1)})}\left(1 + \Delta[\sigma]^p_{\mathrm{Lip}} c^{(3)}_{p,\Delta_{\max},\varepsilon_{k+1}} e^{-\Delta(p[b]_{\mathrm{Lip}}+c_p^{(1)})}\right) \\
&\leq |x-x'|^p e^{\Delta(p[b]_{\mathrm{Lip}}+c_p^{(1)})}\left(1 + \Delta[\sigma]^p_{\mathrm{Lip}} c^{(3)}_{p,\Delta_{\max},\varepsilon_{k+1}}\right) \\
&\leq |x-x'|^p e^{\Delta\left(p[b]_{\mathrm{Lip}}+c_p^{(1)}+[\sigma]^p_{\mathrm{Lip}} c^{(3)}_{p,\Delta_{\max},\varepsilon_{k+1}}\right)}.
\end{aligned}$$

Consequently, $\mathcal{E}_k$ is L^p-lipschitz continuous with $[\mathcal{E}_k]_{\mathrm{Lip}} = e^{\Delta\left(p[b]_{\mathrm{Lip}}+c_p^{(1)}+[\sigma]^p_{\mathrm{Lip}} c^{(3)}_{p,\Delta_{\max},\varepsilon_{k+1}}\right)/p}$, for every $k \in \{1,\dots,n\}$ and $p \in [2, 2+d)$.

• If $1 < p < 2$: Consider $s = p+1 > 2$ so that $p - s < 0$. One has

$$\mathbb{E}|\mathcal{E}_k(x,\varepsilon_{k+1}) - \mathcal{E}_k(x',\varepsilon_{k+1})|^p = \mathbb{E}\left[|\mathcal{E}_k(x,\varepsilon_{k+1}) - \mathcal{E}_k(x',\varepsilon_{k+1})|^s\,|\mathcal{E}_k(x,\varepsilon_{k+1}) - \mathcal{E}_k(x',\varepsilon_{k+1})|^{p-s}\right].$$

On the one hand,

$$\begin{aligned}
|\mathcal{E}_k(x,\varepsilon_{k+1}) - \mathcal{E}_k(x',\varepsilon_{k+1})|^{p-s} &\leq |x-x'|^{p-s}\left(1 + \Delta[b]_{\mathrm{Lip}} + \sqrt{\Delta}\frac{|\sigma(x)-\sigma(x')|}{|x-x'|}|\varepsilon_{k+1}|\right)^{p-s} \\
&\leq |x-x'|^{p-s} e^{(p-s)\left(1+\Delta[b]_{\mathrm{Lip}}+\sqrt{\Delta}\frac{|\sigma(x)-\sigma(x')|}{|x-x'|}|\varepsilon_{k+1}|\right)} \quad (\text{since } 1+x \leq e^x) \\
&\leq |x-x'|^{p-s} \qquad (\text{ since } p-s<0).
\end{aligned}$$

On the other hand, one uses inequality (6.71) from the proof of Lemma 6.2.3 (see Appendix) and denotes $a = x - x' + \Delta[b]_{\mathrm{Lip}}(x-x')$ and $AZ = (\sigma(x)-\sigma(x'))\varepsilon_{k+1}$, to obtain

$$\begin{aligned}
|\mathcal{E}_k(x,\varepsilon_{k+1}) - \mathcal{E}_k(x',\varepsilon_{k+1})|^s \leq& |x-x' + \Delta[b]_{\mathrm{Lip}}(x-x') + \sqrt{\Delta}(\sigma(x)-\sigma(x'))\varepsilon_{k+1}|^s \\
\leq& |a|^s(1+\Delta c_s^{(1)}) + s\left(|a|^{s-1}\frac{a}{|a|}|A\sqrt{\Delta}Z\right) + \Delta c^{(2)}_{s,\Delta_{\max}}|AZ|^s.
\end{aligned}$$

At this stage, one notices that $|a|^s \leq |x-x'|^s(1+\Delta[b]_{\mathrm{Lip}})^s$ and that $|AZ|^s = [\sigma]^s_{\mathrm{Lip}}|x-x'|^s|\varepsilon_{k+1}|^s$. Then, using $1+x \leq e^x$, one deduces

$$\begin{aligned}
|\mathcal{E}_k(x,\varepsilon_{k+1}) - \mathcal{E}_k(x',\varepsilon_{k+1})|^s \leq& |x-x'|^s(1+\Delta c_s^{(1)})(1+\Delta[b]_{\mathrm{Lip}})^s + s\left(|a|^{s-1}\frac{a}{|a|}|A\sqrt{\Delta}Z\right) \\
&+ \Delta c^{(2)}_{s,\Delta_{\max}}[\sigma]^s_{\mathrm{Lip}}|x-x'|^s|\varepsilon_{k+1}|^s \\
\leq& |x-x'|^s e^{\Delta(c_s^{(1)}+s[b]_{\mathrm{Lip}})} + s\left(|a|^{s-1}\frac{a}{|a|}|A\sqrt{\Delta}Z\right) + \Delta c^{(2)}_{s,\Delta_{\max}}[\sigma]^s_{\mathrm{Lip}}|x-x'|^s|\varepsilon_{k+1}|^s.
\end{aligned}$$

Consequently, applying the expectation and keeping in mind that $\mathbb{E}|AZ| = 0$, we obtain

$$\begin{aligned}
\mathbb{E}|\mathcal{E}_k(x,\varepsilon_{k+1}) - \mathcal{E}_k(x',\varepsilon_{k+1})|^p \leq\ & e^{\Delta(c_s^{(1)}+s[b]_{\mathrm{Lip}})}|x-x'|^p + \Delta c^{(2)}_{s,\Delta_{\max}}[\sigma]^s_{\mathrm{Lip}}|x-x'|^p\mathbb{E}|\varepsilon_{k+1}|^s \\
&\leq |x-x'|^p e^{\Delta(c_s^{(1)}+s[b]_{\mathrm{Lip}})}\left(1 + \Delta c^{(2)}_{s,\Delta_{\max}}[\sigma]^s_{\mathrm{Lip}}\mathbb{E}|\varepsilon_{k+1}|^s e^{-\Delta(c_s^{(1)}+s[b]_{\mathrm{Lip}})}\right) \\
&\leq |x-x'|^p e^{\Delta(c_s^{(1)}+s[b]_{\mathrm{Lip}})}\left(1 + \Delta c^{(2)}_{s,\Delta_{\max}}[\sigma]^s_{\mathrm{Lip}}\mathbb{E}|\varepsilon_{k+1}|^s\right) \\
&\leq |x-x'|^p e^{\Delta\left(c_s^{(1)}+s[b]_{\mathrm{Lip}}+c^{(2)}_{s,\Delta_{\max}}[\sigma]^s_{\mathrm{Lip}}\mathbb{E}|\varepsilon_{k+1}|^s\right)}.
\end{aligned}$$

Consequently, $\mathcal{E}_k$ is L^p-Lipschitz continuous, for every $k \in \{1,\dots,n\}$ and $p \in (1,2)$, with $[\mathcal{E}_k]_{\mathrm{Lip}} = e^{\Delta\left(c_s^{(1)}+s[b]_{\mathrm{Lip}}+c_{s,\Delta_{\max}}^{(2)}[\sigma]_{\mathrm{Lip}}^s\mathbb{E}|\varepsilon_{k+1}|^s\right)/p}$.

For the second step, we first note that

$$\begin{aligned}\|\bar{X}_{k+1}-\widetilde{X}_{k+1}\|_p &= \|\mathcal{E}_k(\bar{X}_k,\varepsilon_{k+1})-\mathcal{E}_k(\widehat{X}_k,\varepsilon_{k+1})\|_p\\ &\le [\mathcal{E}_k]_{\mathrm{Lip}}\|\bar{X}_k-\widehat{X}_k\|_p\\ &\le [\mathcal{E}_k]_{\mathrm{Lip}}\|\bar{X}_k-\widetilde{X}_k\|_p+[\mathcal{E}_k]_{\mathrm{Lip}}\|\widetilde{X}_k-\widehat{X}_k\|_p.\end{aligned}$$

Then, we show by induction, since $\widehat{X}_0=\widetilde{X}_0$, that

$$\|\bar{X}_k-\widetilde{X}_k\|_p \le \sum_{l=1}^{k-1}[\mathcal{E}_k]_{\mathrm{Lip}}^{k-l}\|\widetilde{X}_l-\widehat{X}_l\|_p.$$

Consequently,

$$\|\bar{X}_k-\widehat{X}_k\|_p \le \|\bar{X}_k-\widetilde{X}_k\|_p+\|\widetilde{X}_k-\widehat{X}_k\|_p \le \sum_{l=1}^{k}[\mathcal{E}_k]_{\mathrm{Lip}}^{k-l}\|\widetilde{X}_l-\widehat{X}_l\|_p.$$

Now relying on the fact that $\widehat{X}_l$ is an L^2-optimal quantizer of $\widetilde{X}_l$ for every $l\in\{1,\dots,k\}$, we distinguish two cases: one the one hand, if $p\in(1,2)$, we use the monotony of $p\mapsto\|\cdot\|_p$ and Pierce's Lemma (6.14) to conclude that, for every $l\in\{1,\dots,k\}$,

$$\|\widetilde{X}_l-\widehat{X}_l\|_p \le \|\widetilde{X}_l-\widehat{X}_l\|_2 \le \kappa_{d,2,\delta}\|\widetilde{X}_l\|_{2+\delta}N_l^{-\frac{1}{d}},$$

for some $\delta>0$, and, on the other hand, if $p\in[2,2+d)$, we note that $\widetilde{X}_l=F_l(\widehat{X}_{l-1},\varepsilon_l)$ has finite polynomial moments at any order since the innovations $(\varepsilon_k)_{0\le k\le n}$ in the Euler operators are with Gaussian distribution and hence have finite polynomial moments at any order, so one uses section (b) of the distortion mismatch Theorem 6.2.2 to conclude that the quantization $\widehat{X}_l$ of $\widetilde{X}_l$ is L^p-rate optimal for every $p\in[2,2+d)$, in other words, we consider $\delta>0$ such that $r'=2+\delta>\frac{pd}{d+2-p}>2$ so that

$$\|\widetilde{X}_l-\widehat{X}_l\|_p \le \widetilde{K}_{d,2,2+\delta,p}\|\widetilde{X}_l\|_{2+\delta}N_l^{-\frac{1}{d}}.$$

Hence, for every $p\in(1,2+d)$,

$$\|\bar{X}_k-\widehat{X}_k\|_p \le \left(\widetilde{K}_{d,2,2+\delta,p}\vee\kappa_{d,2,\delta}\right)\sum_{l=1}^{k}[\mathcal{E}_k]_{\mathrm{Lip}}^{k-l}\|\widetilde{X}_l\|_{2+\delta}N_l^{-\frac{1}{d}}.$$

The result is obtained by plugging (6.26) for $r=2+\delta>2$ in this last inequality. □

Remark 6.2.5. *In higher dimensions, an approach to obtain the quantization grid of a multidimensional random variable is by taking the tensor product of one-dimensional quantization grids, that is the independent marginals of the distribution. The product quantization grid hence obtained by independent optimal one-dimensional quantizers is stationary and so this problem is solved in the multidimensional case. However, in most cases, the components of the diffusion X_t are not independent so this is not a very useful technique in practice.*

Remark 6.2.6. *We assume that $\widehat{X}_k$ is an L^p-optimal quantizer of $\widetilde{X}_k$ for every $k\in\{1,\dots,n\}$. What differs from L^2-optimal quantizers is that L^p-optimal quantizers are not usually stationary, a property*

that was very useful in the quadratic case. The beginning of the study is exactly similar to the quadratic framework until we obtain

$$\mathbb{E}|\widetilde{X}_k|^r \leq e^{\Delta(C_1+C_2)}\mathbb{E}|\widehat{X}_{k-1}|^r + \Delta C_3.$$

At this stage, one cannot use the stationarity property. Instead, applying inequality (6.27) yields

$$\mathbb{E}|\widehat{X}_{k-1}|^r \leq \mathbb{E}\big(|\widehat{X}_{k-1}-\widetilde{X}_{k-1}|+|\widetilde{X}_{k-1}|\big)^r \leq \mathbb{E}|\widehat{X}_{k-1}-\widetilde{X}_{k-1}|^r e^{C_4\Delta} + \mathbb{E}|\widetilde{X}_{k-1}|^r 2^{r-2}\Big(r+\frac{1}{\Delta^{r-1}}\Big)$$

where we took $\varepsilon = \Delta^{\frac{1}{r}}$ *and denoted* $C_4 = (r-1)2^{r-2}$*. Then,*

$$\mathbb{E}|\widetilde{X}_k|^r \leq \mathbb{E}|\widehat{X}_{k-1}-\widetilde{X}_{k-1}|^r e^{(C_1+C_2+C_4)\Delta} + e^{(C_1+C_2)\Delta}\mathbb{E}|\widetilde{X}_{k-1}|^r 2^{r-2}\Big(r+\frac{1}{\Delta^{r-1}}\Big) + \Delta C_3$$

and an induction yields

$$\begin{aligned}\mathbb{E}|\widetilde{X}_k|^r \leq & e^{k\Delta(C_1+C_2)}\left[2^{r-2}\Big(r+\frac{1}{\Delta^{r-1}}\Big)\right]^k \mathbb{E}|X_0|^r \\ & + \sum_{i=0}^{k} e^{(k-i)\Delta(C_1+C_2)}\left(\mathbb{E}|\widehat{X}_{k-1}-\widetilde{X}_{k-1}|^r e^{(C_1+C_2+C_4)\Delta} + \Delta C_3\right)\left[2^{r-2}\Big(r+\frac{1}{\Delta^{r-1}}\Big)\right]^k\end{aligned}$$

which clearly diverges as n goes to infinity. The fact that it seems impossible to get rid of the factor $\frac{1}{\Delta}$*, without the stationarity property, leads to conclude that we do not obtain satisfactory* L^p*-error bounds with a non-stationary* L^p*-optimal quantizer* $\widehat{X}_k$ *of* $\widetilde{X}_k$*. However, this is not really problematic since this is a very rare situation in practice because, as mentioned previously, one usually uses quadratic optimal quantizers for numerical purposes.*

6.2.3 Hybrid recursive quantization

When the dimension becomes greater than 1, computing the distribution (grids and transition matrices) of $(\widehat{X}_k)_{0\leq k\leq n}$ via the recursive formulas (6.22) cannot be achieved via closed formulas and deterministic optimization procedures. Multi-dimensional extensions can be found in [28] based on product quantization but this approach becomes computationally demanding when the dimension grows, an alternative being to implement a massive "embedded" Monte Carlo simulation. We propose here a third approach based on the quantization of the white noise (here a Gaussian one). This quantization can be part of a pre-processing and kept off line. In the case of a Gaussian noise, highly accurate quantization grids of $\mathcal{N}(0;I_q)$ distributions for dimensions $d=1$ up to 10 and regularly sampled sizes from $N=1$ to 1 000 can be downloaded from the quantization website www.quantize.maths-fi.com (for non-commercial purposes). In other words, we consider, instead of (6.22), the following recursive scheme

$$\left\{\begin{array}{ll}\widetilde{X}_k & = \mathcal{E}_{k-1}(\widehat{X}_{k-1},\widehat{\varepsilon}_k), \\ \widehat{X}_k & = \mathrm{Proj}_{\Gamma_k}(\widetilde{X}_k), \qquad \forall k=1,\ldots,n.\end{array}\right. \tag{6.28}$$

where $(\widehat{\varepsilon}_k)_k$ is now a sequence of optimal quantizers of the Normal distribution $\mathcal{N}(0,I_q)$, which are already computed and kept off line. The main advantage of this approach is that using quantization grids of small size N_k^{ε} approaching the Gaussian random vectors ε_k gives the same precision as a Monte Carlo simulation of much larger size, always having in mind that the optimal quantizers can be computed offline and called when needed. This is a great gain in cost.
In the following, we establish L^p-error bounds of this hybrid recursive quantization scheme, for $p\in(1,2+d)$, in terms of the error between $\widehat{X}_k$ and $\widetilde{X}_k$ and the quantization error between ε_k and $\widehat{\varepsilon}_k$ simultaneously. We recall that $\Delta\in[0,\Delta_{\max})$, $\Delta_{\max}>0$.

Theorem 6.2.7. *Let* $p\in(1,2+d)$ *and* $\delta>0$*. Consider* $(\bar{X}_k)_{0\leq k\leq n}$ *defined by (6.20) and* $(\widehat{X}_k)_{0\leq k\leq n}$ *its hybrid recursive quantization sequence defined by (6.28). Assume that, for every* $k\in\{0,\ldots,n\}$*,* $\widehat{X}_k$ *is a*

stationary L^2-optimal quantization of $\widetilde{X}_k$ of size N_k^X in the sense of (6.15) with $\widehat{X}_0 = \bar{X}_0 = x_0 \in \mathbb{R}^d$ and $(\widehat{\varepsilon}_k)_{0\le k\le n}$ an L^p-optimal quantization sequence of the Gaussian distributed sequence $(\varepsilon_k)_{0\le k\le n}$ of size N_k^ε. For every $k \in \{1,\dots,n\}$,

$$\|\bar{X}_k - \widehat{X}_k\|_p \le \big(\widetilde{K}_{d,2,2+\delta,p} \vee \kappa_{d,2,\delta}\big) \sum_{l=1}^{k} [F_k^x]_{\mathrm{Lip}}^{k-l} C_{2+\delta,b,\sigma,T}^{\frac{1}{2+\eta}} (N_l^X)^{-\frac{1}{d}} + \sum_{l=1}^{k-1} \kappa_{d,p,\eta} [F_k^\varepsilon]_{\mathrm{Lip}}^{k-l} \|\varepsilon_l\|_p (N_l^\varepsilon)^{-\frac{1}{d}}$$

where $\kappa_{d,2,\eta}$ is the constant given by Pierce's Lemma, $\widetilde{K}_{d,2,2+\delta,p}$ is given in Theorem 6.2.2,

$$C_{2+\delta,b,\sigma,T} = e^{t_k(C_1+C_2)}|x_0|^{2+\delta} + \frac{C_3}{C_1+C_2}\Big(e^{t_{k-1}(C_1+C_2)} - 1\Big)$$

with C_1, C_2 and C_3 are defined in Lemma 6.2.4,

$$[F_k^x]_{\mathrm{Lip}} = \begin{cases} e^{\frac{\Delta}{p}\left(c_p^{(1)} + L_{b,\sigma}\left(p + 2^{p-1} c_{p,\Delta_{\max}}^{(2)}\right)\right)} & \text{if } p \in [2, 2+d) \\ e^{\frac{\Delta}{p}\left(c_s^{(1)} + sL_{b,\sigma} + 2^{s-1} L_{b,\sigma}^s c_{s,\Delta_{\max}}^{(2)}\left(\mathbb{E}|\varepsilon|^s + \frac{p-s}{p}\right)\right)} & \text{if } p \in (1,2) \end{cases}$$

and

$$[F_k^\varepsilon]_{\mathrm{Lip}} = \begin{cases} \Delta^{\frac{1}{p}}\left(2^{p-1} c_{p,\Delta_{\max}}^{(2)} L_{b,\sigma}\right)^{\frac{1}{p}} & \text{if } p \in [2, 2+d) \\ \Delta^{\frac{1}{p}}\left(\frac{s}{p} 2^{s-1} c_{s,\Delta_{\max}}^{(2)} L_{b,\sigma}^s\right)^{\frac{1}{p}} & \text{if } p \in (1,2) \end{cases}$$

where $s = p+1$, $c_p^{(1)}$ and $c_{p,\Delta_{\max}}^{(2)}$ are defined in Lemma 6.2.3.

Proof. We start by showing that $\mathcal{E}_k$ is Lipschitz continuous with respect to its two variables. For every $x, x' \in \mathbb{R}^d$ and $\mathbb{R}^d$-valued r.v. ε and ε' with standard Normal distribution, we consider two cases depending on the values of p.

• If $p \in [2, 2+d)$: Always keeping in mind that $\Delta < \Delta_{\max}$, Lemma 6.2.3 yields

$$\begin{aligned} \mathbb{E}|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')|^p &= \mathbb{E}\big|x - x' + \Delta\big(b(x) - b(x')\big) + \sqrt{\Delta}\big(\sigma(x)\varepsilon - \sigma(x')\varepsilon'\big)\big|^p \\ &\le \big|x - x' + \Delta\big(b(x) - b(x')\big)\big|^p \big(1 + c_p^{(1)}\Delta\big) + \Delta c_{p,\Delta_{\max}}^{(2)} \mathbb{E}\big|\sigma(x)\varepsilon - \sigma(x')\varepsilon'\big|^p \\ &\le |x - x'|^p \big(1 + \Delta[b]_{\mathrm{Lip}}\big)^p \big(1 + c_p^{(1)}\Delta\big) + \Delta c_{p,\Delta_{\max}}^{(2)} \mathbb{E}\big|\sigma(x)\varepsilon - \sigma(x')\varepsilon'\big|^p \end{aligned}$$

where $c_p^{(1)}$ and $c_{p,\Delta_{\max}}^{(2)}$ are defined in Lemma 6.2.3. Now, noticing that $|\sigma(x)\varepsilon - \sigma(x')\varepsilon'| = |\sigma(x)\varepsilon - \sigma(x')\varepsilon + \sigma(x')\varepsilon - \sigma(x')\varepsilon'|$ and using $(a+b)^p \le 2^{p-1}(a^p + b^p)$ yield

$$\begin{aligned} \mathbb{E}|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')|^p &\le |x - x'|^p (1 + \Delta[b]_{\mathrm{Lip}})^p (1 + c_p^{(1)}\Delta) \\ &\quad + 2^{p-1} c_{p,\Delta_{\max}}^{(2)} \Delta \Big(\mathbb{E}|\sigma(x)\varepsilon - \sigma(x)\varepsilon'|^p + \mathbb{E}|\sigma(x)\varepsilon' - \sigma(x')\varepsilon'|^p\Big) \\ &\le |x - x'|^p \Big((1 + \Delta[b]_{\mathrm{Lip}})^p (1 + c_p^{(1)}\Delta) + 2^{p-1}\Delta c_{p,\Delta_{\max}}^{(2)} [\sigma]_{\mathrm{Lip}} \mathbb{E}|\varepsilon'|^p\Big) \\ &\quad + 2^{p-1}\Delta c_{p,\Delta_{\max}}^{(2)} \|\sigma\|_\infty \mathbb{E}|\varepsilon - \varepsilon'|^p. \end{aligned}$$

Now, using the fact that $1 + x \le e^x$ yields

$$\mathbb{E}|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')|^p \le e^{\bar{C}\Delta}|x - x'|^p + \Delta\widetilde{C}\mathbb{E}|\varepsilon - \varepsilon'|^p$$

where $\bar{C} = p[b]_{\mathrm{Lip}} + c_p^{(1)} + 2^{(p-3)_+ + p-1}(p-1)\big(1 + \frac{p}{2}\Delta_{\max}^{\frac{p}{2}-1}\big)[\sigma]_{\mathrm{Lip}}$ and $\widetilde{C} = 2^{(p-3)_+ + p-1}(p-1)\big(1 + \frac{p}{2}\Delta_{\max}^{\frac{p}{2}-1}\big)\|\sigma\|_\infty$. Then, applying $(a+b)^{\frac{1}{p}} \le a^{\frac{1}{p}} + b^{\frac{1}{p}}$ for $a, b > 0$ and $p > 1$ yields

$$\|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')\|_p \le e^{\frac{\bar{C}\Delta}{p}}\|x - x'\|_p + (\Delta\widetilde{C})^{\frac{1}{p}}\|\varepsilon - \varepsilon'\|_p.$$

Consequently, $\mathcal{E}_k$ is Lipschitz continuous for $k \in \{1, \dots, n\}$ and for $p \in [2, 2+d)$ with Lipschitz coefficients $[F_k^x]_{\text{Lip}} \leq e^{\Delta \bar{C}/p}$ and $[F_k^\varepsilon]_{\text{Lip}} \leq (\Delta \widetilde{C})^{\frac{1}{p}}$.

• If $p \in (1,2)$: Consider $s = p + 1 > 2$ so that $p - s < 0$. One has

$$\mathbb{E}|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')|^p = \mathbb{E}\left[|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')|^s \, |\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')|^{p-s}\right].$$

On the one hand,

$$\begin{aligned}
|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')|^{p-s} &\leq |x-x'|^{p-s}\left(1 + \Delta[b]_{\text{Lip}} + \sqrt{\Delta}\frac{|\sigma(x)\varepsilon - \sigma(x')\varepsilon'|}{|x-x'|}\right)^{p-s} \\
&\leq |x-x'|^{p-s} e^{(p-s)\left(1+\Delta[b]_{\text{Lip}} + \sqrt{\Delta}\frac{|\sigma(x)\varepsilon - \sigma(x')\varepsilon'|}{|x-x'|}\right)} \quad (\text{since } 1+x \leq e^x) \\
&\leq |x-x'|^{p-s} \qquad (\text{ since } p-s<0).
\end{aligned}$$

On the other hand, using inequality (6.71) from the proof of Lemma 6.2.3 (see Appendix), and noting $a = x - x' + \Delta[b]_{\text{Lip}}(x-x')$ and $AZ = \sigma(x)\varepsilon - \sigma(x')\varepsilon'$, yields

$$\begin{aligned}
|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')|^s \leq& |x - x' + \Delta[b]_{\text{Lip}}(x-x') + \sqrt{\Delta}(\sigma(x)\varepsilon - \sigma(x')\varepsilon')|^s \\
\leq& |a|^s(1 + \Delta c_s^{(1)} + s\left(|a|^{s-1}\frac{a}{|a|}|A\sqrt{\Delta}Z\right) + \Delta c_2^{(s,\Delta_{\max})}|AZ|^s.
\end{aligned}$$

At this stage, one notices that $|a|^s \leq |x-x'|^s(1+\Delta[b]_{\text{Lip}})^s$ and that

$$|AZ| = |\sigma(x)\varepsilon - \sigma(x')\varepsilon'| \leq |\sigma(x)\varepsilon - \sigma(x')\varepsilon| + |\sigma(x')\varepsilon - \sigma(x')\varepsilon'| \leq [\sigma]_{\text{Lip}}|x-x'||\varepsilon| + |\sigma(x')||\varepsilon - \varepsilon'|,$$

so that

$$|AZ|^s \leq 2^{s-1}\left([\sigma]_{\text{Lip}}^s|x-x'|^s|\varepsilon|^s + \|\sigma\|_\infty^s|\varepsilon-\varepsilon'|^s\right).$$

Hence, since $1 + x \leq e^x$,

$$\begin{aligned}
|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')|^s \leq& |x-x'|^s(1+\Delta c_s^{(1)})(1+\Delta[b]_{\text{Lip}})^s + s\left(|a|^{s-1}\frac{a}{|a|}|A\sqrt{\Delta}Z\right) \\
&+ \Delta c_{s,\Delta_{\max}}^{(2)} 2^{s-1}\left([\sigma]_{\text{Lip}}^s|x-x'|^s|\varepsilon|^s + \|\sigma\|_\infty^s|\varepsilon-\varepsilon'|^s\right) \\
\leq& |x-x'|^s e^{\Delta(c_s^{(1)} + s[b]_{\text{Lip}})} + s\left(|a|^{s-1}\frac{a}{|a|}|A\sqrt{\Delta}Z\right) \\
&+ \Delta c_{s,\Delta_{\max}}^{(2)} 2^{s-1}\left([\sigma]_{\text{Lip}}^s|x-x'|^s|\varepsilon|^s + \|\sigma\|_\infty^s|\varepsilon-\varepsilon'|^s\right).
\end{aligned}$$

Consequently, applying the expectation and keeping in mind that $\mathbb{E}|AZ| = 0$, we obtain

$$\begin{aligned}
\mathbb{E}|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')|^p \leq& e^{\Delta(c_s^{(1)} + s[b]_{\text{Lip}})}\mathbb{E}|x-x'|^p \\
&+ \Delta c_{s,\Delta_{\max}}^{(2)} 2^{s-1}\left([\sigma]_{\text{Lip}}^s\mathbb{E}[|x-x'|^p|\varepsilon|^s] + \|\sigma\|_\infty^s\mathbb{E}[|\varepsilon-\varepsilon'|^s|x-x'|^{p-s}]\right).
\end{aligned}$$

Using the fact that ε is independent of $\{x, x'\}$ and applying Young inequality with the conjugate exponents $\frac{p}{s}$ and $\frac{p}{p-s}$ to $\mathbb{E}[|\varepsilon-\varepsilon'|^s|x-x'|^{p-s}]$ yields

$$\begin{aligned}
\mathbb{E}|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')|^p \leq& \mathbb{E}|x-x'|^p\left(e^{\Delta(c_s^{(1)} + s[b]_{\text{Lip}})} + \Delta c_{s,\Delta_{\max}}^{(2)} 2^{s-1}[\sigma]_{\text{Lip}}^s\mathbb{E}|\varepsilon|^s\right) \\
&+ \Delta c_{s,\Delta_{\max}}^{(2)} 2^{s-1}\|\sigma\|_\infty^s\left(\frac{s}{p}\mathbb{E}|\varepsilon-\varepsilon'|^p + \frac{p-s}{p}\mathbb{E}|x-x'|^p\right) \\
\leq& \mathbb{E}|x-x'|^p\left(e^{\Delta(c_s^{(1)} + s[b]_{\text{Lip}})} + \Delta\tilde{\kappa}_1\right) + \Delta\tilde{\kappa}_2\mathbb{E}|\varepsilon-\varepsilon'|^p \\
\leq& \mathbb{E}|x-x'|^p e^{\Delta(c_s^{(1)} + s[b]_{\text{Lip}})}(1 + \Delta\tilde{\kappa}_1 e^{-\Delta(c_s^{(1)} + s[b]_{\text{Lip}})}) + \Delta\tilde{\kappa}_2\mathbb{E}|\varepsilon-\varepsilon'|^p \\
\leq& \mathbb{E}|x-x'|^p e^{\Delta(c_s^{(1)} + s[b]_{\text{Lip}} + \tilde{\kappa}_1)} + \Delta\tilde{\kappa}_2\mathbb{E}|\varepsilon-\varepsilon'|^p
\end{aligned}$$

where $\tilde{\kappa}_1 = c^{(2)}_{s,\Delta_{\max}} 2^{s-1}\left([\sigma]^s_{\mathrm{Lip}}\mathbb{E}|\varepsilon|^s + \|\sigma\|^s_\infty \frac{p-s}{p}\right)$ and $\tilde{\kappa}_2 = c^{(2)}_{s,\Delta_{\max}} 2^{s-1}\|\sigma\|^s_\infty \frac{s}{p}$. Then,

$$\|\mathcal{E}_k(x,\varepsilon) - \mathcal{E}_k(x',\varepsilon')\|_p \le \|x-x'\|_p e^{\Delta\kappa_1} + \|\varepsilon-\varepsilon'\|_p \Delta^{\frac{1}{p}}\kappa_2$$

where $\kappa_1 = (c^{(1)}_s + s[b]_{\mathrm{Lip}} + \tilde{\kappa}_1)/p$ and $\kappa_2 = \tilde{\kappa}_2^{\frac{1}{p}}$. Consequently, $\mathcal{E}_k$ is lipschitz continuous for $k \in \{1,\dots,n\}$ with Lipschitz coefficients $[F^x]_{\mathrm{Lip}} \le e^{\Delta\kappa_1}$ and $[F^\varepsilon]_{\mathrm{Lip}} \le \Delta^{\frac{1}{p}}\kappa_2$, for $p \in (1,2)$.

For the section step, the Lipschitz continuity of $\mathcal{E}_k$ yields

$$\begin{aligned}\|\bar{X}_{k+1} - \widetilde{X}_{k+1}\|_p \le& \|\mathcal{E}_k(\bar{X}_k,\varepsilon_k) - \mathcal{E}_k(\widehat{X}_k,\widehat{\varepsilon}_k)\|_p\\ \le& [F^x]_{\mathrm{Lip}}\|\bar{X}_k - \widehat{X}_k\|_p + [F^\varepsilon]_{\mathrm{Lip}}\|\varepsilon_k - \widehat{\varepsilon}_k\|_p\\ \le& [F^x]_{\mathrm{Lip}}\|\bar{X}_k - \widetilde{X}_k\|_p + [F^x]_{\mathrm{Lip}}\|\widetilde{X}_k - \widehat{X}_k\|_p + [F^\varepsilon]_{\mathrm{Lip}}\|\varepsilon_k - \widehat{\varepsilon}_k\|_p.\end{aligned}$$

Then, by induction, one has

$$\|\bar{X}_k - \widetilde{X}_k\|_p \le \sum_{l=1}^{k-1}[F^x]^{k-l}_{\mathrm{Lip}}\|\widehat{X}_l - \widetilde{X}_l\|_p + [F^\varepsilon]^{k-l}_{\mathrm{Lip}}\|\varepsilon_l - \widehat{\varepsilon}_l\|_p$$

so that

$$\|\bar{X}_k - \widehat{X}_k\|_p \le \|\bar{X}_k - \widetilde{X}_k\|_p + \|\widetilde{X}_k - \widehat{X}_k\|_p \le \sum_{l=1}^{k}[F^x]^{k-l}_{\mathrm{Lip}}\|\widetilde{X}_l - \widehat{X}_l\|_p + \sum_{l=1}^{k-1}[F^\varepsilon]^{k-l}_{\mathrm{Lip}}\|\varepsilon_l - \widehat{\varepsilon}_l\|_p.$$

Now, since $\widehat{\varepsilon}_l$ is an optimal quantization of ε_l of size N^ε_l, then Pierce's Lemma 6.1.1(b) yields

$$\|\bar{X}_k - \widehat{X}_k\|_p \le \sum_{l=1}^{k}[F^x]^{k-l}_{\mathrm{Lip}}\|\widetilde{X}_l - \widehat{X}_l\|_p + \sum_{l=1}^{k-1}[F^\varepsilon]^{k-l}_{\mathrm{Lip}}\kappa_{d,p,\eta}\|\varepsilon_l\|_{p+\eta}(N^\varepsilon_l)^{-\frac{1}{d}}. \tag{6.29}$$

As for the error terms $\|\widetilde{X}_l - \widehat{X}_l\|_p$, one uses the same techniques as in the end of the proof of Theorem 6.2.1, namely the distortion mismatch Theorem 6.2.2 and Lemma 6.2.4, to deduce the result. □

6.3 Time discretization of the RBSDE

We consider the reflected backward stochastic differential equation RBSDE (6.1) with maturity T given in the introduction and recalled below

$$Y_t = g(X_T) + \int_t^T f(s,X_s,Y_s,Z_s)ds + K_T - K_t - \int_t^T Z_s.dW_s\,, \qquad t\in[0,T],$$

$$Y_t \ge h(t,X_t) \quad \text{and} \quad \int_0^T (Y_s - h(s,X_s))dK_s = 0$$

where $(W_t)_{t\ge0}$ is a q-dimensional Brownian motion independent of X_0 and $(X_t)_{t\ge0}$ is an $\mathbb{R}^d$-valued Brownian diffusion process solution to the SDE (6.3) given in the introduction and recalled below

$$X_t = X_0 + \int_0^t b(s,X_s)ds + \int_0^t \sigma(s,X_s)dW_s\,, \qquad X_0 = x_0 \in \mathbb{R}^d,$$

As explained, we need to approximate the solutions of these equations by discretization schemes. The time and space discretization of the forward process $(X_t)_{t\in[0,T]}$ have already been investigated and detailed in Section 6.2. We proceed now with the time discretization of the solution of the RBSDE. Plugging the time-descretized process $(\bar{X}_{t_k})_{0\le k\le n}$ in (6.1) will not make it possible to find an exact solution for

the RBSDE. Another approximation is needed, in which we discretize the term Z_t itself: considering a sequence $(\varepsilon_k)_{0\leq k\leq n}$ of i.i.d. random variables normally distributed, the time discretization scheme associated to (Y_t, Z_t) is given by the following backward recursion

$$\bar{Y}_T = g(\bar{X}_T) \tag{6.30}$$

$$\widetilde{Y}_{t_k} = \mathbb{E}\big(\bar{Y}_{t_{k+1}} \,|\, \mathcal{F}_{t_k}\big) + \Delta \mathcal{E}_k\big(\bar{X}_{t_k}, \mathbb{E}\big(\bar{Y}_{t_{k+1}} \,|\, \mathcal{F}_{t_k}\big), \bar{\zeta}_{t_k}\big), \quad k = 0, \dots, n-1 \tag{6.31}$$

$$\bar{\zeta}_{t_k} = \frac{1}{\sqrt{\Delta}} \mathbb{E}\big(\bar{Y}_{t_{k+1}} \varepsilon_{k+1} \,|\, \mathcal{F}_{t_k}\big), \quad k = 0, \dots, n-1, \tag{6.32}$$

$$\bar{Y}_{t_k} = \widetilde{Y}_{t_k} \vee h_k(\bar{X}_{t_k}), \quad k = 0, \dots, n-1. \tag{6.33}$$

As stated previously, this scheme differs from what was previously studied in the literature (see the references in the Introduction) since the conditional expectation is applied directly to $\bar{Y}_{t_{k+1}}$ inside the driver function which depends itself on the discretization $\bar{\zeta}_{t_k}$ of Z_{t_k}. That is why it is interesting to establish a priori estimates for the error induced by the approximation with such a time discretization scheme. We note that, among others, time discretization errors for RBSDEs with a driver independent of Z_t were establsihed in [3], errors for BSDEs (without reflection) with a driver depending on Z_t and on the conditional expectation of $\bar{Y}_t$ in [65] and those for BSDEs (without reflection) with a driver depending on Z_t but where the conditional expectation is applied to the whole function f were studied in [76].

Since $\bar{X}_{t_k}$ is a Markov chain, one shows that there exists, for every $k \in \{0, \dots, n\}$, Borel functions $\bar{y}_{t_k}$, $\widetilde{y}_{t_k}$ and $\bar{z}_{t_k}$ such that $\bar{Y}_{t_k} = \bar{y}_{t_k}(\bar{X}_{t_k})$, $\widetilde{Y}_{t_k} = \widetilde{y}_{t_k}(\bar{X}_{t_k})$ and $\bar{\zeta}_{t_k} = \bar{z}_{t_k}(\bar{X}_{t_k})$ and defined by

$$\bar{y}_T(x) = g(x), \tag{6.34}$$

$$\widetilde{y}_{t_k}(x) = \mathbb{E}\, \bar{y}_{t_{k+1}}\big(\mathcal{E}_k(x, \varepsilon_{k+1})\big) + \Delta \mathcal{E}_k\left(x, \mathbb{E}\, \bar{y}_{t_{k+1}}\big(\mathcal{E}_k(x, \varepsilon_{k+1})\big), \bar{z}_{t_k}(x)\right) \tag{6.35}$$

$$\bar{z}_{t_k}(x) = \frac{1}{\sqrt{\Delta}} \mathbb{E}\Big(\bar{y}_{t_{k+1}}\big(\mathcal{E}_k(x, \varepsilon_{k+1})\big)\varepsilon_{k+1}\Big) \tag{6.36}$$

$$\bar{y}_{t_k}(x) = \widetilde{y}_{t_k}(x) \vee h_k(x). \tag{6.37}$$

where $\mathcal{E}_k(x, \varepsilon_{k+1}) = x + \Delta b_k(x) + \sqrt{\Delta}\sigma_k(x)\varepsilon_{k+1}$ and $(\varepsilon_k)_{k\geq 0}$ are i.i.d random variables with distribution $\mathcal{N}(0, I_q)$.

In order to establish error bounds between (Y_t, Z_t) and $(\bar{Y}_{t_k}, \bar{Z}_{t_k})$, it is useful to introduce a time continuous process which extends $\bar{Y}_{t_k}$. In fact, one notes that since the variable $\sum_{k=1}^{n-1} \bar{Y}_{t_{k+1}} - \mathbb{E}(\bar{Y}_{t_{k+1}} | \mathcal{F}_{t_k})$ is square integrable and measurable with respect to the augmented Brownian filtration $\mathcal{F}_{t_k}$, then, by the martingale representation Theorem, it can be considered as the terminal value of a Brownian martingale $\int_0^T \bar{Z}_s dW_s$ where the process $\bar{Z}_t$ is such that $\mathbb{E}\sup_{[0,T]} |\bar{Z}_s|^2 \leq \gamma_1 < +\infty$ for a finite constant γ_1. So,

$$\bar{Y}_{t_{k+1}} - \mathbb{E}(\bar{Y}_{t_{k+1}} | \mathcal{F}_{t_k}) = \int_{t_k}^{t_{k+1}} \bar{Z}_s dW_s \qquad \text{for } \; k = 0, \dots, n-1. \tag{6.38}$$

We note that

$$\bar{\zeta}_{t_k} = \frac{1}{\sqrt{\Delta}} \mathbb{E}\left(\bar{Y}_{t_{k+1}} \varepsilon_{k+1} \,|\, \mathcal{F}_{t_k}\right) = \frac{1}{\Delta} \mathbb{E}\left(\int_{t_k}^{t_{k+1}} \bar{Z}_s ds \,|\, \mathcal{F}_{t_k}\right). \tag{6.39}$$

Likewise, we define

$$\zeta_{t_k} = \frac{1}{\Delta} \mathbb{E}\left(\int_{t_k}^{t_{k+1}} Z_s ds | \mathcal{F}_{t_k}\right) \tag{6.40}$$

where Z_s is the solution of the RBSDE (6.1) and one checks that $\bar{\zeta}_t$ is the best approximation of $\bar{Z}_t$ and ζ_t the best approximation of Z_t in $L^2(d\mathbb{P} \times dt)$ among $\mathcal{F}_t$-measurable processes that are piecewise constant on the time intervals $[t_k, t_{k+1}[$.

Consequently, one may define (by a continuous extension) the càdlàg process $\widetilde{Y}_t$ on $[t_k, t_{k+1})$ and the làdcàg process $\bar{Y}_t$ on $(t_k, t_{k+1}]$, by

$$\widetilde{Y}_t = \bar{Y}_t = \bar{Y}_{t_{k+1}} - (t_{k+1} - t)\mathcal{E}_k\big(\bar{X}_{t_k}, \mathbb{E}(\bar{Y}_{t_{k+1}} \,|\, \mathcal{F}_{t_k}), \bar{\zeta}_{t_k}\big) - \int_t^{t_{k+1}} \bar{Z}_s dW_s, \tag{6.41}$$

and the increasing positive process

$$\bar{K}_{t_k} = \sum_{j=0}^{k} \Big(h_j(\bar{X}_{t_j}) - \widetilde{Y}_{t_k}\Big)_+$$

such that $\bar{K}_t = \bar{K}_{t_k}$ for every $t \in (t_k, t_{k+1})$. Finally, we have the following representation

$$\widetilde{Y}_t = \bar{Y}_T + \int_t^T f(\underline{s}, \bar{X}_{\underline{s}}, \mathbb{E}(\bar{Y}_{\bar{s}} \,|\, \mathcal{F}_{\underline{s}}), \bar{\zeta}_{\underline{s}})\, ds - \int_t^{t_{k+1}} \bar{Z}_s dW_s + \bar{K}_T - \bar{K}_t. \tag{6.42}$$

where $\underline{s} = t_k$ and $\bar{s} = t_{k+1}$ if $s \in (t_k, t_{k+1})$. Note that the introduction of $\bar{K}$ is mainly due to the fact that

$$\bar{Y}_{t_k} = \widetilde{Y}_{t_k} \vee h(t_k, \bar{X}_{t_k}) = \widetilde{Y}_{t_k} + \Big(h(t_j, \bar{X}_{t_j}) - \widetilde{Y}_{t_k}\Big)_+ = \widetilde{Y}_{t_k} + \bar{K}_{t_k} - \bar{K}_{t_{k-1}}.$$

In the following, we will denote $\bar{Y}_k, \bar{\zeta}_k, \bar{y}_k, \bar{K}_k$, etc. instead of $\bar{Y}_{t_k}, \bar{\zeta}_{t_k}, \bar{y}_{t_k}, \bar{K}_{t_k}$, etc. to alleviate notations, as well as $\mathbb{E}_k(.)$ instead of $\mathbb{E}(.|\mathcal{F}_{t_k})$. We recall that $\Delta \in [0, \Delta_{\max})$, $\Delta_{\max} > 0$.

Theorem 6.3.1. *Let Y_t be the solution of (6.1) and $(\bar{Y}_k)_{0\le k\le n}$ the corresponding time discretized process defined by (6.33). Assume that the functions f and h are lipschitz continuous. Then, for every $k \in \{1, \dots, n\}$,*

$$\mathbb{E}|Y_k - \bar{Y}_k|^2 \le C_{b,\sigma,f,h,T}\left(\Delta + \int_0^T \mathbb{E}|Z_s - Z_{\underline{s}}|^2 ds\right)$$

where $\underline{s} = t_k$ if $s \in [t_k, t_{k+1})$ and $C_{b,\sigma,f,h,T}$ is a real positive constant.
Furthermore, there exists a finite constant $C > 0$ such that

$$\int_0^T \mathbb{E}|Z_s - Z_{\underline{s}}|^2 ds \le C\sqrt{\Delta}.$$

The second part of the theorem is established in [48], see Theorem 6.3. The proof of the first part is postponed to the appendix (see Appendix B).

6.4 Space discretization of the RBSDE

After the time discretization, we move to the space discretization schemes to approximate the solution of the RBSDE. We rely on the recursive quantization $(\widehat{X}_{t_k})_{0\le k\le n}$ of the time discretized scheme $(\bar{X}_{t_k})_{0\le k\le n}$ to obtain the recursive quantization scheme associated to (6.30)-(6.31)-(6.32)-(6.33). If we consider a sequence $(\varepsilon_k)_{0\le k\le n}$ of i.i.d. random variables with distribution $\mathcal{N}(0, I_q)$, this scheme is defined recursively by

$$\widehat{Y}_T = g(\widehat{X}_T) \tag{6.43}$$

$$\widehat{\zeta}_{t_k} = \frac{1}{\sqrt{\Delta}}\mathbb{E}_k\left(\widehat{Y}_{t_{k+1}}\varepsilon_{k+1}\right), \quad k = 0, \dots, n-1, \tag{6.44}$$

$$\widehat{Y}_{t_k} = \max\left(h_k(\widehat{X}_{t_k}), \mathbb{E}_k\widehat{Y}_{t_{k+1}} + \Delta\mathcal{E}_k\big(\widehat{X}_{t_k}, \mathbb{E}_k\widehat{Y}_{t_{k+1}}, \widehat{\zeta}_{t_k}\big)\right), \quad k = 0, \dots, n-1. \tag{6.45}$$

where $(\widehat{X}_{t_k})_{0\le k\le n}$ is the recursively quantized process associated to $(\bar{X}_{t_k})_{0\le k\le n}$ given by (6.22) or (6.28). This quantization scheme is different than the optimal (or marginal) quantization schemes that were usually applied before in theses situations, in [3, 37, 65] for example. The main difference is that since recursive quantization preserve the Markov property, the process $\widehat{Y}_{t_k}$ is $\mathcal{F}_{t_k}$-measurable for every $k\in\{0,\dots,n\}$ where $\mathcal{F}_{t_k}=\sigma(W_{t_1},\dots,W_{t_k},\mathcal{N}_{\mathbb{P}})$ which is not the case for optimal quantization. More details on the utility of this character of recursive quantization will be presented in Section 6.5.

In the following, we will reconsider the notations with the indices k instead of t_k for every $k\in\{0,\dots,n\}$, and we establish an upper bound for the quantization error induced by approximating $\bar{Y}_k$ by $\widehat{Y}_k$ in L^p for $p\in(1,2+d)$ and $k\in\{1,\dots,n\}$. We recall that $\Delta\in[0,\Delta_{\max})$, $\Delta_{\max}>0$.

Theorem 6.4.1. *Let $(\bar{Y}_k)_{0\le k\le n}$ be the time discretized process defined by (6.33) and $(\widehat{Y}_k)_{0\le k\le n}$ the corresponding recursive quantized process defined by (6.45). For every $p\in(1,2+d)$ and every $k\in\{1,\dots,n\}$,*

$$\|\bar{Y}_k-\widehat{Y}_k\|_p\le\left(\frac{\kappa_2}{\kappa_1}(e^{(T-t_k)\kappa_1}-1)+e^{(T-t_k)\kappa_1}([g]^p_{\mathrm{Lip}}\vee[h]^p_{\mathrm{Lip}})\right)\Big\|\max_{k\le l\le n}\big|\bar{X}_l-\widehat{X}_l\big|\Big\|_p \tag{6.46}$$

where $\kappa_1=p\kappa+(p-1)2^{p-2}$, $\kappa_2=2^{p-2}[f]^p_{\mathrm{Lip}}(1+p\Delta^{p-1})$ and $\kappa=\frac{c_s^{(1)}+s[f]_{\mathrm{Lip}}+[f]^s_{\mathrm{Lip}}c^{(3)}_{s,\Delta_{\max},\varepsilon_{k+1}}}{s}$, the positive finite constants $c_s^{(1)}$ and $c^{(3)}_{s,\Delta_{\max},\varepsilon_{k+1}}$ are defined in Lemmas 6.2.3 and 6.2.4.

Remark 6.4.2. *The norms $\|\bar{X}_l-\widehat{X}_l\|_p$ are recursive quantization errors established in Theorems 6.2.1 and 6.2.7 for $p\in(1,2+d)$. We recall that, for every $l\in\{1,\dots,n\}$, one has $\|\bar{X}_l-\widehat{X}_l\|_p=\mathcal{O}(N_l^{-\frac{1}{d}})$ where N_l is the size of the quantization grid corresponding to $\widehat{X}_l$.*

Proof. For every $k\in\{1,\dots,n\}$, we use the inequality $|\max(a,b)-\max(a',b')|\le\max(|a-a'|,|b-b'|)$ and have

$$|\bar{Y}_k-\widehat{Y}_k|\le\max\Big(|h_k(\bar{X}_k)-h_k(\widehat{X}_k)|,\Big|\mathbb{E}_k\bar{Y}_{k+1}-\mathbb{E}_k\widehat{Y}_{k+1}+\Delta\big(\mathcal{E}_k(\bar{X}_k,\mathbb{E}_k\bar{Y}_{k+1},\bar{\xi}_k)-\mathcal{E}_k(\widehat{X}_k,\mathbb{E}_k\widehat{Y}_{k+1},\widehat{\xi}_k)\big)\Big|\Big)$$

We denote $\beta_k=\mathbb{E}_k(\bar{Y}_{k+1}-\widehat{Y}_{k+1})+\Delta\left(\mathcal{E}_k\big(\bar{X}_k,\mathbb{E}_k\bar{Y}_{k+1},\bar{\xi}_k\big)-\mathcal{E}_k\big(\widehat{X}_k,\mathbb{E}_k\widehat{Y}_{k+1},\widehat{\xi}_k\big)\right)$ and we have

$$\beta_k=\mathbb{E}_k(\bar{Y}_{k+1}-\widehat{Y}_{k+1})+\Delta\Big(\widehat{A}_k(\bar{X}_k-\widehat{X}_k)+\widehat{B}_k\mathbb{E}_k(\bar{Y}_{k+1}-\widehat{Y}_{k+1})+\frac{\widehat{C}_k}{\sqrt{\Delta}}\mathbb{E}_k\big((\bar{Y}_{k+1}-\widehat{Y}_{k+1})\varepsilon_{k+1}\big)\Big)$$

where

$$\widehat{A}_k=\frac{\mathcal{E}_k\big(\bar{X}_k,\mathbb{E}_k\bar{Y}_{k+1},\bar{\xi}_k\big)-\mathcal{E}_k\big(\widehat{X}_k,\mathbb{E}_k\bar{Y}_{k+1},\bar{\xi}_k\big)}{\bar{X}_k-\widehat{X}_k}\mathbb{1}_{\bar{X}_k\neq\widehat{X}_k},$$

$$\widehat{B}_k=\frac{\mathcal{E}_k\big(\widehat{X}_k,\mathbb{E}_k\bar{Y}_{k+1},\bar{\xi}_k\big)-\mathcal{E}_k\big(\widehat{X}_k,\mathbb{E}_k\widehat{Y}_{k+1},\bar{\xi}_k\big)}{\mathbb{E}_k(\bar{Y}_{k+1}-\widehat{Y}_{k+1})}\mathbb{1}_{\mathbb{E}_k\bar{Y}_{k+1}\neq\mathbb{E}\widehat{Y}_{k+1}},$$

$$\widehat{C}_k=\frac{\mathcal{E}_k\big(\widehat{X}_k,\mathbb{E}_k\widehat{Y}_{k+1},\bar{\xi}_k\big)-\mathcal{E}_k\big(\widehat{X}_k,\mathbb{E}_k\widehat{Y}_{k+1},\widehat{\xi}_k\big)}{\mathbb{E}_k\big((\bar{Y}_{k+1}-\widehat{Y}_{k+1})\varepsilon_{k+1}\big)}\mathbb{1}_{\bar{\xi}_k\neq\widehat{\xi}_k}.$$

It is clear that $\max\Big(|\widehat{A}_k|,|\widehat{B}_k|,|\widehat{C}_k|\Big)\le[f]_{\mathrm{Lip}}$, so one has

$$|\beta_k|\le\Delta[f]_{\mathrm{Lip}}|\bar{X}_k-\widehat{X}_k|+\mathbb{E}_k\Big|(1+\Delta\widehat{B}_k+\sqrt{\Delta}\widehat{C}_k\varepsilon_{k+1})(\bar{Y}_{k+1}-\widehat{Y}_{k+1})\Big|.$$

At this stage, we consider two conjugate exponents $r\in(1,2\wedge p)$ and $s=\frac{r}{r-1}>2$ and we apply conditional Hölder's inequality

$$\mathbb{E}_k\Big|(1+\Delta\widehat{B}_k+\sqrt{\Delta}\widehat{C}_k\varepsilon_{k+1})(\bar{Y}_{k+1}-\widehat{Y}_{k+1})\Big|\le\Big(\mathbb{E}_k|1+\Delta\widehat{B}_k+\sqrt{\Delta}\widehat{C}_k\varepsilon_{k+1}|^s\Big)^{\frac{1}{s}}\Big(\mathbb{E}_k|\bar{Y}_{k+1}-\widehat{Y}_{k+1}|^r\Big)^{\frac{1}{r}}.$$

Since $s > 2$, one can apply Lemma 6.2.3 with $a = 1 + \Delta\widehat{B}_k$ and $A = \widehat{C}_k$ and obtains

$$\begin{aligned}\mathbb{E}_k|1 + \Delta\widehat{B}_k + \sqrt{\Delta}\widehat{C}_k\varepsilon_{k+1}|^s &\leq (1 + \Delta[f]_{\text{Lip}})^s(1 + c_s^{(1)}\Delta) + \Delta[f]^s_{\text{Lip}}c^{(3)}_{s,\Delta_{\max},\varepsilon_{k+1}} \\ &\leq e^{s\Delta[f]_{\text{Lip}}+\Delta c_s^{(1)}} + \Delta[f]^s_{\text{Lip}}c^{(3)}_{s,\Delta_{\max},\varepsilon_{k+1}} \\ &\leq e^{\Delta(c_s^{(1)}+s[f]_{\text{Lip}})}(1 + \Delta[f]^s_{\text{Lip}}c^{(3)}_{s,\Delta_{\max},\varepsilon_{k+1}}e^{-\Delta(c_s^{(1)}+s[f]_{\text{Lip}})}) \\ &\leq e^{\Delta(c_s^{(1)}+s[f]_{\text{Lip}}+[f]^s_{\text{Lip}}c^{(3)}_{s,\Delta_{\max},\varepsilon_{k+1}})}\end{aligned}$$

where $c_s^{(1)}$ and $c^{(3)}_{s,\Delta_{\max},\varepsilon_{k+1}}$ are real constants defined in Lemmas 6.2.3 and 6.2.4. Therefore,

$$|\beta_k| \leq \Delta[f]_{\text{Lip}}|\bar{X}_k - \widehat{X}_k| + e^{\kappa\Delta}\left(\mathbb{E}_k|\bar{Y}_{k+1} - \widehat{Y}_{k+1}|^r\right)^{\frac{1}{r}},$$

where $\kappa = \frac{c_s^{(1)}+s[f]_{\text{Lip}}+[f]^s_{\text{Lip}}c^{(3)}_{s,\Delta_{\max},\varepsilon_{k+1}}}{s}$, and

$$|\bar{Y}_k - \widehat{Y}_k|^p \leq \max\left([h]^p_{\text{Lip}}|\bar{X}_k - \widehat{X}_k|^p\,,\,|\beta_k|^p\right).$$

Now, using inequality (6.27) yields

$$|\beta_k|^p \leq e^{p\kappa\Delta}\left(\mathbb{E}_k|\bar{Y}_{k+1} - \widehat{Y}_{k+1}|^r\right)^{\frac{p}{r}}\left(1 + (p-1)2^{p-2}\varepsilon^p\right) + 2^{p-2}[f]^p_{\text{Lip}}|\bar{X}_k - \widehat{X}_k|^p\Delta^p\left(p + \frac{1}{\varepsilon^{p(p-1)}}\right).$$

We choose $\varepsilon = \Delta^{\frac{1}{p}}$ so that $\Delta^p\left(p + \frac{1}{\varepsilon^{p(p-1)}}\right) = \Delta(1 + p\Delta^{p-1})$ and hence

$$|\beta_k|^p \leq e^{\kappa_1\Delta}\left(\mathbb{E}_k|\bar{Y}_{k+1} - \widehat{Y}_{k+1}|^r\right)^{\frac{p}{r}} + \Delta\kappa_2|\bar{X}_k - \widehat{X}_k|^p$$

where $\kappa_1 = p\kappa + (p-1)2^{p-2}$ and $\kappa_2 = 2^{p-2}[f]^p_{\text{Lip}}(1 + p\Delta^{p-1})$. Moreover, by our choice of r, we have that $\frac{p}{r} > 1$ so we apply Jensen's inequality and obtain

$$|\beta_k|^p \leq e^{\kappa_1\Delta}\mathbb{E}_k|\bar{Y}_{k+1} - \widehat{Y}_{k+1}|^p + \Delta\kappa_2|\bar{X}_k - \widehat{X}_k|^p.$$

Hence, having in mind that $\bar{X}_k, \widehat{X}_k, \bar{Y}_k$and $\widehat{Y}_k$ are all $\mathcal{F}_{t_k}$-measurable processes, one has

$$\mathbb{E}_k|\bar{Y}_k - \widehat{Y}_k|^p \leq \max\left([h]^p_{\text{Lip}}\mathbb{E}_k|\bar{X}_k - \widehat{X}_k|^p\,,\,e^{\kappa_1\Delta}\mathbb{E}_k|\bar{Y}_{k+1} - \widehat{Y}_{k+1}|^p + \Delta\kappa_2\mathbb{E}_k|\bar{X}_k - \widehat{X}_k|^p\right). \tag{6.47}$$

At this stage, we aim to prove that $\mathbb{E}_k|\bar{Y}_k - \widehat{Y}_k|^p$ satisfies the following backward induction

$$\mathbb{E}_k|\bar{Y}_k - \widehat{Y}_k|^p \leq e^{(n-k)\kappa_1\Delta}\left([g]^p_{\text{Lip}} \vee [h]^p_{\text{Lip}}\right)\mathbb{E}_k\max_{k\leq i\leq n}|\bar{X}_i - \widehat{X}_i|^p + \Delta\kappa_2\sum_{i=k}^{n-1}e^{(i-k)\kappa_1\Delta}\mathbb{E}_k|\bar{X}_i - \widehat{X}_i|^p. \tag{6.48}$$

First, it is clear that $\mathbb{E}_n|\bar{Y}_n - \widehat{Y}_n|^p \leq [g]^p_{\text{Lip}}\mathbb{E}_n|\bar{X}_n - \widehat{X}_n|^p$ so the induction is satisfied for $k = n$. We assume that (6.48) is true for $k+1$ i.e.

$$\begin{aligned}\mathbb{E}_{k+1}|\bar{Y}_{k+1} - \widehat{Y}_{k+1}|^p &\leq e^{(n-k-1)\kappa_1\Delta}([g]^p_{\text{Lip}} \vee [h]^p_{\text{Lip}})\mathbb{E}_{k+1}\max_{k+1\leq i\leq n}|\bar{X}_i - \widehat{X}_i|^p \\ &\quad + \Delta\kappa_2\sum_{i=k+1}^{n-1}e^{(i-k-1)\kappa_1\Delta}\mathbb{E}_{k+1}|\bar{X}_i - \widehat{X}_i|^p\end{aligned} \tag{6.49}$$

and show it for k. In fact, since $\mathbb{E}_k\mathbb{E}_{k+1}(\cdot) = \mathbb{E}_k(\cdot)$, one has, by merging (6.47) with (6.49), that

$$\begin{aligned}
\mathbb{E}_k|\bar{Y}_k - \widehat{Y}_k|^p &\le \max\Big([h]^p_{\mathrm{Lip}}\mathbb{E}_k|\bar{X}_k - \widehat{X}_k|^p\,,\, e^{\kappa_1\Delta}\mathbb{E}_k\mathbb{E}_{k+1}|\bar{Y}_{k+1} - \widehat{Y}_{k+1}|^p + \Delta\kappa_2\mathbb{E}_k|\bar{X}_k - \widehat{X}_k|^p\Big)\\
&\le \max\Big([h]^p_{\mathrm{Lip}}\mathbb{E}_k|\bar{X}_k - \widehat{X}_k|^p\,,\, \Delta\kappa_2\mathbb{E}_k|\bar{X}_k - \widehat{X}_k|^p + \Delta\kappa_2\sum_{i=k+1}^{n-1} e^{(i-k)\kappa_1\Delta}\mathbb{E}_k\mathbb{E}_{k+1}|\bar{X}_i - \widehat{X}_i|^p\\
&\quad + e^{(n-k)\kappa_1\Delta}([g]^p_{\mathrm{Lip}} \vee [h]^p_{\mathrm{Lip}})\mathbb{E}_k\mathbb{E}_{k+1}\max_{k+1\le i\le n}|\bar{X}_i - \widehat{X}_i|^p\Big)\\
&\le \max\Big([h]^p_{\mathrm{Lip}}\mathbb{E}_k|\bar{X}_k - \widehat{X}_k|^p\,, e^{(n-k)\kappa_1\Delta}([g]^p_{\mathrm{Lip}} \vee [h]^p_{\mathrm{Lip}})\mathbb{E}_k\max_{k\le i\le n}|\bar{X}_i - \widehat{X}_i|^p\\
&\quad + \Delta\kappa_2\sum_{i=k}^{n-1} e^{(i-k)\kappa_1\Delta}\mathbb{E}_k|\bar{X}_i - \widehat{X}_i|^p\Big)
\end{aligned}$$

since $\max_{k+1\le i\le n}\alpha_i \le \max_{k\le i\le n}\alpha_i$ for $\alpha_i > 0$. Furthermore, noticing that

$$[h]^p_{\mathrm{Lip}}\mathbb{E}_k|\bar{X}_k - \widehat{X}_k|^p \le \big([g]^p_{\mathrm{Lip}} \vee [h]^p_{\mathrm{Lip}}\big)\mathbb{E}_k\max_{k\le i\le n}|\bar{X}_i - \widehat{X}_i|^p \le e^{(n-k)\kappa_1\Delta}\big([g]^p_{\mathrm{Lip}} \vee [h]^p_{\mathrm{Lip}}\big)\mathbb{E}_k\max_{k\le i\le n}|\bar{X}_i - \widehat{X}_i|^p$$

because $e^{(n-k)\kappa_1\Delta} > 1$, one concludes the induction (6.48). This yields

$$\mathbb{E}_k|\bar{Y}_k - \widehat{Y}_k|^p \le e^{(T-t_k)\kappa_1}\big([g]^p_{\mathrm{Lip}} \vee [h]^p_{\mathrm{Lip}}\big)\mathbb{E}_k\max_{k\le i\le n}|\bar{X}_i - \widehat{X}_i|^p + \Delta\kappa_2\mathbb{E}_k\max_{k\le i\le n}|\bar{X}_i - \widehat{X}_i|^p\sum_{i=k}^{n-1} e^{(i-k)\kappa_1\Delta}. \quad (6.50)$$

Finally, since $e^x - 1 \ge x$ for $x \ge 0$, one has

$$\sum_{i=k}^{n-1} e^{(i-k)\kappa_1\Delta} = \frac{e^{(n-k)\kappa_1\Delta} - 1}{e^{\kappa_1\Delta} - 1} \le \frac{e^{(T-t_k)\kappa_1} - 1}{\Delta\kappa_1}$$

and then deduces the result by taking the expectation in (6.50) . □

6.5 Algorithmics

Our aim is to write $(\widehat{Y}_k, \widehat{\zeta}_k)$, which approximates the solution of the RBSDE (6.1), in a form that allows us to compute their values. For this, we first note that $(\bar{X}_k)_{0\le k\le n}$ and $(\widehat{X}_k)_{0\le k\le n}$ are both $\mathcal{F}_{t_k}$-Markov chains where $\mathcal{F}_{t_k} = \sigma(W_s, s \le t_k, \mathcal{N}_{\mathbb{P}})$, for every $k \in \{0, \dots, n\}$, with respective transitions $P_k(x, dy) = \mathbb{P}(\bar{X}_{k+1} \in dy | \bar{X}_k = x)$ and $\widehat{P}_k(x, dy) = \mathbb{P}(\widehat{X}_{k+1} \in dy | \widehat{X}_k = x)$. The main advantage of recursive quantization is that it preserves the Markovian property of $(\widehat{X}_k)_{0\le k\le n}$ with respect to the filtration $(\mathcal{F}_{t_k})_{0\le k\le n} = \big(\sigma(W_s, s \le t_k, \mathcal{N}_{\mathbb{P}})\big)_{0\le k\le n}$. Note that, for optimal quantization, the trick was to force the Markov property by conditioning with respect to the filtration $\widehat{\mathcal{F}}_{t_k} = \sigma(\widehat{X}_0, \dots, \widehat{X}_k)$ instead of $\mathcal{F}_{t_k}$ in (6.44)-(6.45). The price to pay is that the approximations $\|\bar{X}_k - \widehat{X}_k\|_p$, for every $k \in \{1, \dots, n\}$, are less accurate (but not in a drastic way). This point is discussed in details in [65].

For every bounded or non-negative Borel function f, one has $P_k f(x) = \displaystyle\int_{\mathbb{R}^d} f(y)P_k(x, dy)$, so that

$$\mathbb{E}\big(f(\bar{X}_{k+1})\,|\,\mathcal{F}_{t_k}\big) = P_k f(\bar{X}_k) \qquad \text{and} \qquad \mathbb{E}\big(f(\widehat{X}_{k+1})\,|\,\mathcal{F}_{t_k}\big) = \widehat{P}_k f(\widehat{X}_k).$$

Moreover, we introduce

$$Q_k f(\bar{X}_k) = \frac{1}{\sqrt{\Delta}}\mathbb{E}\big(f(\bar{X}_{k+1})\varepsilon_{k+1}\,|\,\mathcal{F}_{t_k}\big) \qquad \text{and} \qquad \widehat{Q}_k f(\widehat{X}_k) = \frac{1}{\sqrt{\Delta}}\mathbb{E}\big(f(\widehat{X}_{k+1})\varepsilon_{k+1}\,|\,\mathcal{F}_{t_k}\big)$$

where $(\varepsilon_k)_{0\le k\le n}$ are i.i.d. with Normal distribution $\mathcal{N}(0, I_q)$.

Similarly to the functions $(\bar{y}_k)_{0\le k\le n}$ defined by (6.37), one shows that there exists Borel functions $(\widehat{y}_k)_{0\le k\le n}$ such that $\widehat{Y}_k = \widehat{y}_k(\widehat{X}_k)$ for every $k \in \{0,\dots,n\}$. They are defined recursively by the following Backward Dynamic Programming Principle (BDPP)

$$\left\{\begin{array}{rl} \widehat{y}_n & = h_n \\ \widehat{y}_k & = \max\Big(h_k,\, \widehat{P}_k\widehat{y}_{k+1} + \Delta\mathcal{E}_k\big(.,\widehat{P}_k\widehat{y}_{k+1}, \widehat{Q}_k\widehat{y}_{k+1}\big)\Big), \quad k=0,\dots,n-1, \end{array}\right. \tag{6.51}$$

This BDPP can also be written in distribution, one can write $(\bar{y}_k)_{0\le k\le n}$ as

$$\left\{\begin{array}{rl} \bar{y}_n & = h_n \\ \bar{y}_k & = \max\Big(h_k,\, P_k\bar{y}_{k+1} + \Delta\mathcal{E}_k\big(.,P_k\bar{y}_{k+1}, Q_k\bar{y}_{k+1}\big)\Big), \quad k=0,\dots,n-1, \end{array}\right.$$

The fact that $\bar{Y}_k = \bar{y}_k(\bar{X}_k)$ and $\widehat{Y}_k = \widehat{y}_k(\widehat{X}_k)$ can easily be checked by a backward induction relying on (6.30)-(6.31)-(6.33) and (6.43)-(6.45) respectively. Furthermore, there exists functions $\bar{z}_k$ and $\widehat{z}_k$ such that $\bar{\zeta}_k = \bar{z}_k(\bar{X}_k)$ and $\widehat{\zeta}_k = \widehat{z}_k(\widehat{X}_k)$, defined by

$$\bar{z}_k = Q_k\bar{y}_{k+1} \qquad \text{and} \qquad \widehat{z}_k = \widehat{Q}_k\widehat{y}_{k+1}.$$

In order to compute $\widehat{Y}_k$ and $\widehat{\zeta}_k$, we first need to compute the optimal (or at least optimized) recursive quantization $\widehat{X}_k$ of $\bar{X}_k$ for every $k \in \{0,\dots,n\}$ and the corresponding transition weights. We will consider the quadratic case $p = 2$ for all numerical aspects.

6.5.1 Computation of the recursive quantizers

As defined previously, the recursive quantization of $(\bar{X}_k)_{0\le k\le n}$ is realized via (6.22) (or (6.28)). In a quadratic framework, the computation of the optimal quantization grids Γ_k of $\widetilde{X}_k$ of size N_k, at each time step t_k, is achieved by algorithms such as CLVQ (Competitive Learning Vector Quantization), Lloyd's algorithm or Newton-Raphson. These algorithms are presented in details in [61] for example. Here, we expose a variant of Lloyd's algorithm for recursive quantization.

For $k \in \{1,\dots,n\}$, computing an optimal quantizer $\widehat{X}_k^{\Gamma_k}$ of $\widetilde{X}_k$ consists in computing the grid Γ_k solution to the minimization problem

$$\Gamma_k \in \operatorname{argmin}\Big\{\|\widehat{X}_k^{\Gamma} - \widetilde{X}_k\|_2^2,\, \Gamma\subset\mathbb{R}^d,\, \operatorname{card}(\Gamma)\le N_k\Big\}.$$

The construction of these grids is performed recursively at each step t_k in a forward way. It is somehow an *embedded* optimization. We suppose that, at time t_k, the grid $\Gamma_k = \{x_1^k,\dots,x_{N_k}^k\}$ is already computed (optimized) and that $\widetilde{X}_k$ has been quantized by $\widehat{X}_k = \sum_{i=1}^{N_k} x_i^k \mathbb{1}_{C_i(\Gamma_k)}$ where $(C_i(\Gamma_k))_{1\le i\le N_k}$ is the Voronoï diagram associated to $\widehat{X}_k$ and defined by (6.11). Then, at time step t_{k+1}, we build the grid Γ_{k+1} that minimizes the quadratic distortion $G_{k+1}^2(\Gamma)$ defined by (6.23) and written as a function of the grid $\Gamma_k = \{x_1^k,\dots,x_{N_k}^k\}$ computed at the previous step. So, if $\Gamma_{k+1} = \{x_1^{k+1},\dots,x_{N_{k+1}}^{k+1}\}$, then one has, for every $j \in \{1,\dots,N_{k+1}\}$,

$$x_j^{k+1} = \mathbb{E}\Big(\widetilde{X}_{k+1} \,|\, \widehat{X}_{k+1}\in C_j(\Gamma_{k+1})\Big) = \frac{\sum_{i=1}^{N_k} p_i^k \mathbb{E}\Big(\mathcal{E}_k(x_i^k,\varepsilon_{k+1})\mathbb{1}_{\{\mathcal{E}_k(x_i^k,\varepsilon_{k+1})\in C_j(\Gamma_{k+1})\}}\Big)}{p_j^{k+1}}. \tag{6.52}$$

Recalling that $\mathcal{E}_k(x,\varepsilon_{k+1}) = x + \Delta b_k(x) + \sqrt{\Delta}\sigma_k(x)\varepsilon_{k+1}$, it is important to notice that, for every $k \in \{1,\dots,n\}$ and $i \in \{1,\dots,N_k\}$, $\mathcal{E}_k(x_i^k,\varepsilon_{k+1}) \sim \mathcal{N}(m_i^k,\Sigma_i^k)$ where $m_i^k = x_i^k + \Delta b_k(x_i^k)$ and $\Sigma_i^k = \sqrt{\Delta}\sigma_k(x_i^k)$.

We are interested in more than just computing the distribution of $(\widehat{X}_k)_{0\le k\le n}$, the computation of the transition matrices $P_k = (p_{ij}^k)_{ij}$ is even more fundamental among the companion parameters in view of our applications. For every $k \in \{1,\dots,n\}$ and $i,j \in \{1,\dots,N_k\}$, the transition probability p_{ij}^k from x_i^k to x_j^{k+1} is given by

$$p_{ij}^k = \mathbb{P}\left(\widetilde{X}_{k+1} \in C_j(\Gamma_{k+1}) \,|\, \widetilde{X}_k \in C_i(\Gamma_k)\right) = \mathbb{P}\left(\mathcal{E}_k(x_i^k, \varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\right). \tag{6.53}$$

This identity allows the computation of the weights p_j^{k+1} of the Voronoï cells $C_j(\Gamma_{k+1})$, for every $j \in \{1,\dots,N_{k+1}\}$, via the classical (discrete time) forward Kolmogorov equation. They are given by

$$p_j^{k+1} = \mathbb{P}\big(\widetilde{X}^{k+1} \in C_j(\Gamma_{k+1})\big) = \sum_{i=1}^{N_k} p_i^k \mathbb{P}\Big(\mathcal{E}_k(x_i^k, \varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\Big). \tag{6.54}$$

One-dimensional setting $q = d = 1$: The transition weights p_{ij}^k can be computed in a direct way as follows: for every $i \in \{1,\dots,N_k\}$ and $j \in \{1,\dots,N_{k+1}\}$

$$p_{ij}^k = \mathbb{P}\Big(\widetilde{X}_{k+1} \le x_{j+\frac{1}{2}}^{k+1} \,|\, \widehat{X}_k = x_i^k\Big) - \mathbb{P}\Big(\widetilde{X}_{k+1} \le x_{j-\frac{1}{2}}^{k+1} \,|\, \widehat{X}_k = x_i^k\Big) = \Phi_0\big(x_{i,j_+}^{k+1}\big) - \Phi_0\big(x_{i,j_-}^{k+1}\big)$$

where Φ_0 is the cumulative distribution function of the standard Normal distribution $\mathcal{N}(0,1)$ and

$$x_{i,j_+}^{k+1} = \frac{x_{j+\frac{1}{2}}^{k+1} - x_i^k - \Delta b_k(x_i^k)}{\sqrt{\Delta}\sigma_k(x_i^k)} \quad \text{and} \quad x_{i,j_-}^{k+1} = \frac{x_{j-\frac{1}{2}}^{k+1} - x_i^k - \Delta b_k(x_i^k)}{\sqrt{\Delta}\sigma_k(x_i^k)}$$

with $x_{j+\frac{1}{2}}^{k+1} = \frac{x_j^{k+1}+x_{j+1}^{k+1}}{2}$, $x_{\frac{1}{2}}^{k+1} = -\infty$ and $x_{N_{k+1}-\frac{1}{2}}^{k+1} = +\infty$.

General setting: In order to approximate the transition probabilities and the weights of the Voronoï cells when $d > 1$, one may proceed with Monte Carlo simulations or rely on Markovian and componentwise product quantization (see [28]). A very interesting alternative is the hybrid recursive quantization, studied in Section 6.2.3, where we replaced the white Gaussian noise by its optimal quantization sequences. The principle on which we rely to design the hybrid recursive quantizers is the same as the one for the standard recursive quantization. The only difference is with the computation of the expectations and probabilities in (6.52),(6.53) and (6.54). Instead of resorting to large and slow Monte Carlo simulations, we consider sequences of optimal quantizers $(\hat{\varepsilon}_l^k)_{1\le l\le N_\varepsilon}$ of size N_ε of the Gaussian distribution $\mathcal{N}(0, I_d)$, available on the quantization website www.quantize.maths-fi.com, and compute the sequence and its companion parameters based on the following formulas

$$\mathbb{E}\Big(\mathcal{E}_k(x_i^k, \varepsilon_k)\mathbb{1}_{\mathcal{E}_k(x_i^k,\varepsilon_k)\in C_j(\Gamma_{k+1})}\Big) = \sum_{l=1}^{N_\varepsilon} p_{\varepsilon_l}^k \mathcal{E}_k(x_i^k, \hat{\varepsilon}_l^k)\mathbb{1}_{\mathcal{E}_k(x_i^k,\hat{\varepsilon}_l^k)\in C_j(\Gamma_{k+1})} \tag{6.55}$$

and

$$P\big(\mathcal{E}_k(x_i^k, \varepsilon_k) \in C_j(\Gamma_{k+1})\big) = \sum_{l=1}^{N_\varepsilon} p_{\varepsilon_l}^k \mathbb{1}_{\mathcal{E}_k(x_i^k,\hat{\varepsilon}_l^k)\in C_j(\Gamma_{k+1})} \tag{6.56}$$

where $p_{\varepsilon_l}^k$ is the weight of the Voronoï cell of centroid $\hat{\varepsilon}_l^k$, also available on the quantization website.

6.5.2 Computation of the quantized solution of the RBSDE

Having already computed the recursive quantization $(\widehat{X}_k)_{0\le k\le n}$ of $(\bar{X}_k)_{0\le k\le n}$ as described in the previous section 6.5.1, as well as the corresponding companion parameters (the weights $(p_i^k)_{1\le i\le N_k}$ of Voronoï cells

and the transition weights $(p_{ij}^k)_{1\le i\le N_k, 1\le j\le N_{k+1}}$, we proceed with the computation of $(\widehat{Y}_k)_{0\le k\le n}$ and rely on the BDPP (6.51) allowing us to compute $\widehat{Y}_k = \widehat{y}_k(\widehat{X}_k)$ as a function of the quantizer $\Gamma_k = \{x_1^k, \dots, x_{N_k}^k\}$. For every $k \in \{0, \dots, n-1\}$ and $i \in \{1, \dots, N_k\}$, we denote

$$\widehat{\alpha}_k(x_i^k) = \sum_{j=1}^{N_{k+1}} \widehat{y}_{k+1}(x_j^{k+1}) p_{ij}^k \qquad \text{and} \qquad \widehat{\beta}_k(x_i^k) = \frac{1}{\Delta} \sum_{j=1}^{N_{k+1}} \widehat{y}_{k+1}(x_j^{k+1}) \pi_{ij}^k$$

where

$$\pi_{ij}^k = \frac{\sqrt{\Delta}}{p_i^k} \mathbb{E}\Big(\varepsilon_{k+1} \mathbb{1}_{\{\widehat{X}_{k+1} = x_j^{k+1}, \widehat{X}_k = x_i^k\}}\Big) = \sqrt{\Delta}\mathbb{E}\Big(\varepsilon_{k+1} \mathbb{1}_{\mathcal{E}_k(x_i^k, \varepsilon_{k+1}) \in C_j(\Gamma_{k+1})}\Big) \tag{6.57}$$

and $\mathcal{E}_k(x, \varepsilon_{k+1}) = x + \Delta b_k(x) + \sqrt{\Delta}\sigma_k(x)\varepsilon_{k+1}$. Note that the quantities $(\pi_{ij}^k)_{1\le i,j\le N_k}$ are computed online at the same time as the transition weight matrices $(p_{ij}^k)_{1\le i,j\le N_k}$ for every $k \in \{0, \dots, n-1\}$, so that they can be stored and used instantly in the computations of the solution of the RBSDE.

Therefore, the solution Y_0 of the RBSDE is approximated by the value $\widehat{y}_0$ at time t_0 of the following recursive quantized scheme

$$\begin{cases} \widehat{y}_n(x_i^n) & = h_n(x_i^n), \qquad i = 1, \dots, N_n, \\ \widehat{y}_k(x_i^k) & = \max\Big(h_k(x_i^k),\, \widehat{\alpha}_k(x_i^k) + \Delta \mathcal{E}_k\big(x_i^k, \widehat{\alpha}_k(x_i^k), \widehat{\beta}_k(x_i^k)\big)\Big), \qquad i = 1, \dots, N_k, \end{cases} \tag{6.58}$$

And, the function $\hat{z}_k$ used to approximate $\hat{\zeta}_k$ is computed via the following sum

$$\hat{z}_k(x_i^k) = \frac{1}{\Delta} \sum_{j=1}^{N_{k+1}} \hat{y}_{k+1}(x_j^{k+1}) \pi_{ij}^k.$$

Remark 6.5.1. *One should mention that, once the recursive quantization grids and the corresponding companion parameters are computed, the computation of the solution of the RBSDE is almost instantaneous, we can even say that its computational cost is negligible.*

6.6 Numerical examples

We carry out some numerical experiments to illustrate the rate of convergence of the recursive quantization-based discretized scheme and to compare its performances with other schemes based on optimal quantization, greedy quantization and greedy recursive quantization. We start by explaining how to obtain the quantizers and their companions parameters (Voronoï and transition weights) by optimal, greedy and recursive greedy quantization. Concerning the time discretization, we consider the Euler scheme of the forward diffusion $(X_t)_{0\le t\le T}$ defined by (6.20).

6.6.1 Various quantization methods

Quanization tree with optimal marginal quantization

In this section, we aim to build optimal quantizers $\widehat{X}_k^{\Gamma_k}$ of $\bar{X}_k$ for every $k \in \{0, \dots, n\}$. At time t_0, we start with $\widehat{X}_0 = X_0 = x_0 \in \mathbb{R}^d$. Then, at each time step t_k, we rely on a sequence of optimal quantizers $(z_i^k)_{1\le i\le N_k}$ of size N_k of the Normal distribution $\mathcal{N}(0, I_d)$ and we compute the quantizer $\Gamma_k = (x_1^k, \dots, x_{N_k}^k)$ via

$$x_i^k = x_0 + t_k b(x_0) + \sqrt{t_k}\sigma(x_0) z_i^k, \qquad i \in \{1, \dots, N_k\}.$$

In particular, if $(\bar{X}_k)_{0,\leq k\leq n}$ evolves following a Black-Scholes model with interest rate r and volatility σ, then the quantizers are computed as follows

$$x_i^k = x_0 \exp\Big((r - \tfrac{\sigma^2}{2})t_k + \sigma\sqrt{t_k} z_i^k\Big).$$

The weights of the Voronoï cells are obtained by the forward Kolmogorov equation (6.54). In the one-dimensional case, they are easily computed relying on the c.d.f. of the Gaussian distribution.

The challenge in this method is the computation of the transition weights p_{ij}^k, which are mandatory for our cause. By optimal quantization, $(\widehat{X}_k)_{0\leq k\leq n}$ is not a Markov chain so one cannot use its distribution to compute p_{ij}^k like for recursive quanization. One usually compute them by Monte Carlo simulations, but, in the one-dimensional case, there exist some closed formulas. In the following, we present such closed formulas in the case of a Black-Scholes model (the case that interests us the most for our numerical examples), i.e. a case where, for an the interest rate r and a volatility σ, the process is given by

$$\widehat{X}_k = \widehat{X}_0 \exp\Big((r - \frac{\sigma^2}{2})t_k + \sigma\sqrt{t_k}\varepsilon_k\Big)$$

where $(\varepsilon_k)_{1\leq k\leq n}$ is an i.i.d. sequence of random variables with distribution $\mathcal{N}(0,1)$.

Exact computation of the transition weights Assume that the quantizers $\Gamma_k = (x_i^k)_{1\leq i\leq N_k}$ of size N_k of $\bar{X}_k$ are already computed for every $k \in \{1,\dots,n\}$ and that the sizes of the grids N_k, $k = 1,\dots,n$, are all equal to $N \in \mathbb{N}$. Note that this hypothesis is not optimal but turns out to be optimal in terms of complexity for a given budget $N_1 + \cdots + N_n$. It is not sharp in terms of error estimates (up to a multiplicative constant) but remains a good compromise which is convenient in practice for the implementation. The goal is to compute the transition weights

$$p_{ij}^k = \mathbb{P}\Big(\widehat{X}_{k+1} = x_j^{k+1} \,|\, \widehat{X}_k = x_i^k\Big) = \frac{\bar{p}_{ij}^k}{p_i^k}$$

where

$$\bar{p}_{ij}^k = \mathbb{P}\Big(\widehat{X}_{k+1} = x_j^{k+1},\, \widehat{X}_k = x_i^k\Big) \qquad \text{and} \qquad p_i^k = \mathbb{P}\big(\widehat{X}_k = x_i^k\big).$$

The weights p_i^k are computed via the forward Kolmogorov equation, using the transition weights p_{ij}^k, as follows

$$p_j^{k+1} = \sum_{i=1}^{N_k} p_{ij}^k\, p_i^k = \sum_{i=1}^{N_k} \bar{p}_{ij}^k,$$

keeping in mind that the Voronoï weight at time t_0 (i.e. $k = 0$) is equal to 1 since $\widehat{X}_0 = X_0 = x_0$ is deterministic. So, our main concern is the computation of $\bar{p}_{ij}^k$ for every $k \in \{1,\dots,n\}$ and $i,j \in \{1,\dots,N\}$. We start by noticing that

$$\widehat{X}_{k+1} = \widehat{X}_k\Big(1 + rh + \sigma\sqrt{h}\varepsilon_k\Big)$$

where $h = \frac{T}{n}$ is the time step of the discretization scheme. Note that highly accurate quantization grids of $\mathcal{N}(0,1)$ for regularly sampled sizes from $N = 1$ to $1\,000$ are available and can be downloaded from the quantization website www.quantize.maths-fi.com (for non-commercial purposes). Then, considering two independent random variables z_1 and z_2 with distribution $\mathcal{N}(0,1)$, one has

$$\begin{aligned}\bar{p}_{ij}^k &= \mathbb{P}\Big(\widehat{X}_{k+1} \in \big[x_{j-\frac{1}{2}}^{k+1}, x_{j+\frac{1}{2}}^{k+1}\big]\,,\ \widehat{X}_k \in \big[x_{i-\frac{1}{2}}^k, x_{i+\frac{1}{2}}^k\big]\Big)\\ &= \mathbb{P}\left(\widehat{X}_k(1 + rh + \sigma\sqrt{h}z_2) \in C_j(\Gamma_{k+1}),\, z_1 \in \big[\underline{x}_i^k,\, \overline{x}_i^k\big]\right)\end{aligned}$$

where

$$\underline{x}_i^k = \frac{\ln\left(x_{i-\frac{1}{2}}^k\right) + \left(\frac{\sigma^2}{2} - r\right)t_k - \ln(x_0)}{\sigma\sqrt{t_k}} \quad \text{and} \quad \overline{x}_i^k = \frac{\ln\left(x_{i+\frac{1}{2}}^k\right) + \left(\frac{\sigma^2}{2} - r\right)t_k - \ln(x_0)}{\sigma\sqrt{t_k}}, \tag{6.59}$$

Then, the independence of z_1 and z_2 yields

$$\begin{aligned}
\bar{p}_{ij}^k &= \int_{\underline{x}_i^k}^{\overline{x}_i^k} \mathbb{P}\Big(x_0(1 + rh + \sigma\sqrt{h}z_2)\exp\Big((r - \tfrac{\sigma^2}{2})t_k + \sigma\sqrt{t_k}z\Big) \in \Big[x_{j-\frac{1}{2}}^{k+1}, x_{j+\frac{1}{2}}^{k+1}\Big]\Big)\, e^{-\frac{z^2}{2}}\frac{dz}{\sqrt{2\pi}} \\
&= \int_{\underline{x}_i^k}^{\overline{x}_i^k} \mathbb{P}\left(z_2 \in \left[\frac{x_{j-\frac{1}{2}}^{k+1} e^{(\frac{\sigma^2}{2}-r)t_k - \sigma\sqrt{t_k}z} - x_0 - rhx_0}{\sigma x_0\sqrt{h}}\,,\ \frac{x_{j+\frac{1}{2}}^{k+1} e^{(\frac{\sigma^2}{2}-r)t_k - \sigma\sqrt{t_k}z} - x_0 - rhx_0}{\sigma x_0\sqrt{h}}\right]\right) e^{-\frac{z^2}{2}}\frac{dz}{\sqrt{2\pi}} \\
&= \int_{\underline{x}_i^k}^{\overline{x}_i^k} \left(\Phi_0(\overline{x}_j^{k+1}) - \Phi_0(\underline{x}_j^{k+1})\right) e^{-\frac{z^2}{2}}\frac{dz}{\sqrt{2\pi}},
\end{aligned} \tag{6.60}$$

where

$$\underline{x}_j^{k+1} = \frac{x_{j-\frac{1}{2}}^{k+1} e^{(\frac{\sigma^2}{2}-r)t_k - \sigma\sqrt{t_k}z} - x_0 - rhx_0}{\sigma x_0\sqrt{h}} \quad \text{and} \quad \overline{x}_j^{k+1} = \frac{x_{j+\frac{1}{2}}^{k+1} e^{(\frac{\sigma^2}{2}-r)t_k - \sigma\sqrt{t_k}z} - x_0 - rhx_0}{\sigma x_0\sqrt{h}}. \tag{6.61}$$

These integrals can be computed via Gaussian quadrature formulas, mainly Gauss-Legendre quadrature formulas for integrals on closed intervals and Gauss-Laguerre quadrature formulas for integrals on semi-closed intervals. So, if $i = 1$ or $i = N$, one uses Gauss-Laguerre formulas since the Voronoï cells (over which we are integrating) are of the form $(-\infty, a)$ or $(a, +\infty)$ for some $a \in \mathbb{R}$. Otherwise, the Voronoï cells are closed intervals so one relies on Gauss-Legendre quadrature formula. Let us detail these computations.

▷ Integration on a closed interval $[a, b]$: Gauss Legendre fomula

Considering $f(z) = \left(\Phi_0(\overline{x}_j^{k+1}) - \Phi_0(\underline{x}_j^{k+1})\right)\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$, $a = \underline{x}_i^k$ and $b = \overline{x}_i^k$, the goal is to compute $I = \int_a^b f(z)dz$. Applying the change of variables $z = \frac{b-a}{2}x + \frac{a+b}{2}$, I can be written and computed as follows

$$I = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}x + \frac{a+b}{2}\right)dx = \frac{b-a}{2}\sum_{i=1}^{n} w_i f\left(\frac{b-a}{2}x_i + \frac{a+b}{2}\right)$$

where $(x_i)_{1\le i\le n}$ are the roots of the n^{th} Legendre polynomial $P_n(x) = \frac{1}{2^n}\sum_{k=0}^{\lfloor \frac{n}{2}\rfloor}(-1)^k\frac{(2n-2k)!}{k!(n-k)!(n-2k)!}x^{n-2k}$ and the weights $(w_i)_{1\le i\le n}$ are given by

$$w_i = \frac{2}{(1-x_i^2)P_n'(x_i)^2} = \frac{2(1-x_i^2)}{(n+1)^2 P_{n+1}(x_i)^2}.$$

▷ Integration on intervals of the form $[a, +\infty)$ or $(-\infty, a]$: Gauss Laguerre quadrature

We consider $f(z) = \Phi_0(\overline{x}_j^{k+1}) - \Phi_0(\underline{x}_j^{k+1})$ and distinguish two cases.

• *Integration on* $[a, +\infty)$

The goal is to compute $I = \int_a^{+\infty} f(z)e^{-\frac{z^2}{2}}dz$ where $a = \underline{x}_i^k$. Applying the change of variables $x = \frac{z^2}{2}$ and denoting $g(x) = \frac{f(x)}{x}$ yield

$$I = \int_{\frac{a^2}{2}}^{+\infty} \frac{f(\sqrt{2x})}{\sqrt{2x}}e^{-x}dx = \int_{\frac{a^2}{2}}^{+\infty} g(\sqrt{2x})e^{-x}dx = e^{-\frac{a^2}{2}}\int_0^{+\infty} g\left(\sqrt{2x + a^2}\right)e^{-x}dx$$

where we applied in the last equality the change of variables $y = x - \frac{a^2}{2}$. Hence, we use Gauss-Legendre quadrature formula to obtain

$$I = e^{-\frac{a^2}{2}} \sum_{i=1}^{N} w_i g\left(\sqrt{2x_i + a^2}\right)$$

where $(x_i)_{1\le i\le n}$ are the roots of the n^{th} Laguerre polynomial $L_n(x) = \sum_{k=0}^{n}(-1)^k \frac{n!}{k!(n-k)!^2}x^k$ and the weights $(w_i)_{1\le i\le n}$ are given by

$$w_i = \frac{1}{(n+1)L_n'(x_i)L_{n+1}(x_i)} = \frac{x_i}{(n+1)^2 L_{n+1}(x_i)^2}. \tag{6.62}$$

• *Integration on* $(-\infty, a]$

The goal is to compute $I = \int_{-\infty}^{a} f(x)e^{-\frac{x^2}{2}}dx$ where $a = \overline{x}_i^k$. Similarly to the previous case, I can be written as follows

$$I = \int_{-a}^{+\infty} f(-x)e^{-\frac{x^2}{2}}dx = \int_{\frac{a^2}{2}}^{+\infty} \frac{f(-\sqrt{2z})}{\sqrt{2z}}e^{-z}dz = \int_{\frac{a^2}{2}}^{+\infty} g(\sqrt{2z})e^{-z}dz = e^{-\frac{a^2}{2}}\int_0^{+\infty} g\left(\sqrt{2z+a^2}\right)e^{-z}dz$$

where $g(x) = \frac{f(-x)}{x}$. Hence, Gauss-Legendre quadrature formula yields

$$I = e^{-\frac{a^2}{2}} \sum_{i=1}^{N} w_i g\left(\sqrt{2x_i + a^2}\right)$$

where $(x_i)_{1\le i\le n}$ are the roots of $L_n(x)$ and $(w_i)_{1\le i\le n}$ are given by (6.62).

Approximation of the transition weights If the goal is not necessarily the highest level of precision, then one approximates the transition weights p_{ij}^k by $g_j(z_i^k)$ where the function $g_j(z)$ is defined by

$$g_j(z) = \Phi_0(\overline{x}_j^{k+1}) - \Phi_0(\underline{x}_j^{k+1}). \tag{6.63}$$

and $\overline{x}_j^{k+1}$ and $\underline{x}_j^{k+1}$ are given by (6.61). In fact, based on (6.60) and then applying Taylor-Lagrange formula, one has

$$\begin{aligned}\bar{p}_{ij}^k &= \int_{z^k_{i-\frac12}}^{z^k_{i+\frac12}} g_j(z)e^{-\frac{z^2}{2}}\frac{dz}{\sqrt{2\pi}}\\ &= g_j(z_i^k)p_i^k + g_j'(z_i^k)\int_{z^k_{i-\frac12}}^{z^k_{i+\frac12}}(z-z_i^k)e^{-\frac{z^2}{2}}\frac{dz}{\sqrt{2\pi}} + \int_{z^k_{i-\frac12}}^{z^k_{i+\frac12}} g_j''(\xi(z))\frac{(z-z_i^k)^2}{2}e^{-\frac{z^2}{2}}\frac{dz}{\sqrt{2\pi}}.\end{aligned}$$

Since $(z_i^k)_{1\le i\le N}$ is a quadratic optimal quantization sequence of the standard Normal distribution, then it is stationary and the second term of the above inequality is equal to 0. Moreover,

$$g_j'(z) = \frac{k}{x_0\sqrt{2\pi}}\left(x_{j+\frac12}e^{-\sigma\sqrt{t_k}z-\frac12\bar{x}_j^2} - x_{j-\frac12}e^{-\sigma\sqrt{t_k}z-\frac12\underline{x}_j^2}\right)$$

and

$$g_j''(z) = \frac{k}{x_0\sqrt{2\pi}}\left[\sigma\sqrt{t_k}e^{-\sigma\sqrt{t_k}z}\left(x_{j-\frac12}e^{-\frac12\underline{x}_j^2} - x_{j+\frac12}e^{-\frac12\overline{x}_j^2}\right) + \frac{k}{x_0}e^{-2\sigma\sqrt{t_k}z}\left(x_{j+\frac12}^2\overline{x}_j^2e^{-\frac12\overline{x}_j^2} - x_{j-\frac12}^2\underline{x}_j^2e^{-\frac12\underline{x}_j^2}\right)\right].$$

At this stage, one notices that $\gamma(z) := \exp(-2z - \frac12 e^{-2z}) \le \kappa$ for every $z \in \mathbb{R}$ for some finite positive constant κ and that $|g_j''(z)| \le \bar{\kappa}$ for a finite positive constant $\bar{\kappa}$. Consequently, $\left|p_{ij}^k - g_j(z_i^k)\right|$ is bounded.

It is important to note that when we estimate the transition weight by $g_j(z_i^k)$, we formally get the transition weight from x_i^k to x_j^{k+1} obtained by recursive quantization, even though they are not the same grids.

Remark 6.6.1. *For local volatility models (CEV models for example), it becomes more complicated to establish such closed formulas for the computations of the transition matrix. One tends to approximate them by Monte Carlo simulations, for example.*

Greedy quantization

Another technique is greedy vector quantization introduced in [45] and developed in [24]. It consists in building a *sequence* of points $(a_n)_{n\geq 1}$ in $\mathbb{R}^d$ recursively optimal step by step, in the following *greedy* sense: having computed the first n points $a_1, \ldots, a_n$ of the sequence and defining the resulting grid $a^{(n)} = \{a_1, \ldots, a_n\}$ for $n \geq 1$, we compute the $(n+1)$-th point as a solution to the minimization problem

$$a_{n+1} \in \operatorname{argmin}_{\xi \in \mathbb{R}^d} e_p(a^{(n)} \cup \{\xi\}, X), \tag{6.64}$$

with the convention $a^{(0)} = \varnothing$. Quadratic greedy quantization sequences are obtained by implementing "freezing" avatars of usual stochastic optimization algorithms used for optimal quantization, these variants are exposed in details in [46]. In this paragraph, we give a quick idea on the computation of the greedy quantization sequence of $(\bar{X}_k)_{0\leq k\leq n}$. Starting at $\widehat{X}_0 = \bar{X}_0 = x_0$, the process $\bar{X}_k$ can be written, for every $k \in \{1, \ldots, n\}$, as follows

$$\bar{X}_k = x_0 + t_k b(x_0) + \sqrt{t_k}\sigma(x_0)\varepsilon_k$$

where ε_k is a random variable with distribution $\mathcal{N}(0, I_q)$. So $\bar{X}_k$ is with Normal distribution $\mathcal{N}(m_k, \Sigma_k)$ where $m_k = x_0 + t_k b(x_0)$ and $\Sigma_k = \sqrt{t_k}\sigma(x_0)$ and hence this is the distribution that needs to be discretized by greedy quantization. The transition weights in the one-dimensional case are computed via Gaussian quadrature formula like explained for the optimal quantization, and the weights of the Voronoï cells by the forward Kolmogorov equation.

In the high-dimensional framework ($d > 1$), the computations become too demanding. So, instead of designing pure greedy quantization sequences, one tends to build greedy product quantization sequences which are obtained as a result of the tensor product of one-dimensional sequences, when the target law is a tensor product of its independent marginal laws. We refer to [24] for further details.

Greedy recursive quantization

In the algorithm described in Section 6.5, the recursive quantization scheme (6.22) is based on an optimal quantization of the sequences $(\widetilde{X}_k)_{0\leq k\leq n}$ at each time step t_k. Here, we consider, as an alternative, greedy optimal quantization grids $\widehat{X}_k$ of $\widetilde{X}_k$. They are designed as follows: At time t_{k+1}, assuming that the N_k-tuple $(x_1^k, \ldots, x_{N_k}^k)$ and its companion parameters are already computed, one needs to build, step by step by greedy quantization, the N_{k+1}-tuple $(x_1^{k+1}, \ldots, x_{N_{k+1}}^{k+1})$ which approaches best $\widetilde{X}_{k+1} = \mathcal{E}_k(\widehat{X}_k, \varepsilon_{k+1})$. Since $\mathcal{E}_k(x_i^k, \varepsilon_{k+1}) \sim \mathcal{N}(m_i^k, \Sigma_i^k)$ with $m_i^k = x_i^k + \Delta b_k(x_i^k)$ and $\Sigma_i^k = \sqrt{\Delta}\sigma_k(x_i^k)$, the first point of the sequence is $x_1^{k+1} = \mathbb{E}\big[\widehat{X}_k + \Delta b_k(\widehat{X}_k)\big] = \sum_{i=1}^{N_k} p_i^k \big(x_i^k + \Delta b_k(x_i^k)\big)$ and then, at each iteration N, $N \in \{2, \ldots, N_{k+1}\}$, one adds one point x_N^{k+1} following the steps of the greedy variant of Lloyd's algorithm detailed in [46]. One should take in consideration that the local interpoint inertia are computed, at each time step t_{k+1}, by

$$\sigma_j^2 = \sum_{i=1}^{N_k} p_i^k \left(\int_{x_j^{k+1,N}}^{x_{j+\frac{1}{2}}^{k+1,N}} \big(\xi - x_j^{k+1,N}\big)^2 P(d\xi) + \int_{x_{j+\frac{1}{2}}^{k+1,N}}^{x_{j+1}^{k+1,N}} \big(\xi - x_{j+1}^{k+1,N}\big)^2 P(d\xi) \right) := \sum_{i=1}^{N_k} p_i^k s_{ij} \tag{6.65}$$

where $x_{j+\frac{1}{2}}^{k+1,N} = \frac{x_j^{k+1,N}+x_{j+1}^{k+1,N}}{2}$ with $x_0^{k+1,N} = x_{\frac{1}{2}}^{k+1,N} = -\infty$ and $x_N^{k+1,N} = x_{N-\frac{1}{2}}^{k+1,N} = +\infty$. Likewise, the recurrence of the algorithm is given by

$$x_{\ell+1} = \frac{\sum_{i=1}^{N_k} p_i^k \mathbb{E}\Big(\mathcal{E}_k(x_i^k, \varepsilon_{k+1}) \mathbb{1}_{\left\{\mathcal{E}_k(x_i^k,\varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\right\}}\Big)}{\sum_{i=1}^{N_k} p_i^k \mathbb{P}\Big(\mathcal{E}_k(x_i^k, \varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\Big)}, \tag{6.66}$$

The companion parameters are computed following the same principle as for the standard recrusive quantization.

6.6.2 Examples

American call option in a market with bid-ask spread on interest rates

We are interested in the valuation of an American call option with maturity T in a market with a bid-ask spread on interest rates with a borrowing rate R and a lending rate $r \leq R$. The stock price is represented by the process $(X_t)_{t\in[0,T]}$ given by the SDE (6.3) and the dynamics of the portfolio are given by

$$-dY_t = \left(-rY_t - \frac{b_t(X_t) - r}{\sigma_t(X_t)} Z_t - (R-r)\min\left(Y_t - \frac{Z_t}{\sigma_t(X_t)}, 0\right)\right) dt - Z_t dW_t$$

$$Y_T = h(X_T) \qquad \text{and} \qquad Y_t \geq g(X_t)$$

where $h(x) = g(x) = \max(x - K, 0)$, K being the strike price.

Black-Scholes model We consider that $(X_t)_{t\in[0,T]}$ evolves following the Black-Scholes dynamics and is time discretized following the Euler scheme, i.e. for every $k \in \{0, \dots, n-1\}$,

$$\bar{X}_{k+1} = \bar{X}_k + \mu \Delta \bar{X}_k + \sigma\sqrt{\Delta}\bar{X}_k\, \varepsilon_{k+1} \tag{6.67}$$

where μ is the drift and σ is the volatility. The space discretization is established via recursive quantization (RQ), optimal quantization (OQ), greedy quantization (GQ) and greedy recursive quantization (GRQ). We consider $n = 20$ time steps and build corresponding quantization grids of size $N = 100$ and their companion parameters as explained in the different sections previously in the chapter. Then, we rely on the backward recursion (6.58) to compute the value Y_0 of the underlying option. Note that the quantities π_{ij}^k are computed, for every $k \in \{1, \dots, n\}$, as a companion parameter with the diffusion $\widehat{X}_k$ via a Monte Carlo simulation of size 10^6. We consider the following parameters

$$X_0 = 100\,, \quad T = 0.25\,, \quad \sigma = 0.2\,, \quad \mu = 0.05\,, \quad r = 0.01\,, \quad R = 0.06$$

and we compare the values obtained by the different methods for different values of K varying between 100 and 120. As a benchmark, we will assume that the optimal quantization converges to the exact value and, under this hypothesis, we consider the fastest and most accurate version of optimal quantization, which is the quantization-based Richardson-Romberg extrapolation. The idea is the following:
If the goal is to approximate $\mathbb{E}f(X)$ for a function f and a random variable X, one considers two optimal quantization sequences $\widehat{X}^{N_1}$ of size N_1 and $\widehat{X}^{N_2}$ of size N_2 of the random variable X and hence $\mathbb{E}f(X)$ is given by

$$\mathbb{E}f(X) = \frac{N_2^2 \mathbb{E}f(\widehat{X}^{N_2}) - N_1^2 \mathbb{E}f(\widehat{X}^{N_1})}{N_2^2 - N_1^2}. \tag{6.68}$$

From a practical point of view, one usually considers $N_1 = N$ and $N_2 = \frac{N}{2}$. Furthermore, when the dimension $d = 1$, the standard quantization error is of the form

$$e_2(X, \mu) \approx c_1\sqrt{n} + c_2\sqrt{n}N^{-1}$$

and the Romberg-quantization error is of the form

$$e_2(X,\mu) \approx c_2\sqrt{n}\left(\frac{1}{N_1}-\frac{1}{N_2}\right) \approx \frac{c_1\sqrt{n}}{2N_1}.$$

So, by studying the values of this error for different values of n and N_1, we realize that the best technique is to consider a small number of time steps n and a large size N of the quantizer.

In our example, we consider an optimal quantization-based Richardson Romberg extrapolation with $n = 5$ and $N = 1\,000$. We observe in Table 6.1 the results and the errors obtained by the various methods. Here, we emphasize on the computational time of these simulations which are performed on a CPU 2.7

K	**RQ**		**GRQ**		**OQ**		**GQ**		**Romberg**
	Value	Error	Value	Error	Value	Error	Value	Error	
100	4.719	0.026	4.728	0.017	4.747	0.002	4.704	0.041	4.745
105	2.538	0.012	2.548	0.002	2.561	0.011	2.529	0.021	2.55
110	1.222	0.003	1.225	0.006	1.234	0.015	1.212	0.007	1.219
115	0.526	0.008	0.526	0.008	0.532	0.014	0.518	0	0.518
120	0.203	0.007	0.202	0.006	0.206	0.01	0.198	0.002	0.196
Average		0.0112		0.0078		0.0104		0.0142	

Table 6.1: Pricing of an American call option in a market with bid-ask spread for interest rates in a Black-Scholes model by recursive (RQ), greedy recursive (GRQ), optimal (OQ) and greedy (GQ) quantization.

GHz and 8 GB memory computer. The optimal quantizer and its companion parameters are obtained in about 40 seconds while the greedy quantization sequence and its companions in about 30 seconds. This is approximately a 25% gain in time in favor of greedy quantization whose results are comparable (a little less precise) than optimal quantization. As for the recursive quantization, the standard simulations (RQ) are obtained in about 2.3 minutes and the greedy simulations (GRQ) in about 2 minutes. Hence, the greedy character introduced in the recursive algorithm brings a 13% gain in time. The additional cost in time is compensated by the preservation of the Markovian property and the precision of the results.
Figure 6.1 depicts the convergence of the error induced by the approximation of Y_0 based on a recursive quantization of the forward process $\bar{X}_k$. For this illustration, we consider a strike $K = 100$ and we make the size N of the grids vary between 10 and 100. The graph is represented in a log-log-scale scale and an $\mathcal{O}(N^{-1})$ rate of convergence is clearly observed.

CEV model We consider a local volatility model, the CEV model, in which $(X_t)_{0\le t\le T}$ evolves following

$$dX_t = \mu X_t dt + \vartheta X_t^{\delta} dW_t, \qquad X_0 = x_0, \tag{6.69}$$

for some $\delta \in (0,1)$ and $\vartheta \in (0,\overline{\vartheta}]$ with $\overline{\vartheta} > 0$. $\sigma(x) = \vartheta x^{\delta}$ is the local volatility function. The discretized Euler scheme associated to $(X_t)_{t\in[0,T]}$ is given, for every $k \in \{0,\dots,n-1\}$, by

$$\bar{X}_{k+1} = \bar{X}_k + \mu\Delta\bar{X}_k + \vartheta\bar{X}_k^{\delta}\sqrt{\Delta}\,\varepsilon_k \tag{6.70}$$

where $(\varepsilon_k)_{1\le k\le n}$ is an i.i.d sequence of random variables with distribution $\mathcal{N}(0,1)$.
The construction of the quantizers and the computation of the companion parameters by recursive and greedy recursive quantization is similar to what was done for the Black-Scholes model. As for optimal and greedy quantization, closed forms for the companion parameters are no longer available in this model, we estimate them by Monte Carlo simulations of size 10^5 coupled with a nearest neighbor search. We build

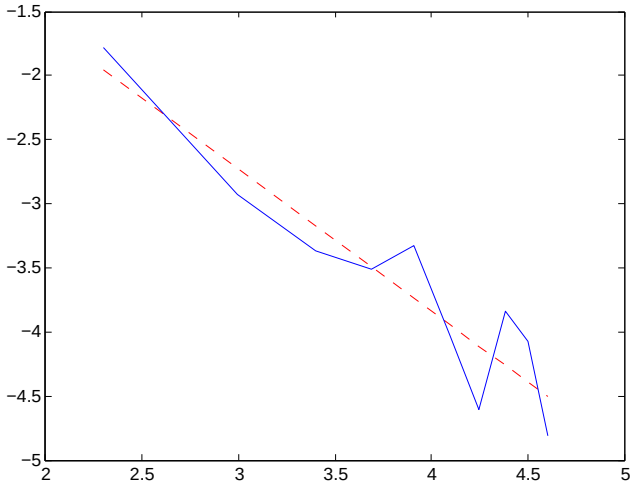

Figure 6.1: Convergence rate of the error induced by the approximation of the Bid-ask spread Call option in a Black-Scholes model discretized by recursive quantization for different sizes $N = 10, \ldots, 100$.

corresponding quantization grids of size $N = 150$ and consider $n = 15$ time steps. The parameters are the following

$$X_0 = 100\,, \quad T = 0.25\,, \quad \vartheta = 4\,, \quad \delta = 0.5\,, \quad \varepsilon = 1\,, \quad \mu = 0.05\,, \quad r = 0.01\,, \quad R = 0.06$$

and we compare the values obtained by the different methods for different values of K between 100 and 120. The benchmark is given by an optimal quantization-based Richardson-Romberg extrapolation (6.68). We observe in Table 6.2 the results and errors obtained by such comparisons. As for the computation time, we note that the optimal quantizer and its companion parameters are obtained in about 100 seconds while the greedy quantization sequence and its companions in about 70 seconds. The fact that these computations take more time for the CEV model than for the Black-Scholes model is due to the non-existence of closed formulas for the computation of the companion parameters in the CEV model, the computation of the quantizers themselves is almost instantaneous. Moreover, the recursive quantizer and its companions are computed in about 3.5 minutes while the greedy recursive quantizers in about 3 minutes.

K	**RQ**		**GRQ**		**OQ**		**GQ**		**Romberg**
	Value	Error	Value	Error	Value	Error	Value	Error	
100	8.517	0.074	8.524	0.067	8.536	0.055	8.593	0.002	8.591
105	6.262	0.049	6.272	0.039	6.288	0.023	6.321	0.01	6.311
110	4.479	0.023	4.483	0.019	4.498	0.004	4.522	0.02	4.502
115	3.11	0.006	3.113	0.003	3.125	0.009	3.128	0.012	3.116
120	2.094	0.003	2.1	0.009	2.109	0.018	2.103	0.012	2.091
Average		0.031		0.0274		0.0218		0.0112	

Table 6.2: Pricing of an American call option in a market with bid-ask spread for interest rates in a CEV model by recursive (RQ), greedy recursive (GRQ), optimal (OQ) and greedy (GQ) quantization.

Two-dimensional American exchange options

We are interested in pricing an American exchange option with exchange rate μ and maturity T. This price is given by the value Y_0 at time t_0 of the solution of the RBSDE (6.1) with driver $f = 0$ and $h_t(x) = g_t(x) = \max\left(e^{-\lambda t}X_t^1 - \mu X_t^2, 0\right)$. X_t^1 and X_t^2 are two assets, such that X_t^1 is with a geometric dividend rate λ and X_t^2 is without dividend, both following a Black-Scholes model. The discretized Euler scheme $(\bar{X}_k^1, \bar{X}_k^2)$ is given, for every $k \in \{0, \dots, n-1\}$, by

$$\begin{aligned} \bar{X}_{k+1}^1 &= \bar{X}_k^1 e^{(r-\frac{\sigma^2}{2})\Delta + \sigma\sqrt{\Delta}\varepsilon_k^1} \\ \bar{X}_{k+1}^2 &= \bar{X}_k^2 e^{(r-\frac{\sigma^2}{2})\Delta + \sigma\sqrt{\Delta}(\rho\varepsilon_k^1 + \sqrt{1-\rho^2}\varepsilon_k^2)} \end{aligned}$$

where r is the interest rate, σ the volatility, ρ is a correlation coefficient and $(\varepsilon_k^1, \varepsilon_k^2)_{1\le k\le n}$ is a sequence of i.i.d. random variables with distribution $\mathcal{N}(0, I_2)$.
From a numerical point of view, we discretize in $n = 10$ time steps, build quantizers of size $N_X = 100$ and consider the following parameters

$$X_0^1 = 40\,, \quad T = 1\,, \quad r = 0\,, \quad \sigma = 0.2\,, \quad \lambda = 0.05\,, \quad \mu = 1\,.$$

In high dimensions ($d > 1$), the implementation of the recursive quantization algorithm is too expensive and its cost in time is very high. We consider, instead, the hybrid recursive quantization, introduced in Section 6.2.3 and use sequences of optimal quantizers $(\hat{\varepsilon}_l^k)_{1\le l\le N_\varepsilon}$ of size $N^\varepsilon = 1000$ to compute the sequence and the companion parameters as detailed in Section 6.5. We also build optimal quantizers and greedy product quantization sequences (see Section 6.6.1). We compute the price of the option by these methods for $X_0^2 \in \{36; 44\}$ and $\rho \in \{-0.8; 0; 0.8\}$ and compare the results obtained to those computed by a finite difference algorithm in [74] and expose the errors hence induced in Table 6.3.
Similarly to the one-dimensional Example 6.6.2, a gain in the computation time appears in favor of the

X_0^2	ρ	**OQ**		**HRQ**		**GPQ**		**Benchmark**
		Value	Error	Value	Error	Value	Error	
36	-0.8	7.062	0.087	6.979	0.004	6.926	0.049	6.975
36	0	5.832	0.186	5.706	0.06	5.763	0.117	5.646
36	0.8	4.076	0.076	4.008	0.008	4	0	4
Average error			0.116		0.024		0.055	
44	-0.8	3.834	0.065	3.741	0.028	3.609	0.16	3.769
44	0	2.453	0.117	2.329	0.007	2.042	0.294	2.336
44	0.8	0.426	0.067	0.282	0.077	0.401	0.042	0.359
Average error			0.083		0.037		0.165	

Table 6.3: Pricing of an American exchange option for $d = 2$ in a BS model by hybrid recursive (HRQ), optimal (OQ) and greedy product quantization (GPQ).

greedy quantization. In fact, greedy product quantization sequences are obtained in about 55 seconds whereas optimal and hybrid recursive quantizers in about 70 seconds and 3.75 minutes respectively, and hence the gain is about 20% compared to optimal quantization and 75% compared to hybrid recursive quantization. Moreover, we remark that hybrid recursive quantization gives the most precise results while an expected gain in precision for optimal quantization compared to greedy quantization is observed.

6.7 Appendix

6.7.1 Appendix A: The proof of Lemma 6.2.3

First note that the function $f : u \mapsto |u|^r$ satisfies (since $r \geq 2$)

$$\nabla |u|^r = r|u|^{r-1}\frac{u}{|u|} \quad \text{and} \quad \nabla^2|u|^r = r|u|^{r-2}\left((r-2)\frac{u}{|u|}\frac{u^*}{|u|} + I_d\right)$$

(convention $\frac{0}{|0|} = 0$). Consequently, Taylor's Theorem with Lagrange remainder applied to f reads

$$f(u+v) = f(u) + \langle \nabla f(u), v\rangle + \frac{1}{2}v^*\nabla^2 f(\xi_{u,v})v$$

for some $\xi_{u,v} = \lambda_{u,v}u + (1-\lambda_{u,v})(u+v)$, $\lambda_{u,v} \in (0,1)$. Note that

$$v^*\nabla^2 f(\xi_{u,v})v = r|\xi_{u,v}|^{r-2}\Big((r-2)\frac{\langle v, \xi_{u,v}\rangle^2}{|\xi_{u,v}|^2} + |v|^2\Big) \leq r|\xi_{u,v}|^{r-2}(r-1)|v|^2$$

owing to Cauchy-Schwartz inequality. Then, noting that $|\xi_{u,v}| \leq |u| \vee |u+v| \leq |u| + |v|$, we obtain

$$\begin{aligned}
|u+v|^r &\leq |u|^r + \big\langle r|u|^{r-1}\frac{u}{|u|}, v\big\rangle + \frac{r(r-1)}{2}\big(|u|+|v|\big)^{r-2}|v|^2 \\
&\leq |u|^r + \big\langle r|u|^{r-1}\frac{u}{|u|}, v\big\rangle + \frac{r(r-1)}{2}2^{(r-3)_+}\big(|u|^{r-2} + |v|^{r-2}\big)|v|^2 \\
&= |u|^r + \big\langle r|u|^{r-1}\frac{u}{|u|}, v\big\rangle + \frac{r(r-1)}{2}2^{(r-3)_+}\big(|u|^{r-2}|v|^2 + |v|^r\big).
\end{aligned}$$

Applying the above inequality to $u = a$ and $v = \sqrt{h}\,AZ$ yields

$$\big|a + A\sqrt{h}Z\big|^r \leq |a|^r + r\Big\langle |a|^{r-1}\frac{a}{|a|}, A\sqrt{h}\,Z\Big\rangle + 2^{(r-3)_+}\frac{r(r-1)}{2}\big(h|a|^{r-2}|AZ|^2 + h^{\frac{r}{2}}|AZ|^r\big).$$

Applying Young's inequality (when $r > 2$) to the product $|a|^{r-2}|AZ|^2$ with conjugate exponents $r' = \frac{r}{r-2}$ and $s' = \frac{r}{2}$ yields

$$\begin{aligned}
\big|a + A\sqrt{h}Z\big|^r &\leq |a|^r + r\Big\langle |a|^{r-1}\frac{a}{|a|}, A\sqrt{h}\,Z\Big\rangle + 2^{(r-3)_+}\frac{r(r-1)}{2}\Big(\frac{h}{r}\big((r-2)|a|^r + 2|AZ|^r\big) + h^{\frac{r}{2}}|AZ|^r\Big) \\
&\leq |a|^r\left(1 + 2^{(r-3)_+}\frac{(r-1)(r-2)}{2}h\right) + r\Big\langle |a|^{r-1}\frac{a}{|a|}, A\sqrt{h}\,Z\Big\rangle \\
&\quad + 2^{(r-3)_+}(r-1)h\|A\|^r|Z|^r\Big(1 + \frac{r}{2}\,h^{\frac{r-2}{2}}\Big). \qquad (6.71)
\end{aligned}$$

Finally taking expectation and using that $\mathbb{E}\,Z = 0$ and $h < h_0$ yields the announced result.

6.7.2 Appendix B: Proof of Theorem 6.3.1

To get into the core of the proof of the first part of Theorem 6.3.1, we need to show some properties of the functions $\bar{y}_k$ and $\bar{z}_k$.

Lemme 6.7.1. *The functions $\bar{y}_k$ and $\bar{z}_k$ defined by (6.36)-(6.37) are Lipschitz continuous with $[\bar{y}_k]_{\mathrm{Lip}}$ and $[\bar{z}_k]_{\mathrm{Lip}}$ their respective Lipschitz coefficients given by*

$$[\bar{y}_k]_{\mathrm{Lip}} \leq [h]_{\mathrm{Lip}} + \Delta_{\max}(1+\Delta_{\max})[f]_{\mathrm{Lip}} + e^{(1+C_f+C_{b,\sigma})\Delta_{\max}}[\bar{y}_{k+1}]_{\mathrm{Lip}}$$

and

$$[\bar{z}_k]_{\mathrm{Lip}} \leq \frac{1}{\sqrt{\Delta}}[\bar{y}_{k+1}]_{\mathrm{Lip}}e^{C_{b,\sigma}\Delta}$$

where $C_{b,\sigma} = 1 + \Delta_{\max}(2[b_k]_{\mathrm{Lip}} + [\sigma_k]_{\mathrm{Lip}}) + \Delta^2_{\max}[b_k]^2_{\mathrm{Lip}}$ and $C_f = 2[f]_{\mathrm{Lip}} + [f]^2_{\mathrm{Lip}}$.

Proof. **STEP 1:** We show that $\bar{y}_k$ and $\widetilde{y}_k$ are Lipschitz continuous. We rely on a backward induction. In this part, we denote $\mathcal{E}_k^x = \mathcal{E}_k(x, \varepsilon_{k+1})$ for every x to alleviate notations. It is clear that $[\bar{y}_n]_{\mathrm{Lip}} = [g]_{\mathrm{Lip}}$. We assume that $\bar{y}_{k+1}$ is $[\bar{y}_{k+1}]_{\mathrm{Lip}}$-Lipschitz continuous and show the Lipschitz continuity of $\bar{y}_k$. For every x, x', we start by noticing that

$$|\widetilde{y}_k(x) - \widetilde{y}_k(x')| = \Big|\mathbb{E}_k \bar{y}_{k+1}\big(\mathcal{E}_k^x\big) - \mathbb{E}_k \bar{y}_{k+1}\big(\mathcal{E}_k^{x'}\big) + \Delta\Big(A_k(x-x') + B_k \mathbb{E}_k\Big(\bar{y}_{k+1}\big(\mathcal{E}_k^x\big) - \bar{y}_{k+1}\big(\mathcal{E}_k^{x'}\big)\Big) + \frac{C_k}{\sqrt{\Delta}} \mathbb{E}_k\Big(\bar{y}_{k+1}\big(\mathcal{E}_k^x\big) - \bar{y}_{k+1}\big(\mathcal{E}_k^{x'}\big)\Big)\varepsilon_{k+1}\Big)\Big|$$

where

$$A_k = \frac{\mathcal{E}_k\big(x, \mathbb{E}_k\, \bar{y}_{k+1}(\mathcal{E}_k^x), \bar{z}_k(x)\big) - \mathcal{E}_k\big(x', \mathbb{E}_k\, \bar{y}_{k+1}(\mathcal{E}_k^x), \bar{z}_k(x)\big)}{x - x'}\, \mathbb{1}_{x \neq x'},$$

$$B_k = \frac{\mathcal{E}_k\big(x', \mathbb{E}_k\, \bar{y}_{k+1}(\mathcal{E}_k^x), \bar{z}_k(x)\big) - \mathcal{E}_k\big(x', \mathbb{E}_k\, \bar{y}_{k+1}(\mathcal{E}_k^{x'}), \bar{z}_k(x)\big)}{\mathbb{E}_k\big(\bar{y}_{k+1}\big(\mathcal{E}_k^x\big) - \bar{y}_{k+1}\big(\mathcal{E}_k^{x'}\big)\big)}\, \mathbb{1}_{\mathbb{E}_k\, \bar{y}_{k+1}\left(\mathcal{E}_k^x\right) \neq \mathbb{E}\, \bar{y}_{k+1}\left(\mathcal{E}_k^{x'}\right)},$$

$$C_k = \frac{\mathcal{E}_k\Big(x', \mathbb{E}_k\, \bar{y}_{k+1}(\mathcal{E}_k^{x'}), \bar{z}_k(x)\Big) - \mathcal{E}_k\big(x', \mathbb{E}_k\, \bar{y}_{k+1}(\mathcal{E}_k^{x'}), \bar{z}_k(x')\big)}{\mathbb{E}_k\big(\bar{y}_{k+1}\big(\mathcal{E}_k^x\big) - \bar{y}_{k+1}\big(\mathcal{E}_k^{x'}\big)\varepsilon_{k+1}\big)}\, \mathbb{1}_{\bar{z}_k(x) \neq \bar{z}_k(x')}.$$

It is clear that these quantities are $\mathcal{F}_{t_k}$-measurable and that $\max\big(|A_k|, |B_k|, |C_k|\big) \leq [f]_{\mathrm{Lip}}$ so

$$|\widetilde{y}_k(x) - \widetilde{y}_k(x')| \leq \Delta[f]_{\mathrm{Lip}}|x - x'| + \mathbb{E}_k\Big|\Big(\bar{y}_{k+1}(\mathcal{E}_k^x) - \bar{y}_{k+1}(\mathcal{E}_k^{x'})\Big)\Big(1 + \Delta B_k + C_k\sqrt{\Delta}\varepsilon_{k+1}\Big)\Big|.$$

Now, using the inequality $(a+b)^2 \leq a^2(1+\Delta) + b^2(1+\frac{1}{\Delta})$, one obtains

$$\begin{aligned}\big|\widetilde{y}_k(x) - \widetilde{y}_k(x')\big|^2 &\leq \Delta^2[f]_{\mathrm{Lip}}^2|x-x'|^2(1+\tfrac{1}{\Delta}) + (1+\Delta)\mathbb{E}_k\Big|\big(\bar{y}_{k+1}(\mathcal{E}_k^x) - \bar{y}_{k+1}(\mathcal{E}_k^{x'})\big)\big(1 + \Delta B_k + C_k\sqrt{\Delta}\varepsilon_{k+1}\big)\Big|^2 \\ &\leq \Delta(1+\Delta)[f]_{\mathrm{Lip}}^2|x-x'|^2 + (1+\Delta)\mathbb{E}_k\Big|\bar{y}_{k+1}(\mathcal{E}_k^x) - \bar{y}_{k+1}(\mathcal{E}_k^{x'})\Big|^2 \mathbb{E}_k\big(1 + \Delta B_k + C_k\sqrt{\Delta}\varepsilon_{k+1}\big)^2.\end{aligned}$$

Since, $(\varepsilon_k)_{k\geq 0}$ is a sequence of i.i.d. random variables, then

$$\mathbb{E}_k\Big(1 + \Delta B_k + C_k\sqrt{\Delta}\varepsilon_{k+1}\Big)^2 = (1 + [f]_{\mathrm{Lip}}\Delta)^2 + \Delta[f]_{\mathrm{Lip}}^2\mathbb{E}|\varepsilon_{k+1}|^2 \leq 1 + 2\Delta[f]_{\mathrm{Lip}} + \Delta[f]_{\mathrm{Lip}}^2 \leq e^{C_f\Delta},$$

so that

$$|\widetilde{y}_k(x) - \widetilde{y}_k(x')|^2 \leq \Delta(1+\Delta)[f]_{\mathrm{Lip}}^2|x-x'|^2 + e^{(C_f+1)\Delta}\mathbb{E}_k\Big|\bar{y}_{k+1}(\mathcal{E}_k^x) - \bar{y}_{k+1}(\mathcal{E}_k^{x'})\Big|^2.$$

At this stage, one notes that if $a, b \geq 0$, then $\max(a,b)^2 \leq \max(a^2, b^2)$ so

$$\begin{aligned}|\bar{y}_k(x) - \bar{y}_k(x')|^2 &\leq \max\Big(|h_k(x) - h_k(x')|^2, |\widetilde{y}_k(x) - \widetilde{y}_k(x')|^2\Big) \\ &\leq \max\Big([h]_{\mathrm{Lip}}^2|x-x'|^2, \Delta(1+\Delta)[f]_{\mathrm{Lip}}^2|x-x'|^2 + e^{\Delta(1+C_f)}\mathbb{E}_k\Big|\bar{y}_{k+1}(\mathcal{E}_k^x) - \bar{y}_{k+1}(\mathcal{E}_k^{x'})\Big|^2\Big)\end{aligned}$$

We use the fact that $\bar{y}_{k+1}$ is Lipschitz continuous and write

$$\begin{aligned}\mathbb{E}_k\Big|\bar{y}_{k+1}(\mathcal{E}_k^x) - \bar{y}_{k+1}(\mathcal{E}_k^{x'})\Big|^2 &\leq [\bar{y}_{k+1}]_{\mathrm{Lip}}\mathbb{E}|x - x' + \Delta(b_k(x) - b_k(x')) + \sqrt{\Delta}(\sigma_k(x) - \sigma_k(x'))\varepsilon_{k+1}|^2 \\ &\leq [\bar{y}_{k+1}]_{\mathrm{Lip}}|x-x'|^2(1 + \Delta(2[b_k]_{\mathrm{Lip}} + [\sigma_k]_{\mathrm{Lip}}) + \Delta^2[b_k]_{\mathrm{Lip}}^2) \\ &\leq [\bar{y}_{k+1}]_{\mathrm{Lip}}\, e^{C_{b,\sigma}\Delta}|x-x'|^2 \qquad (6.72)\end{aligned}$$

where $C_{b,\sigma} = 2[b_k]_{\text{Lip}} + [\sigma_k]_{\text{Lip}} + \Delta_{\max}[b_k]^2_{\text{Lip}}$. Therefore, one has

$$|\bar{y}_k(x) - \bar{y}_k(x')|^2 \le \max\Big([h]^2_{\text{Lip}}|x-x'|^2, \Delta(1+\Delta)[f]_{\text{Lip}}|x-x'|^2 + e^{(1+C_f+C_{b,\sigma})\Delta}[\bar{y}_{k+1}]_{\text{Lip}}|x-x'|^2\Big).$$

Now, since $\Delta \le \Delta_{\max}$, one deduces that $\bar{y}_k$ is $[\bar{y}_k]_{\text{Lip}}$-Lipschitz continuous with

$$[\bar{y}_k]_{\text{Lip}} \le [h]_{\text{Lip}} + \Delta_{\max}(1+\Delta_{\max})[f]_{\text{Lip}} + e^{(1+C_f+C_{b,\sigma})\Delta_{\max}}[\bar{y}_{k+1}]_{\text{Lip}}.$$

STEP 2: For the Lipschitz continuity of $\bar{z}_k$, we will use the same property of $\bar{y}_{k+1}$, more precisely inequality (6.72). For every $x, x' \in \mathbb{R}^d$,

$$|\bar{z}_k(x) - \bar{z}_k(x')|^2 \le \frac{1}{\sqrt{\Delta}}\mathbb{E}\Big|(\bar{y}_{k+1}(\mathcal{E}^x_k) - \bar{y}_{k+1}(\mathcal{E}^{x'}_k))\varepsilon_{k+1}\Big| \le \frac{1}{\sqrt{\Delta}}[\bar{y}_{k+1}]_{\text{Lip}}e^{C_{b,\sigma}\Delta}|x-x'|^2.$$

□

Proof of theorem 6.3.1. We denote $\delta V_t = V_t - \bar{V}_t$ for any process V. We consider the following stopping times

$$\tau^c = \inf\Big\{u \ge t;\ \int_t^u \mathbb{1}_{\delta Y_s > 0} dK_s > 0\Big\} \wedge T, \tag{6.73}$$

$$\tau^d = \min\Big\{t_j \ge t;\ \mathbb{1}_{\delta Y_i < 0}\big(h_i(\bar{X}_i) - \widetilde{Y}_i\big)_+ > 0\Big\} \wedge T \tag{6.74}$$

and

$$\tau = \tau^c \wedge \tau^d.$$

Keeping in mind that $(\bar{Y}_t)_t$ is a càglàd process (see (6.41)), we use Itô's formula between t and τ to write

$$\begin{aligned}
|\delta Y_\tau|^2 &= |\delta Y_t|^2 + 2\int_{[t,\tau)} \delta Y_s d\delta Y_s + \int_{[t,\tau)} |\delta Z^2_s| ds + \sum_{t\le s<\tau}(\delta Y_s - \delta Y_{s^-})^2 \\
&= |\delta Y_t|^2 - 2\int_{[t,\tau)} \delta Y_s\big(f(\Theta_s) - f(\bar{\Theta}_{\underline{s}})\big)ds - 2\int_{[t,\tau)} \delta Y_s dK_s + 2\int_{[t,\tau)} \delta Y_s d\bar{K}_s \\
&\quad + \int_{[t,\tau)} (Z_s - \bar{Z}_s)dW_s + \int_{[t,\tau)} |\delta Z^2_s| ds + \sum_{t\le s<\tau}(\delta Y_s - \delta Y_{s^-})^2
\end{aligned}$$

where $\Theta_s = (X_s, Y_s, Z_s)$, $\bar{\Theta}_{\underline{s}} = (\bar{X}_{\underline{s}}, \mathbb{E}_{\underline{s}}\bar{Y}_{\bar{s}}, \bar{\zeta}_{\underline{s}})$, $\underline{s} = t_i$ and $\bar{s} = t_{i+1}$ if $s \in (t_i, t_{i+1})$. One notes that $(\delta Y_s - \delta Y_{s^-})^2 = (\bar{Y}_s - \widetilde{Y}_s)^2$ so that, by the definition of the process $\bar{K}_s$, one has

$$\begin{aligned}
|\delta Y_t|^2 =& |\delta Y_\tau|^2 + 2\int_t^\tau \delta Y_s\Big(f\big(\Theta_s\big) - f\big(\bar{\Theta}_{\underline{s}}\big)\Big)ds + 2\int_{[t,\tau)} \delta Y_s dK_s - \int_{[t,\tau)} (Z_s - \bar{Z}_s)dW_s \\
&- \int_{[t,\tau)} |\delta Z^2_s| ds - \sum_{t\le t_i<\tau}\Big(2\delta Y_i(h_i(\bar{X}_i) - \widetilde{Y}_i)_+ + (\bar{Y}_i - \widetilde{Y}_i)^2\Big).
\end{aligned} \tag{6.75}$$

For every $t_i < \tau$, we set $\alpha_i = 2\delta Y_i\,(h_i(\bar{X}_i) - \widetilde{Y}_i)_+ + (\bar{Y}_i - \widetilde{Y}_i)^2$ for convenience. It can be written as follows:

$$\begin{aligned}
\alpha_i &= 2(Y_i - \bar{Y}_i)\big(h_i(\bar{X}_i) \vee \widetilde{Y}_i - \widetilde{Y}_i\big) + (\bar{Y}_i - \widetilde{Y}_i)^2 \\
&= 2(Y_i - \bar{Y}_i)(\bar{Y}_i - \widetilde{Y}_i) + (\bar{Y}_i - \widetilde{Y}_i)^2 \\
&= (Y_i - \widetilde{Y}_i)^2 - (Y_i - \bar{Y}_i)^2 \\
&= (Y_i - \widetilde{Y}_i)^2 - (\delta Y_i)^2.
\end{aligned}$$

where we used, in the third line, the equality $2(a-b)(b-c) + (b-c)^2 = (a-c)^2 - (a-b)^2$.

Let us evaluate this term α_i. For every $t_i < \tau \leq \tau^d$, we have, by (6.74), two choices: Either $h_i(\bar{X}_i) < \widetilde{Y}_i$ so that $\bar{Y}_i = \widetilde{Y}_i$ and hence, $\delta Y_i = Y_i - \widetilde{Y}_i$ and $(\delta Y_i)^2 = (Y_i - \widetilde{Y}_i)^2$, or, $\delta Y_i > 0$ so, since $\widetilde{Y}_t < \bar{Y}_t$ for every t, we have $\widetilde{Y}_i - Y_i < \bar{Y}_i - Y_i < 0$ and then, $(\delta Y_i)^2 < (Y_i - \widetilde{Y}_i)^2$.
Consequently, for every $t_i \in [t, \tau[$,

$$\alpha_i = (Y_i - \widetilde{Y}_i)^2 - (\delta Y_i)^2 \geq 0.$$

Moreover, for $s \in [t, \tau[$, $s < \tau^c$ so that, by (6.73), we have $\delta Y_s < 0$ dK_s-a.e. Hence,

$$\int_{[t,\tau)} \delta Y_s dK_s < 0.$$

This yields

$$|\delta Y_t|^2 \leq |\delta Y_\tau|^2 + 2\int_t^\tau \delta Y_s \big(f(\Theta_s) - f(\bar{\Theta}_{\underline{s}})\big)ds - \int_{[t,\tau)} (Z_s - \bar{Z}_s)dW_s - \int_{[t,\tau)} |\delta Z_s^2|ds.$$

Now, we evaluate $|\delta Y_\tau|^2$ depending on the value of τ.
• If $\tau = \tau^d$, then, by (6.74), $\delta Y_\tau < 0$ and $h_\tau(\bar{X}_\tau) > \widetilde{Y}_\tau$. This means $\bar{Y}_\tau = h_\tau(\bar{X}_\tau)$ and, since $Y_t \geq h_t(X_t)$ for every $t \in [0, T]$,

$$0 \leq |\delta Y_\tau| = \bar{Y}_\tau - Y_\tau = h_\tau(\bar{X}_\tau) - Y_\tau \leq h_\tau(\bar{X}_\tau) - h_\tau(X_\tau).$$

Hence, $|\delta Y_\tau|^2 \leq [h]^2_{\text{Lip}} |X_\tau - \bar{X}_\tau|^2$.
• If $\tau = \tau^c$, then, by (6.73), $\delta Y_\tau > 0$ and K_s changes its value so $Y_\tau = h_\tau(X_\tau)$. Consequently,

$$0 \leq \delta Y_\tau = Y_\tau - \bar{Y}_\tau = h_\tau(X_\tau) - \bar{Y}_\tau \leq h_\tau(X_\tau) - h_\tau(\bar{X}_\tau)$$

since $\bar{Y}_t \geq h_t(\bar{X}_t)$ for every $t \in [0, T]$. So, $|\delta Y_\tau|^2 \leq [h]^2_{\text{Lip}} |X_\tau - \bar{X}_\tau|^2$.
• If $\tau = T$, $\delta Y_T = g(X_T) - g(\bar{X}_T)$ so $|\delta Y_\tau|^2 \leq [g]^2_{\text{Lip}} |X_\tau - \bar{X}_\tau|^2$. Consequently, for all the possible values of τ, we have

$$|\delta Y_\tau|^2 \leq C_{h,g,b,T,\sigma} \Delta$$

where $C_{h,g,b,T,\sigma}$ is a constant related to the Euler discretization error and depending on h and g. Thus, taking the conditional expectation with respect to t leads to

$$\mathbb{E}_t \left(|\delta Y_t|^2 + \int_t^\tau |\delta Z_s|^2 ds \right) \leq C_{h,g,b,T,\sigma} \Delta + 2\mathbb{E}_t \int_t^\tau \delta Y_s \big(f_s(\Theta_s) - f_{\underline{s}}(\bar{\Theta}_{\underline{s}})\big) - \mathbb{E}_t \int_{[t,\tau)} (Z_s - \bar{Z}_s)dW_s. \quad (6.76)$$

It remains to study the term $2\mathbb{E}_t \int_t^\tau \delta Y_s \big(f_s(\Theta_s) - f_{\underline{s}}(\bar{\Theta}_{\underline{s}})\big)$. As f is Lipschitz continuous, we use Young's inequality $ab \leq \frac{a^2}{2\alpha} + \frac{\alpha b^2}{2}$ and the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ to write

$$\begin{aligned} 2\mathbb{E}_t \int_t^\tau \delta Y_s \big(f_s(\Theta_s) - f_{\underline{s}}(\bar{\Theta}_{\underline{s}})\big) \leq \frac{3[f]_{\text{Lip}}}{\alpha} \Bigg(& \int_t^\tau \mathbb{E}_t |X_s - \bar{X}_{\underline{s}}|^2 ds + \int_t^\tau \mathbb{E}_t |Y_s - \mathbb{E}_{\underline{s}} \bar{Y}_{\bar{s}}|^2 ds \\ & + \mathbb{E}_t \int_t^\tau |Z_s - \bar{\zeta}_{\underline{s}}|^2 ds \Bigg) + \alpha [f]_{\text{Lip}} \mathbb{E}_t \int_t^\tau |\delta Y_s|^2 ds. \end{aligned} \quad (6.77)$$

On the one hand,

$$\mathbb{E}_t |X_s - \bar{X}_{\underline{s}}|^2 \leq 2\mathbb{E}_t |X_s - X_{\underline{s}}|^2 + 2\mathbb{E}_t |X_{\underline{s}} - \bar{X}_{\underline{s}}|^2$$

where $\mathbb{E}_t |X_s - X_{\underline{s}}|^2$ is bounded as follows: from (6.3) taken between $\underline{s}$ and s, we have

$$\begin{aligned} \mathbb{E}_t |X_s - X_{\underline{s}}|^2 &\leq 2\mathbb{E}_t \int_{\underline{s}}^s b_u(X_u)^2 du + 2\mathbb{E}_t \int_{\underline{s}}^s \sigma_u(X_u)^2 du \\ &\leq 4L^2_{b,\sigma} \mathbb{E}_t \int_{\underline{s}}^s (1 + |X_u|)^2 du \\ &\leq 4L^2_{b,\sigma} \Delta \mathbb{E}_t \sup_{\underline{s} \leq u \leq s} (1 + |X_u|)^2. \end{aligned}$$

Hence, denoting $C_X = 4L^2_{b,\sigma}(\tau - t)$,

$$\int_t^\tau \mathbb{E}_t|X_s - \bar{X}_{\underline{s}}|^2 ds \leq C_X \Delta \mathbb{E}_t \sup_{\underline{s}\leq u\leq s}(1+|X_u|)^2 + 2\int_t^\tau \mathbb{E}_t|X_{\underline{s}} - \bar{X}_{\underline{s}}|^2 ds. \tag{6.78}$$

On the other hand,

$$\mathbb{E}_t|Y_s - \mathbb{E}_{\underline{s}}\bar{Y}_{\bar{s}}|^2 \leq 2\mathbb{E}_t|Y_s - \bar{Y}_s|^2 + 4\mathbb{E}_t|\bar{Y}_s - \widetilde{Y}_{\underline{s}}|^2 + 4\mathbb{E}_t\mathbb{E}_{\underline{s}}|\widetilde{Y}_{\underline{s}} - \bar{Y}_{\bar{s}}|^2. \tag{6.79}$$

For every v, v' such that $v < v'$ and $|v - v'| \leq \Delta$, (6.41) at v and v' yields

$$\widetilde{Y}_v - \bar{Y}_{v'} = (v'-v) f\big(\underline{v},\, \bar{X}_{\underline{v}},\, \mathbb{E}_{\underline{v}}\bar{Y}_{\bar{v}},\, \bar{\zeta}_{\underline{v}}\big) - \int_v^{v'} \bar{Z}_s dW_s + \bar{K}_{v'} - \bar{K}_v$$

so that taking the conditional expectations w.r.t. t yields

$$\begin{aligned}
\mathbb{E}_t|\widetilde{Y}_v - \bar{Y}_{v'}|^2 \leq& 2(v'-v)^2\mathbb{E}_t f(\bar{\theta}_{\underline{v}})^2 + 2\mathbb{E}_t\left(\int_v^{v'} \bar{Z}_s dWs\right)^2 + 2\mathbb{E}_t(\bar{K}_{v'} - \bar{K}_v)^2 \\
\leq& 2\Delta^2\mathbb{E}_t f(\bar{\theta}_{\underline{v}})^2 + 2\mathbb{E}_t\left(\int_v^{v'} \bar{Z}_s dWs\right)^2 + 2\mathbb{E}_t(\bar{K}_{v'} - \bar{K}_v)^2.
\end{aligned}$$

Since $\bar{K}_v \geq 0$ for every $v \in [0,T]$, we have $-\bar{K}_v < \bar{K}_v$ so that $\bar{K}_{v'} - \bar{K}_v < \bar{K}_{v'} + \bar{K}_v$. Then, owing the fact that $\bar{K}_v$ is non decreasing, $\bar{K}_{v'} - \bar{K}_v \geq 0$ so

$$(\bar{K}_{v'} - \bar{K}_v)^2 \leq (\bar{K}_{v'} - \bar{K}_v)(\bar{K}_{v'} + \bar{K}_v) = \bar{K}^2_{v'} - \bar{K}^2_v.$$

Hence, noting that $f(\bar{\Theta}_{\underline{s}}) = f(\bar{X}_{\underline{s}}, \mathbb{E}_{\underline{s}}\bar{Y}_{\bar{s}}, \bar{\zeta}_{\underline{s}})$ is a composition of the functions f, $\bar{y}_{\underline{s}}$ and $\bar{z}_{\underline{s}}$ which are all Lipschitz continuous according to Lemma 6.7.1 and recalling that if a function g is Lipschitz continuous then it has linear growth i.e. there exists a finite constant C_0 such that $g(x) \leq C(1+|x|)$, one has

$$\mathbb{E}_t|\widetilde{Y}_v - \bar{Y}_{v'}|^2 \leq 2\Delta^2 C_0 \mathbb{E}_t(1 + \sup_{v\leq s\leq v'}|\bar{X}_s|)^2 + 2\mathbb{E}_t\left(\int_v^{v'} \bar{Z}_s dWs\right)^2 + 2\mathbb{E}_t(\bar{K}^2_{v'} - \bar{K}^2_v).$$

Combining this with (6.79) twice yields

$$\begin{aligned}
\int_t^\tau \mathbb{E}_t|Y_s - \mathbb{E}_{\underline{s}}\bar{Y}_{\bar{s}}|^2 ds \leq& 2\int_t^\tau \mathbb{E}_t|\delta Y_s|^2 ds + 4\sum_{i=\underline{t}/\Delta}^{(\bar{\tau}/\Delta)-1}\int_{t_i}^{t_{i+1}} \mathbb{E}_t|\bar{Y}_s - \widetilde{Y}_{\underline{s}}|^2 ds + 4\sum_{i=\underline{t}/\Delta}^{(\bar{\tau}/\Delta)-1}\int_{t_i}^{t_{i+1}} \mathbb{E}_t|\bar{Y}_{\bar{s}} - \widetilde{Y}_{\underline{s}}|^2 ds \\
\leq& 2\int_t^\tau \mathbb{E}_t|\delta Y_s|^2 ds + 8\Delta^2 C_0(\tau - t)\mathbb{E}_t(1 + \sup_{\underline{s}\leq u\leq \bar{s}}|\bar{X}_u|)^2 \\
&+ 8\sum_{i=\underline{t}/\Delta}^{(\bar{\tau}/\Delta)-1}\int_{t_i}^{t_{i+1}} \mathbb{E}_t(\bar{K}^2_s - \bar{K}^2_{\underline{s}}) + \mathbb{E}_t(\bar{K}^2_{\bar{s}} - \bar{K}^2_{\underline{s}}) \\
&+ 8\sum_{i=\underline{t}/\Delta}^{(\bar{\tau}/\Delta)-1}\int_{t_i}^{t_{i+1}} \mathbb{E}_t\left(\int_{\underline{s}}^s \bar{Z}_u dWu\right)^2 ds + 8\sum_{i=\underline{t}/\Delta}^{(\bar{\tau}/\Delta)-1}\int_{t_i}^{t_{i+1}} \mathbb{E}_t\left(\int_{\underline{s}}^{\bar{s}} \bar{Z}_u dWu\right)^2 ds \\
\leq& 2\int_t^\tau \mathbb{E}_t|\delta Y_s|^2 ds + 8\Delta(\bar{\tau} - \underline{t})\mathbb{E}_t|\bar{K}_T|^2 + 8\Delta^2 C_0 C_f(\tau - t)\mathbb{E}_t(1 + \sup_{\underline{s}\leq u\leq \bar{s}}|\bar{X}_u|)^2 \\
&+ 8\sum_{i=\underline{t}/\Delta}^{(\bar{\tau}/\Delta)-1}\int_{t_i}^{t_{i+1}} \left(\mathbb{E}_t\Big(\int_{\underline{s}}^s \bar{Z}_u dWu\Big)^2 + \mathbb{E}_t\Big(\int_{\underline{s}}^{\bar{s}} \bar{Z}_u dWu\Big)^2\right) ds
\end{aligned} \tag{6.80}$$

where we used the fact that $\bar{K}_t$ is a non-decreasing positive process so for every $t \in [0,T]$, $\bar{K}_t < \bar{K}_T$ and the fact that $\sup_{\underline{s}\le u\le s}\alpha_u \le \sup_{\underline{s}\le u\le \bar{s}}\alpha_u$.

Thirdly,

$$\begin{aligned}\mathbb{E}_t\int_t^\tau |Z_s-\bar{\zeta}_{\underline{s}}|^2ds &\le \sum_{i=\underline{t}/\Delta}^{(\bar{\tau}/\Delta)-1}\mathbb{E}_t\int_{t_i}^{t_{i+1}}|Z_s-\bar{\zeta}_{\underline{s}}|^2ds\\ &\le \sum_{i=\underline{t}/\Delta}^{(\bar{\tau}/\Delta)-1}\mathbb{E}_t\left(4\int_{t_i}^{t_{i+1}}|Z_s-Z_{\underline{s}}|^2ds+4\int_{t_i}^{t_{i+1}}|Z_{\underline{s}}-\zeta_{\underline{s}}|^2ds+2\int_{t_i}^{t_{i+1}}|\zeta_{\underline{s}}-\bar{\zeta}_{\underline{s}}|^2ds\right).\end{aligned}$$

By the definitions (6.40) and (6.39) of ζ_s and $\bar{\zeta}_s$, we have

$$|Z_{\underline{s}}-\zeta_{\underline{s}}|^2=\Big|Z_{\underline{s}}-\frac{1}{\Delta}\mathbb{E}_{\underline{s}}\int_{\underline{s}}^{\bar{s}}Z_sds\Big|^2=\frac{1}{\Delta^2}\Big|\mathbb{E}_{\underline{s}}\int_{\underline{s}}^{\bar{s}}(Z_s-Z_{\underline{s}})\Big|^2\le\frac{1}{\Delta}\mathbb{E}_{\underline{s}}\int_{\underline{s}}^{\bar{s}}|Z_s-Z_{\underline{s}}|^2ds$$

where the last inequality was obtained by using Cauchy-Schwarz inequality. Hence, we use Fubini's Theorem to deduce

$$\int_{t_i}^{t_{i+1}}|Z_{\underline{s}}-\zeta_{\underline{s}}|^2ds\le\mathbb{E}_{\underline{s}}\int_{\underline{s}}^{\bar{s}}|Z_s-Z_{\underline{s}}|^2ds.$$

Likewise,

$$|\zeta_{\underline{s}}-\bar{\zeta}_{\underline{s}}|^2=\Big|\frac{1}{\Delta}\mathbb{E}_{\underline{s}}\int_{\underline{s}}^{\bar{s}}(Z_s-\bar{Z}_s)\Big|^2\le\frac{1}{\Delta}\mathbb{E}_{\underline{s}}\int_{\underline{s}}^{\bar{s}}|Z_s-\bar{Z}_s|^2ds$$

so that

$$\int_{t_i}^{t_{i+1}}|\zeta_{\underline{s}}-\bar{\zeta}_{\underline{s}}|^2ds\le\mathbb{E}_{\underline{s}}\int_{\underline{s}}^{\bar{s}}|Z_s-\bar{Z}_s|^2ds.$$

Consequently,

$$\mathbb{E}_t\int_t^\tau|Z_s-\bar{\zeta}_{\underline{s}}|^2ds\le 8\mathbb{E}_t\int_{\underline{t}}^{\bar{\tau}}|Z_s-Z_{\underline{s}}|^2ds+2\mathbb{E}_t\int_{\underline{t}}^{\bar{\tau}}|Z_s-\bar{Z}_s|^2ds. \tag{6.81}$$

At this stage, we merge the 3 equations (6.78), (6.80) and (6.81) with (6.77) and take the expectation to obtain

$$\begin{aligned}\mathbb{E}\Big(|\delta Y_t|^2+\int_t^\tau|\delta Z_t|^2\Big)\le&\Delta\Big(C_{h,g,b,T,\sigma}+\frac{6[f]}{\alpha}\big(C_X+C_0(\bar{\tau}-\underline{t})\big)\mathbb{E}(1+\sup_{\underline{s}\le u\le\bar{s}}|\bar{X}_u|)^2+\frac{12[f]}{\alpha}(\bar{\tau}-\underline{t})\mathbb{E}|\bar{K}_T|^2\Big)\\&+\left(2\alpha[f]_{\mathrm{Lip}}+\frac{6[f]_{\mathrm{Lip}}}{\alpha}\right)\int_t^\tau\mathbb{E}|\delta Y_s|^2ds+\frac{6[f]_{\mathrm{Lip}}}{\alpha}\int_t^\tau\mathbb{E}|X_{\underline{s}}-\bar{X}_{\underline{s}}|^2ds\\&+\frac{24[f]_{\mathrm{Lip}}}{\alpha}\mathbb{E}\int_{\underline{t}}^{\bar{\tau}}|Z_s-Z_{\underline{s}}|^2ds+\frac{6[f]_{\mathrm{Lip}}}{\alpha}\mathbb{E}\int_{\underline{t}}^{\bar{\tau}}|Z_s-\bar{Z}_s|^2ds\\&+\frac{24[f]_{\mathrm{Lip}}}{\alpha}\sum_{i=\underline{t}/\Delta}^{(\bar{\tau}/\Delta)-1}\int_{t_i}^{t_{i+1}}\left(\mathbb{E}\int_{\underline{s}}^{s}|\bar{Z}_u|^2du+\mathbb{E}\int_{\underline{s}}^{\bar{s}}|\bar{Z}_u|^2du\right)ds.\end{aligned}$$

As stated in (6.4), $\mathbb{E}|\bar{K}_T|^2\le\gamma_0$ and by the classical properties of the Euler scheme, we have

$$\mathbb{E}(1+\sup_u|\bar{X}_u|)^2\le C_{b,T,\sigma}(1+|x_0|)^2\qquad\text{and}\qquad\mathbb{E}|X_{\underline{s}}-\bar{X}_{\underline{s}}|^2\le C_{b,T,\sigma}\Delta(1+|x_0|)^2.$$

Moreover,

$$\sum_{i=\underline{t}/\Delta}^{(\bar{\tau}/\Delta)-1}\int_{t_i}^{t_{i+1}}\left(\mathbb{E}\int_{\underline{s}}^{s}|\bar{Z}_u|^2du+\mathbb{E}\int_{\underline{s}}^{\bar{s}}|\bar{Z}_u|^2du\right)ds\le\sum_{i=\underline{t}/\Delta}^{(\bar{\tau}/\Delta)-1}\int_{t_i}^{t_{i+1}}2\Delta\mathbb{E}\sup_{\underline{s}\le u\le\bar{s}}|\overline{Z}_u|^2\le 2\Delta^2(\bar{\tau}-\underline{t})\gamma_1.$$

Hence, if we consider $\alpha = 6[f]_{\mathrm{Lip}}$ and denote $\bar{C} = C_{h,g,b,T,\sigma} + C_X + (\bar{\tau} - \underline{t})\Big(2\gamma_0 + 4\gamma_1\Delta_{\max} + (1 + C_0)C_{b,T,\sigma}(1 + |x_0|)^2\Big)$ and $\widetilde{C} = 1 + 12[f]^2_{\mathrm{Lip}}$, then we obtain

$$\begin{aligned}\mathbb{E}\left(|\delta Y_t|^2 + \int_t^\tau |\delta Z_s|^2\right) \leq & \Delta\bar{C} + \widetilde{C}\int_t^\tau \mathbb{E}|\delta Y_s|^2 ds + 4\mathbb{E}\int_{\underline{t}}^{\bar{\tau}} |Z_s - Z_{\underline{s}}|^2 ds \\ & + \mathbb{E}\int_{\underline{t}}^t |Z_s - \bar{Z}_s|^2 ds + \mathbb{E}\int_t^\tau |Z_s - \bar{Z}_s|^2 ds + \mathbb{E}\int_\tau^{\bar{\tau}} |Z_s - \bar{Z}_s|^2 ds.\end{aligned}$$

Consequently,

$$\mathbb{E}|\delta Y_t|^2 \leq \widetilde{C}\int_t^\tau \mathbb{E}|\delta Y_s|^2 ds + K$$

where $K = \Delta\bar{C} + 4\mathbb{E}\int_{\underline{t}}^{\bar{\tau}} |Z_s - Z_{\underline{s}}|^2 ds + \mathbb{E}\int_{\underline{t}}^t |Z_s - \bar{Z}_s|^2 ds + \mathbb{E}\int_\tau^{\bar{\tau}} |Z_s - \bar{Z}_s|^2 ds$. Let us denote $f(t) = \mathbb{E}|\delta Y_t|^2$. This function satisfies

$$f(t) \leq \widetilde{C}\int_t^\tau f(s) ds + K.$$

We consider $g(t) = f(T - t)$ which satisfies also

$$g(t) \leq \widetilde{C}\int_0^t g(s) ds + K.$$

Hence, Gronwall's Lemma yields $g(t) \leq e^{\widetilde{C}t}K$ so that

$$f(t) \leq e^{\widetilde{C}(T-t)}K.$$

Consequently,

$$\mathbb{E}|Y_t - \bar{Y}_t|^2 \leq e^{\widetilde{C}(T-t)}\left(\Delta\bar{C} + 4\mathbb{E}\int_0^T |Z_s - Z_{\underline{s}}|^2 ds + \mathbb{E}\int_{\underline{t}}^t |Z_s - \bar{Z}_s|^2 ds + \mathbb{E}\int_\tau^{\bar{\tau}} |Z_s - \bar{Z}_s|^2 ds\right).$$

In particular, if $t = t_k$ and $\tau = t_{k'}$, $k, k' \in \{1, \ldots, n\}$, then $\underline{t} = t$ and $\bar{\tau} = \tau$ so

$$\mathbb{E}|Y_k - \bar{Y}_k|^2 \leq e^{\widetilde{C}(T-t_k)}\left(\Delta\bar{C} + 4\int_0^T \mathbb{E}|Z_s - Z_{\underline{s}}|^2 ds + 0 + 0\right).$$

This completes the proof. □

Chapter 7

Barrier options and details on recursive quantization

7.1 Introduction

In the first part of this chapter, we detail the numerical section of Chapter 6 and give further details on the numerical computation of the recursive quantization of a diffusion $(X_t)_{t\in[0,T]}$ given by

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \quad X_0 = x_0, \tag{7.1}$$

evolving according to certain models, mainly Black-Scholes model and the CEV model, and discretized following an Euler scheme with time step $\Delta = \frac{T}{n}$, $n \in \mathbb{N}$ as follows

$$\bar{X}_{t_{k+1}} = \bar{X}_{t_k} + b_t(\bar{X}_{t_k})\Delta + \sigma_t(\bar{X}_{t_k})\sqrt{\Delta}\varepsilon_{k+1} := \mathcal{E}_k(\bar{X}_{t_k}, \varepsilon_{k+1}) \tag{7.2}$$

where $(\varepsilon_k)_{1\leq k\leq n}$ is a sequence of i.i.d. random variables with distribution $\mathcal{N}(0, I_d)$. We give more numerical examples in order to illustrate the convergence of a recursive quantization-based discretization scheme of a reflected BSDE, mainly in the valuation of the price of American options.

In the second part, we attack another application of recursive quantization. We are interested in the pricing of a class of path-dependent payoffs, the Barrier options. Just like in the previous chapter where the goal was to approximate the solution of a reflected BSDE with a forward process given by (7.1), we start with a time discretization to obtain (7.2) and then, we apply a space discretization by recursive vector quantization to compute the corresponding recursive quantization $(\widehat{X}_{t_k})_{0\leq k\leq n}$. This technique was deeply explained in Section 6.2 of the previous chapter and upper error bounds were established in L^p, $p \in (1, 2 + d)$.

We will replace, most of the times, the indices t_k by k, for example, $\bar{X}_{t_k}$ will be denoted $\bar{X}_k$.

7.2 Numerical implementation of specific models

As explained in Section 6.5 of Chapter 6, the recursive quantization of $\bar{X}_k$ is given by $\widehat{X}_k = \sum_{i=1}^{N_k} x_i^k \mathbb{1}_{C_i(\Gamma_k)}$ where the grid Γ_k is the optimal quantizer of $\widetilde{X}_k$ of size N_k. Assuming that all the grids $\Gamma_k = \{x_1^k, \ldots, x_{N_k}^k\}$ are already computed up to time t_k, the grid $\Gamma_{k+1} = \{x_1^{k+1}, \ldots, x_{N_{k+1}}^{k+1}\}$ is computed via

$$x_j^{k+1} = \mathbb{E}\left(\widetilde{X}_{k+1} \,|\, \widehat{X}_{k+1} \in C_j(\Gamma_{k+1})\right) = \frac{\sum_{i=1}^{N_k} p_i^k \mathbb{E}\left(\mathcal{E}_k(x_i^k, \varepsilon_{k+1})\mathbb{1}_{\{\mathcal{E}_k(x_i^k,\varepsilon_{k+1})\in C_j(\Gamma_{k+1})\}}\right)}{p_j^{k+1}}. \tag{7.3}$$

where $\mathcal{E}_k(x,\varepsilon_{k+1}) = x + \Delta b_k(x) + \sqrt{\Delta}\sigma_k(x)\varepsilon_{k+1}$. For every $k \in \{1,\dots,n\}$ and $i,j \in \{1,\dots,N_k\}$, the transition probability p_{ij}^k from x_i^k to x_j^{k+1} is given by

$$p_{ij}^k = \mathbb{P}\left(\widehat{X}_{k+1} \in C_j(\Gamma_{k+1}) \,|\, \widehat{X}_k \in C_i(\Gamma_k)\right) = \mathbb{P}\left(\mathcal{E}_k(x_i^k,\varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\right). \tag{7.4}$$

And the weights p_j^{k+1} of the Voronoï cells $C_j(\Gamma_{k+1})$ are given, for every $j \in \{1,\dots,N_{k+1}\}$, via the classical (discrete time) forward Kolmogorov equation, as follows

$$p_j^{k+1} = \mathbb{P}\big(\widetilde{X}^{k+1} \in C_j(\Gamma_{k+1})\big) = \sum_{i=1}^{N_k} p_i^k \mathbb{P}\Big(\mathcal{E}_k(x_i^k,\varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\Big). \tag{7.5}$$

In this section, we give details on the computations of the recursive (and greedy recursive) quantizers of a process $(\bar{X}_k)_{0\le k\le n}$ following a Black-Scholes model and a CEV model, and discretized following an Euler scheme with time step $\Delta = \frac{T}{n}$, $n > 0$. The following closed formulas are available in the one-dimensional framework and are used in the pricing of American options and Barrier options in the end of this chapter.

7.2.1 Black-Scholes model

Consider a process $(X_t)_{0\le t\le T}$ evolving following a Black-Scholes model

$$dX_t = rX_t dt + \sigma X_t dW_t, \qquad X_0 = x_0 \in \mathbb{R},$$

where r is the interest rate, σ the volatility, T the maturity and $(W_t)_{0\le t\le T}$ a standard Brownian motion. It is discretized following the Euler scheme

$$\bar{X}_{k+1} = \bar{X}_k + r\Delta\bar{X}_k + \sigma\sqrt{\Delta}\bar{X}_k\,\varepsilon_{k+1} := \mathcal{E}_k(\bar{X}_k,\varepsilon_{k+1}) \tag{7.6}$$

where $(\varepsilon_k)_{1\le k\le n}$ is a sequence of i.i.d. random variables with Normal distribution $\mathcal{N}(0,1)$.

Relying on the fact that, for every $i \in \{1,\dots,N_k\}$, $\mathcal{E}_k(x_i^k,\varepsilon_{k+1}) \sim \mathcal{N}(m_i^k,\Sigma_i^k)$ where $m_i^k = x_i^k + \Delta b_k(x_i^k)$ and $\Sigma_i^k = \sqrt{\Delta}\sigma_k(x_i^k)$, the expectations and probabilities in (7.3), (7.4) and (7.5) are computed, for every $i \in \{1,\dots,N_k\}$ and $j \in \{1,\dots,N_{k+1}\}$, as follows

$$\begin{aligned}\mathbb{E}\Big(\mathcal{E}_k(x_i^k,\varepsilon_{k+1})\mathbb{1}_{\{\mathcal{E}_k(x_i^k,\varepsilon_{k+1})\in C_j(\Gamma^{k+1})\}}\Big) =& \frac{\sigma\sqrt{\Delta}x_i^k}{\sqrt{2\pi}}\Big(e^{-\frac{\left(x_{j,\inf}^{k+1}\right)^2}{2}} - e^{-\frac{\left(x_{j,\sup}^{k+1}\right)^2}{2}}\Big)\\ &+ x_i^k(1+\Delta r)\Big(\Phi_0\big(x_{j,\sup}^{k+1}\big) - \Phi_0\big(x_{j,\inf}^{k+1}\big)\Big)\end{aligned} \tag{7.7}$$

and

$$\mathbb{P}\left(\mathcal{E}_k(x_i^k,\varepsilon_{k+1}) \in C_j(\Gamma^{k+1})\right) = \Phi_0\left(x_{j,\sup}^{k+1}\right) - \Phi_0\left(x_{j,\inf}^{k+1}\right) \tag{7.8}$$

where Φ_0 is the c.d.f. of $\mathcal{N}(0,1)$,

$$x_{j,\inf}^{k+1} = \frac{x_{j-\frac{1}{2}}^{k+1} - m_i^k}{\Sigma_i^k} \quad \text{and} \quad x_{j,\sup}^{k+1} = \frac{x_{j+\frac{1}{2}}^{k+1} - m_i^k}{\Sigma_i^k}$$

and $x_{j+\frac{1}{2}}^{k+1} = \frac{x_j^{k+1}+x_{j+1}^{k+1}}{2}$ with the conventions $x_{\frac{1}{2}}^{k+1} = -\infty$ and $x_{N_{k+1}+\frac{1}{2}}^{k+1} = +\infty$.

Next, we deal with the computation of the quantizers of $(\bar{X}_k)_{0\le k\le n}$ by greedy recursive quantization. The steps to follow in this case are detailed in Section 6.6.1 of Chapter 6. Here, we will present the formulas needed to compute the local inter-point inertia. Denote $m_j^{k-1} = x_j^{k-1} + r\Delta x_j^{k-1}$ and $\Sigma_j^{k-1} = \sigma\sqrt{\Delta}x_j^{k-1}$, then, for every $i \in \{1,\dots,N\}$, these inertias are computed via

$$\sigma_i^2 = \sum_{j=1}^{N} p_j^{k-1} s_{ij} \tag{7.9}$$

where s_{ij} are given, for every $j \in \{1, \dots, N_k\}$, by
• If $i = 0$

$$s_{1j} = \Phi_0\left(\frac{x_1^k - m_j^{k-1}}{\Sigma_j^{k-1}}\right)\left((x_1^k - m_j^{k-1})^2 + (\Sigma_j^{k-1})^2\right) + \frac{\Sigma_j^{k-1}}{\sqrt{2\pi}}(x_1^k - m_j^{k-1})e^{-\frac{(x_i^k - m_j^{k-1})^2}{2(\Sigma_j^{k-1})^2}},$$

• If $i = N_k - 1$

$$s_{N_k-1\,j} = \left(1 - \Phi_0\left(\frac{x_{N_k}^k - m_j^{k-1}}{\Sigma_j^{k-1}}\right)\right)\left((x_{N_k}^k - m_j^{k-1})^2 + (\Sigma_j^{k-1})^2\right) - \frac{\sigma_j^{k-1}}{\sqrt{2\pi}}(x_{N_k}^k - m_j^{k-1})e^{-\frac{(x_{N_k}^k - m_j^{k-1})^2}{2(\Sigma_j^{k-1})^2}},$$

• If $0 < i < N_k - 1$

$$\begin{aligned} s_{ij} = {} & \frac{\Sigma_j^{k-1}}{\sqrt{2\pi}}(m_j^{k-1} - x_i^k)e^{-\frac{(x_i^k - m_j^{k-1})^2}{2(\Sigma_j^{k-1})^2}} + \frac{\Sigma_j^{k-1}}{\sqrt{2\pi}}(x_{i+1}^k - m_j^{k-1})e^{-\frac{(x_{i+1}^k - m_j^{k-1})^2}{2(\Sigma_j^{k-1})^2}} + \frac{\Sigma_j^{k-1}}{\sqrt{2\pi}}(x_i^k - x_{i+1}^k)e^{-\frac{(x_{i+\frac{1}{2}}^k - m_j^{k-1})^2}{2(\Sigma_j^{k-1})^2}} \\ & + \left((x_i^k - m_j^{k-1})^2 + (\Sigma_j^{k-1})^2\right)\left(\Phi_0\left(\frac{x_{i+\frac{1}{2}}^k - m_j^{k-1}}{\Sigma_j^{k-1}}\right) - \Phi_0\left(\frac{x_i^k - m_j^{k-1}}{\Sigma_j^{k-1}}\right)\right) \\ & + \left((x_{i+1}^k - m_j^{k-1})^2 + (\Sigma_j^{k-1})^2\right)\left(\Phi_0\left(\frac{x_{i+1}^k - m_j^{k-1}}{\Sigma_j^{k-1}}\right) - \Phi_0\left(\frac{x_{i+\frac{1}{2}}^k - m_j^{k-1}}{\Sigma_j^{k-1}}\right)\right). \end{aligned}$$

Then, for implementing Lloyd's algorithm, one uses the formulas (7.7) and (7.8) to compute the quantization sequence as well as the companion parameters (transition weights and Voronoï weights).

For an example, we consider

$$T = 1, \quad X_0 = 100, \quad r = 0.006, \quad \sigma = 0.2$$

discretize in $n = 30$ time steps and build quantizers of size $N_k = 50$ for every $k \in \{1, \dots, n\}$. In Figure 7.1, we observe the functions $x_i^k \mapsto p_i^k$, for every $k \in \{1, \dots, n\}$ where $(x_i^k)_{1 \le i \le N_k}$ is the recursive quantization grid and $(p_i^k)_{1 \le i \le N_k}$ are the corresponding Voronoï weights.

7.2.2 CEV model

Consider that the process $(X_t)_{0 \le t \le T}$ evolves following a CEV model, a local volatility model, according to

$$dX_t = rXtdt + \vartheta X_t^\delta dW_t, \qquad X_0 = x_0, \tag{7.10}$$

for $\delta \in (0, 1)$ and $\vartheta \in (0, \overline{\vartheta}]$, $\overline{\vartheta} > 0$, where r represents the interest rate and $\sigma(x) = \vartheta x^\delta$ represents the local volatility function. The corresponding Euler scheme with timestep $\Delta = \frac{T}{n}$, $n > 0$, is given by

$$\bar{X}_{k+1} = \bar{X}_k + r\Delta\bar{X}_k + \vartheta\bar{X}_k^\delta\sqrt{\Delta}\varepsilon_k \tag{7.11}$$

where $(\varepsilon_k)_k$ is a sequence of i.i.d. random variables with distribution $\mathcal{N}(0, 1)$.

The recursive and greedy recursive quantization of this process is identical to the Black-Scholes model framework. The only difference is replacing the constant volatility σ with $\sigma(x) = \vartheta x^\delta$, especially in the expressions of m_i^k and Σ_i^k.

In Figure 7.2, we represent the functions $x_i^k \mapsto p_i^k$ for every $k \in \{1, \dots, n\}$ where $(x_i^k)_{1 \le i \le N_k}$ is the recursive quantization grid and $(p_i^k)_{1 \le k \le N_k}$ the corresponding Voronoï weights. From a practical point

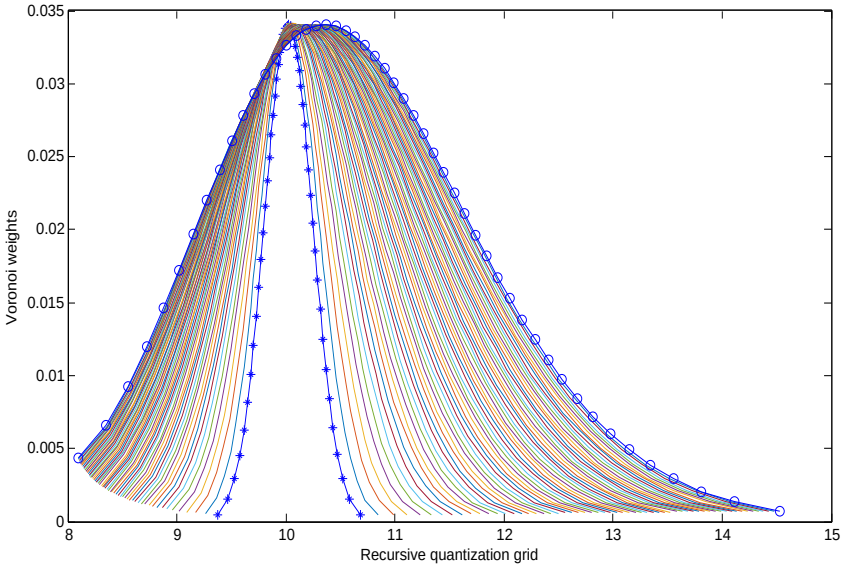

Figure 7.1: Representation of $x_i^k \mapsto p_i^k$ where $(x_i^k)_{1\le i\le N_k}$ is the recursive quantization grid, for every $k \in \{1,\dots,n\}$, in a Black-Scholes model (* corresponds to $k=2$ and $\circ$ corresponds to $k=n=30$).

of view, we build quantizers of size $N_k = 50$ for every $k \in \{1,\dots,n\}$, discretize in $n=20$ time steps and consider

$$T=1, \quad X0=100, \quad r=0.15, \quad \delta=0.5, \quad \vartheta=4.$$

7.3 Optimal quantization of a Brownian motion

In Section 6.6.1 of Chapter 6, we gave exact formulas for the computation of the transition matrices of an optimal quantization tree corresponding to $(\widehat{X}_k)_{0\le k\le n}$ following a Black-Scholes model. Here, our aim is to establish similar formulas to compute the transition weights of an optimal quantization tree corresponding to a standard Brownian motion, which is a more general case.

Let $(W_{t_k})_{0\le k\le n}$ be a standard Brownian motion (sampled at time t_k) and $(\widehat{W}_{t_k})_{0\le k\le n}$ its optimal marginal quantization sequence in the sense that each $\widehat{W}_{t_k}$ is an optimal quadratic quantization of W_{t_k}. Assume that the size of the grids N_k, $k=1,\dots,n$, are all equal to $N \in \mathbb{N}$. Note that this hypothesis is not optimal but turns out to be optimal in terms of complexity for a given budget $N_1+\cdots+N_n$. It is not sharp in terms of error estimates (up to a multiplicative constant) but remains a good compromise which is convenient in practice for the implementation.

We start by noticing that, at time t_k, one has $W_{t_k} = W_{t_k} - W_{t_0} = \sqrt{t_k}\varepsilon_k$ where $(\varepsilon_k)_{0\le k\le n}$ is a sequence of i.i.d. random variables with distribution $\mathcal{N}(0,1)$. Hence, the optimal quadratic quantizer $\Gamma_k = (x_1^k,\dots,x_N^k)$ of W_{t_k} is obtained by simple dilatation from an optimal quantizer $(z_1^k,\dots,z_N^k)$ of the standard Normal distribution. In other words, one has, for every $i \in \{1,\dots,N\}$,

$$x_i^k = \sqrt{t_k} z_i^k.$$

Note that highly accurate quantization grids of $\mathcal{N}(0,1)$ for regularly sampled sizes from $N=1$ to $1\,000$

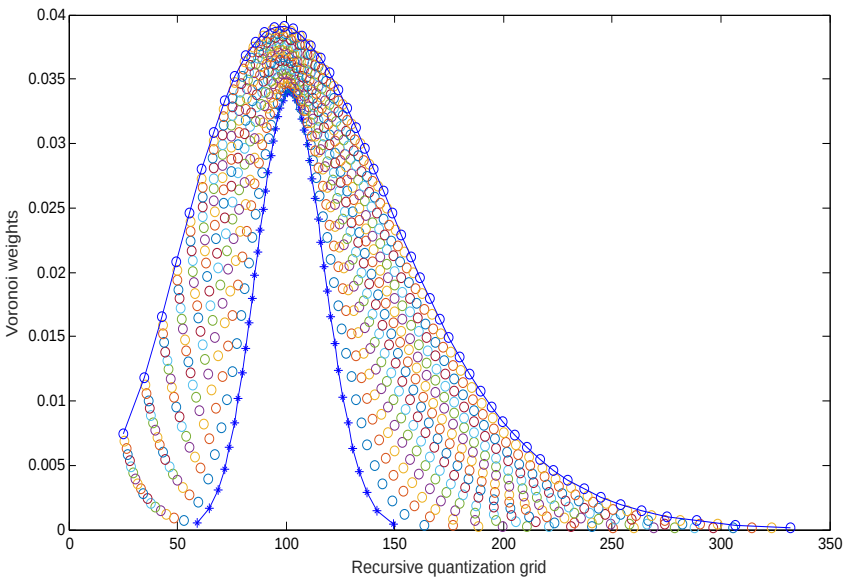

Figure 7.2: Representation of $x_i^k \mapsto p_i^k$ where $(x_i^k)_{1\le i\le N_k}$ is the recursive quantization grid, for every $k \in \{1,\dots,n\}$, in a CEV model (* corresponds to $k=2$ and $\circ$ corresponds to $k=n=20$).

are available and can be downloaded from the quantization website www.quantize.maths-fi.com (for non-commercial purposes).

Exact computation of the transition weights The goal is to compute the transition weights

$$p_{ij}^k = \mathbb{P}\Big(\widehat{W}_{t_{k+1}} = x_j^{k+1} \,|\, \widehat{W}_{t_k} = x_i^k\Big) = \frac{\bar{p}_{ij}^k}{p_i^k}$$

where

$$\bar{p}_{ij}^k = \mathbb{P}\Big(\widehat{W}_{t_{k+1}} = x_j^{k+1},\, \widehat{W}_{t_k} = x_i^k\Big) \qquad \text{and} \qquad p_i^k = \mathbb{P}\big(\widehat{W}_{t_k} = x_i^k\big).$$

The weights p_i^k are equal to the weights of the Voronoï cells induced by the quantizer $(z_1^k,\dots,z_N^k)$ of the standard Normal distribution and are available with the pre-computed sequences on the quantization website. However, they can always be computed via the forward Kolmogorov equation, using the transition weights p_{ij}^k, as follows

$$p_j^{k+1} = \sum_{i=1}^{N_k} p_{ij}^k\, p_i^k = \sum_{i=1}^{N_k} \bar{p}_{ij}^k,$$

keeping in mind that the Voronoï weight at time t_0 (i.e. $k=0$) is equal to 1 since $\widehat{X}_0 = X_0 = x_0$ is deterministic. So, our main concern is the computation of $\bar{p}_{ij}^k$ for every $k \in \{1,\dots,n\}$ and $i,j \in \{1,\dots,N\}$. We rely on the fact that $W_{t_k} = \sqrt{t_k}\varepsilon_1$ and $W_{t_{k+1}} = W_{t_k} + (W_{t_{k+1}} - W_{t_k}) = \sqrt{t_k}\varepsilon_1 + \sqrt{\frac{T}{n}}\varepsilon_2$ where ε_1 and ε_2 are two independent random variables with distribution $\mathcal{N}(0,1)$. Hence, denoting $\Delta = \frac{T}{n}$, one obtains

$$\begin{aligned}
\bar{p}_{ij}^k &= \mathbb{P}\left(\sqrt{t_k}\varepsilon_1 + \sqrt{\Delta}\varepsilon_2 \in C_j(\Gamma_{k+1}),\, \sqrt{t_k}\varepsilon_1 \in C_i(\Gamma_k)\right) \\
&= \mathbb{P}\left(\sqrt{t_k}\varepsilon_1 + \sqrt{\Delta}\varepsilon_2 \in \left[\sqrt{t_{k+1}}z_{j-\frac12}^{k+1},\, \sqrt{t_{k+1}}z_{j+\frac12}^{k+1}\right],\, \sqrt{t_k}\varepsilon_1 \in \left[\sqrt{t_k}z_{i-\frac12}^{k},\, \sqrt{t_k}z_{i+\frac12}^{k}\right]\right) \\
&= \mathbb{P}\left(\varepsilon_2 \in \left[\sqrt{k+1}z_{j-\frac12}^{k+1} - \sqrt{k}\varepsilon_1,\, \sqrt{k+1}z_{j+\frac12}^{k+1} - \sqrt{k}\varepsilon_1\right],\, \varepsilon_1 \in \left[z_{i-\frac12}^{k}, z_{i+\frac12}^{k}\right]\right)
\end{aligned}$$

where we used in the last inequality the fact that $\sqrt{\frac{t_{k+1}}{\Delta}} = \sqrt{k+1}$ and $\sqrt{\frac{t_k}{\Delta}} = \sqrt{k}$. Then, the independence of ε_1 and ε_2 yields

$$\begin{aligned}\bar{p}_{ij}^k &= \int_{z_{i-\frac{1}{2}}^k}^{z_{i+\frac{1}{2}}^k} \mathbb{P}\left(\varepsilon_2 \in \left[\sqrt{k+1}z_{j-\frac{1}{2}}^{k+1} - \sqrt{k}z,\ \sqrt{k+1}z_{j+\frac{1}{2}}^{k+1} - \sqrt{k}z\right]\right) e^{-\frac{z^2}{2}} \frac{dz}{\sqrt{2\pi}} \\ &= \int_{z_{i-\frac{1}{2}}^k}^{z_{i+\frac{1}{2}}^k} \left(\Phi_0\left(\sqrt{k+1}z_{j+\frac{1}{2}}^{k+1} - \sqrt{k}z\right) - \Phi_0\left(\sqrt{k+1}z_{j-\frac{1}{2}}^{k+1} - \sqrt{k}z\right)\right) e^{-\frac{z^2}{2}} \frac{dz}{\sqrt{2\pi}}. \end{aligned} \tag{7.12}$$

These integrals can be computed via Gaussian quadrature formulas as explained in Chapter 6, mainly Gauss-Legendre quadrature formulas for integrals on closed intervals and Gauss-Laguerre quadrature formulas for integrals on semi-closed intervals.

Approximation of the transition weights If the goal is not necessarily the highest level of precision, then one approximates the transition weights p_{ij}^k by $g_j(z_i^k)$ where the function g_j is given by

$$g_j(z) = \Phi_0\left(\sqrt{k}z - \sqrt{k+1}z_{j+\frac{1}{2}}^{k+1}\right) - \Phi_0\left(\sqrt{k}z - \sqrt{k+1}z_{j-\frac{1}{2}}^{k+1}\right).$$

The reasoning is similar to the case of the optimal quantization tree associated to a diffusion evolving following a Black-Scholes model in Section 6.6.1.

7.4 Further numerical examples

We present further numerical examples illustrating the theoretical results obtained in the previous chapter on the recursive quantization-based discretization scheme of a reflected BSDE.

7.4.1 American put options under the historical probability

We are interested in the computation of an American put option price with maturity T and strike price K. This (risk-neutral) price is given by the initial value Y_0 of the following RBSDE under the historical (real world) probability $\mathbb{P}$

$$\begin{gathered} -dY_t = \Big(- rY_t - \frac{b_t(X_t) - r}{\sigma_t(X_t)} Z_t \Big) dt - Z_t dW_t + dK_t \\ Y_T = h(X_T) \qquad \text{and} \qquad Y_t \geq g(X_t) \end{gathered}$$

where $g(x) = h(x) = \max(K - x, 0)$ and $b_t(X_t)$ and $\sigma_t(X_t)$ are the coefficients of the diffusion $(X_t)_{t\in[0,T]}$ representing the stock price.

Black-Scholes model We consider that the forward process $(X_t)_{t\in[0,T]}$ evolves following the Black-Scholes dynamics. The corresponding Euler scheme is given by (7.6). We compute the quantizers of $\bar{X}_k$ for every $k \in \{0, \dots, n\}$ by recursive quantization (RQ), optimal quantization (OQ), greedy quantization (GQ) and greedy recursive quantization (GRQ) as explained in Chapter 6 and in the sections above. Then, we compute the price of the underlying option Y_0 via the backward recursion (6.58). We consider $n = 15$ time steps and a size $N = 150$ of the quantizers. The parameters of the model are the following

$$X_0 = 40\,, \quad T = 0.5833\,, \quad \sigma = 0.3\,, \quad \mu = 0\,, \quad r = 0.0488\,.$$

We compute the desired values by the different types of quantization for $K \in \{35; 40; 45\}$ and compare the results with the prices obtained in [36]. The results and the errors induced by this comparison are displayed in Table 7.1. Figure 7.3 depicts the convergence of the error induced by the approximation of Y_0 based on a recursive quantization of $\bar{X}_k$. We fix the strike $K = 40$ and vary the size N of the grids between 10 and 100. The graph is represented in a log-log-scale and an $\mathcal{O}(N^{-1})$ rate of convergence is clearly observed.

K	**RQ**		**GRQ**		**OQ**		**GQ**		**Benchmark**
	Value	Error	Value	Error	Value	Error	Value	Error	
35	1.228	0.082	1.23	0.0102	1.23	0.0102	1.217	0.0028	1.2198
40	3.17	0.0004	3.166	0.0036	3.165	0.0046	3.157	0.0126	3.1696
45	6.232	0.0116	6.225	0.0186	6.223	0.0206	6.232	0.0116	6.2436
Average		0.0067		0.0108		0.0118		0.009	

Table 7.1: Pricing of an American put option under the historical probability in a Black-Scholes model discretized according to an Euler scheme and recursive (RQ), greedy recursive (GRQ), optimal (OQ) and greedy (GQ) quantization for different values of K.

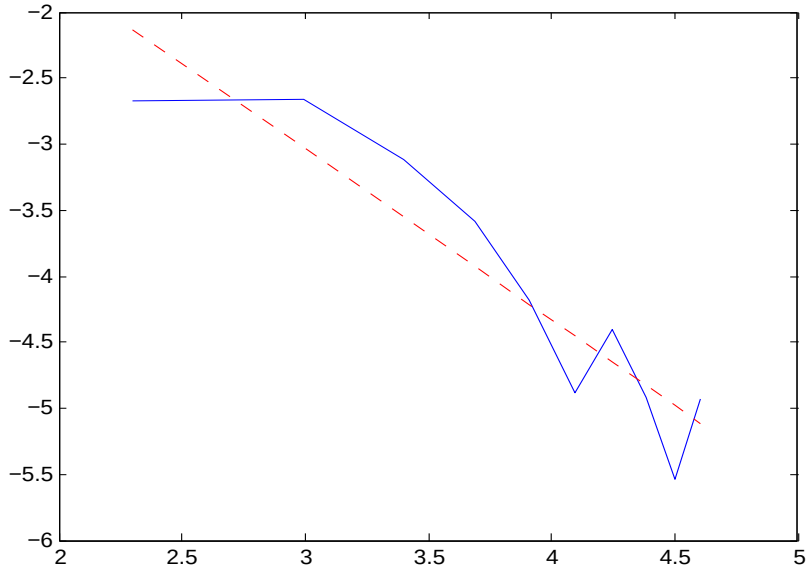

Figure 7.3: Convergence rate of the quantization error for the American put under historical probability in a Black-Scholes model for different sizes $N = 10, \dots, 100$.

CEV model Now consider that $(X_t)_{t\in[0,T]}$ evolves following the CEV model, i.e.

$$dX_t = \mu X_t dt + \vartheta X_t^{\delta} dW_t, \qquad X_0 = x_0, \tag{7.13}$$

for some $\delta \in (0,1)$ and $\vartheta \in (0, \overline{\vartheta}]$ with $\overline{\vartheta} > 0$. $\sigma(x) = x^{\delta}$ is the local volatility function. The discretized Euler scheme associated to $(X_t)_{t\in[0,T]}$ is given, for every $k \in \{0, \dots, n-1\}$, by

$$\bar{X}_{k+1} = \bar{X}_k + \mu \Delta \bar{X}_k + \vartheta \bar{X}_k^{\delta} \sqrt{\Delta}\, \varepsilon_k \tag{7.14}$$

where $(\varepsilon_k)_{1\le k\le n}$ is an i.i.d sequence of random variables with distribution $\mathcal{N}(0,1)$. We compute the quantizers of $\bar{X}_k$ for every $k \in \{0, \dots, n\}$ by recursive quantization (RQ), optimal quantization (OQ), greedy quantization (GQ) and greedy recursive quantization (GRQ) as explained in Chapter 6 and in the sections above. For this model, the transition weights of the optimal and greedy quantization tree are computed via Monte Carlo simulations and not with closed formulas. We discretize with $n = 15$ time steps, build quantizers of size $N = 150$ and consider the following parameters

$$X_0 = 40\,, \quad T = 0.5833\,, \quad \vartheta = 2\,, \quad \delta = 0.5\,, \quad \varepsilon = 1\,, \quad \mu = 0\,, \quad r = 0.0488\,.$$

We compute the price Y_0 via (6.58) for the different types of quantization for different values of K between 30 and 50. The benchmark in this case is the price obtained by an optimal quantization-based Richardson

Romberg extrapolation, as explained in section 6.6.2 of Chapter 6, with $n = 5$ and $N = 1\,000$. The results and the errors hence induced are exposed in Table 7.2.

K	**RQ**		**GRQ**		**OQ**		**GQ**		**Romberg**
	Value	Error	Value	Error	Value	Error	Value	Error	
30	0.593	0.025	0.597	0.021	0.605	0.013	0.594	0.024	0.618
35	1.707	0.023	1.709	0.021	1.726	0.004	1.705	0.025	1.73
40	3.76	0.017	3.758	0.019	3.779	0.002	3.756	0.021	3.777
45	6.81	0.02	6.806	0.024	6.835	0.005	6.807	0.023	6.83
50	10.698	0.042	10.692	0.048	10.722	0.018	10.694	0.046	10.74
Average		0.0254		0.0266		0.0084		0.0278	

Table 7.2: Pricing of an American put option under the historical probability in a CEV model discretized according to an Euler scheme and recursive (RQ), greedy recursive (GRQ), optimal (OQ) and greedy (GQ) quantization for different values of K.

7.4.2 American put options

The price of an American put option is given by the RBSDE (6.1) with a driver f equal to 0 and $h(x) = g(x) = \max(K - x, 0)$. This means that the time discretized backward recursion is given by

$$\bar{Y}_T = h_T(\bar{X}_T) \qquad \text{and} \qquad \bar{Y}_{t_k} = \max\left(h_k(\bar{X}_{t_k}), \mathbb{E}_k\, \bar{Y}_{t_{k+1}}\right)$$

and the space discretization backward recursion by

$$\widehat{Y}_T = h_T(\widehat{X}_T) \qquad \text{and} \qquad \widehat{Y}_{t_k} = \max\left(h_k(\widehat{X}_{t_k}), \mathbb{E}_k\, \widehat{Y}_{t_{k+1}}\right)$$

where $(\widehat{X}_k)_{0\le k\le n}$ is the recursive quantization sequence associated to $(\bar{X}_k)_{0\le k\le n}$ so that $\bar{Y}_k$ and $\widehat{Y}_k$ are both $\mathcal{F}_{t_k}$-measurable processes for every $k \in \{1, \dots, n\}$. The solutions of these recursions can be written respectively as the Snell envelopes

$$\bar{Y}_k = \mathbb{P}\text{-esssup}\big\{\mathbb{E}\left(h_\tau(\bar{X}_\tau)\,|\,\mathcal{F}_\tau\right),\ \tau \in \{t_k, \dots, T\}\ \mathcal{F}_\tau\text{-stopping time}\big\}$$

and

$$\widehat{Y}_k = \mathbb{P}\text{-esssup}\big\{\mathbb{E}\left(h_\tau(\widehat{X}_\tau)\,|\,\mathcal{F}_\tau\right),\ \tau \in \{t_k, \dots, T\}\ \mathcal{F}_\tau\text{-stopping time}\big\}.$$

This allows to estimate an upper bound for the L^p-space discretization error $\|\bar{Y}_k - \widehat{Y}_k\|_p$ as follows

$$\|\bar{Y}_k - \widehat{Y}_k\|_p \le [h]_{\text{Lip}} \big\| \max_{l\ge k} |\bar{X}_l - \widehat{X}_l| \big\|_p \le [h]_{\text{Lip}} \left(\sum_{l=k}^{n} \|\bar{X}_l - \widehat{X}_l\|_p^p\right)^{\frac{1}{p}}$$

where $\|\bar{X}_l - \widehat{X}_l\|_p$ is the L^p-recursive quantization error estimated in Section 6.2. In fact, by the definitions of $(\bar{Y}_k)_{0\le k\le n}$ and $(\widehat{Y}_k)_{0\le k\le n}$, one has, for every $k \in \{0, \dots, n\}$

$$\begin{aligned} |\bar{Y}_k - \widehat{Y}_k| &\le \mathbb{P} - \text{esssup}\Big\{\mathbb{E}_k\, |h_\tau(\bar{X}_\tau) - h_\tau(\widehat{X}_\tau|,\ \tau \in \{k, \dots, n\}\ \mathcal{F}_{t_k}\text{-stopping time}\Big\} \\ &\le [h]_{\text{Lip}} \mathbb{E}_k \Big(\max_{l\ge k} |\bar{X}_l - \widehat{X}_l| \Big). \end{aligned}$$

From a numerical point of view, we proceed like for the previous examples and build quantizers via recursive, greedy recursive and optimal quantization. Particularly in this example, one does not need to compute the parameters π_{ij}^k since the driver f is equal to 0 and hence they are not needed to implement the backward recursion (6.58). The quantizers, the weights of the Voronoï cells and the transition weights matrices are computed similarly to the previous example.

Black-Scholes model - discretization according to an Euler scheme In this paragraph, $(X_t)_{t\in[0,T]}$ evolves following a Black-Scholes model. The corresponding Euler scheme is given by (7.6). We consider $n = 15$ time steps, build quantizers of size $N = 100$ and consider

$$K = 110\,, \quad T = 1\,, \quad \sigma = 0.2\,, \quad \mu = 0.006\,.$$

We compare the prices obtained by the different quantization techniques to a benchmark obtained by a binomial tree with $\bar{n} = 10^4$ time steps. The principle is the following: Starting with X_0 at time t_0, one computes, at time $t_k = k\Delta = k\frac{T}{\bar{n}}$, the process $X_{k+1} = (x_j^{k+1})_{1\le j\le k+1}$ from $X_k = (x_i^k)_{1\le i\le n}$

$$x_j^{k+1} = \begin{cases} ux_i^k & \text{if the price increases} \\ dx_i^k & \text{if the price decreases} \end{cases}$$

where $u = e^{(r-\frac{\sigma^2}{2})\Delta+\sigma\sqrt{\Delta}}$ and $d = e^{(r-\frac{\sigma^2}{2})\Delta-\sigma\sqrt{\Delta}}$. Then, once we have computed all the values for every $k \in \{1,\dots,\bar{n}\}$, we proceed with the valuation of the price of the American put option via a backward recursion as follows:

$$\begin{cases} v_i^{\overline{n}} & = h_{\overline{n}}(x_i^{\overline{n}}), \quad i = 1,\dots,N_{\overline{n}}, \\ v_i^k & = \max\left(h_k(x_i^k),\, \mathbb{E}(v^{k+1}|v_i^k)\right), \quad i = 1,\dots,N_k,\ k = 0,\dots,\overline{n}-1. \end{cases} \tag{7.15}$$

Note that the transition probability from x_i^k to x_j^{k+1} is equal to $p = 0,5$ and to x_{j+1}^{k+1} is equal to $1-p$. The results and the errors hence obtained are exposed in Table 7.3 for different values of X_0 between 90 and 120.

X_0	**RQ**		**GRQ**		**OQ**		**Binomial tree**
	Value	error	Value	Error	Value	Error	
90	21.275	0.02	21.256	0.001	21.233	0.022	21.255
95	17.378	0.018	17.346	0.014	17.341	0.019	17.36
100	13.936	0.022	13.886	0.028	13.902	0.012	13.914
105	10.971	0.025	10.915	0.031	10.945	0.001	10.946
110	8.483	0.028	8.425	0.03	8.467	0.012	8.455
120	4.819	0.028	4.771	0.032	4.825	0.034	4.791
Average error		0.024		0.021		0.017	

Table 7.3: Pricing of an American put option in a Black-Scholes model discretized according to an Euler scheme and recursive (RQ), greedy recursive (GRQ) and optimal (OQ) quantization for different values of X_0.

Furthermore, we are interested, in this example, in comparing the results obtained when the transition weight matrices of the optimal quantization tree are computed exactly via Gaussian quadrature formulas versus when they are approximated by the function g_j (see Section 6.6.1 of Chapter 6). We observe the errors induced in Table 7.4 and deduce that using closed formulas or approximating formulas to compute the transition weights give almost the same results, the differences between the two approaches are negligible. However, it should be mentioned that the approximating computation is much faster so the choice depends on the interest of the implementation and on the values of X_0. For comparison purposes, we compute the difference between the exact transition weights and the approximated transition weights and observe an error equal to 0.0614 which is more or less significant and mostly due to the differences in the transition weights of the Voronoï cells at the edges. This interpretation is deduced by the numerical experiments but it is also intuitive since these cells are of the form $[a,+\infty)$ or $(-\infty,a]$ and their centroids cannot represent the whole unbounded cell.

X_0	**Approximated weights**	**Exact weights**
90	0.02	0.022
95	0.019	0.019
100	0.014	0.012
105	0.004	0.001
110	0.008	0.012
120	0.029	0.034
Average error	0.016	0.017

Table 7.4: Errors induced by the pricing of an American put option in a Black-Scholes model discretized according to an Euler scheme and optimal quantization with transition weights computed exactly and approximately for different values of X_0.

CEV model - discretization according to a Milstein scheme We consider that $(X_t)_{t\in[0,T]}$ evolves following a CEV model and the time discretization is established according to a Milstein scheme with time step $\Delta = \frac{T}{n}$, i.e.

$$\bar{X}_{k+1} = m_k\varepsilon_{k+1} + c_k := \mathcal{U}(\bar{X}_k, \varepsilon_{k+1}) \tag{7.16}$$

where

$$c_k = \bar{X}_k\Big(1 - \frac{1}{2\delta} + r\Delta - \frac{1}{2}h\vartheta^2\delta\bar{X}_k^{2(\delta-1)}\Big) \quad , \quad m_k = \frac{1}{2}\Delta\vartheta^2\delta\bar{X}_k^{2\delta-1}$$

and ε_{k+1} is a random variable with distribution $\chi^2(1,\mu_k^2)$ with 1 degree of freedom and

$$\mu_k = \frac{\bar{X}_k^{1-\delta}}{\vartheta\delta\sqrt{\Delta}}.$$

Let us first give some details about the computation of the quantizers of $(\bar{X}_k)_{1\le k\le n}$ in the Milstein scheme case. For the optimal quantization, the quantizers are obtained as explained previously for the Euler scheme and the transition weights are computed by Monte Carlo simulations. However, for the recursive and greedy recursive quantization, one should note some differences.

We start with the computation of the inter-point inertia in the case of greedy recursive quantization before proceeding with the expectations and probabilities common to both methods. According to (7.9), these inter-point inertia are given, for every $i,j\in\{1,\dots,N\}$, by $\sigma_j^2 = \sum_{i=1}^N p_i^{k-1}s_{ij}$ where s_{ij} is computed based on the following formulas:

- The cumulative distribution function of $\chi^2(1,\mu^2)$: $F_\varepsilon(x) = \Phi_0(x^+) - \Phi_0(x^-)$,
- The first order moment of $\chi^2(1,\mu^2)$

$$M_\varepsilon^1(x) = (1+\mu^2)\left(\Phi_0(x^+) - \Phi_0(x^-)\right) + (2\mu + x^-)\frac{e^{-\frac{x^{-2}}{2}}}{\sqrt{2\pi}} - (2\mu + x^+)\frac{e^{-\frac{x^{+2}}{2}}}{\sqrt{2\pi}},$$

- The second order moment of $\chi^2(1,\mu^2)$

$$\begin{aligned} M_\varepsilon^2(x) = \left(\Phi_0(x^+) - \Phi_0(x^-)\right)\left(\frac{\mu^4}{2} + 3\mu^2 + 3\right) + \frac{e^{-\frac{x^{-2}}{2}}}{\sqrt{2\pi}}\left(4\mu^3 + 6\mu^2x^- + 4\mu(2+x^{-^2}) + x^-(3+x^{-^2})\right) \\ - \frac{e^{-\frac{x^{+2}}{2}}}{\sqrt{2\pi}}\left(4\mu^3 + 6\mu^2x^+ + 4\mu(2+x^{+^2}) + x^+(3+x^{+^2})\right). \end{aligned}$$

where Φ_0 is the c.d.f of the standard Normal distribution, $x^+ = \sqrt{x} - \mu$ and $x^- = -\sqrt{x} - \mu$.

Then, in order to compute the quantizer $\Gamma_{k+1} = \{x_1^{k+1}, \dots, x_N^{k+1}\}$ and its companion parameters for every $k \in \{0, \dots, n\}$, we start by denoting

$$x_{j,i+} = \frac{x_{j+\frac{1}{2}}^{k+1} - c_k^i}{m_k^i} \qquad \text{and} \qquad x_{j,i-} = \frac{x_{j-\frac{1}{2}}^{k+1} - c_k^i}{m_k^i}$$

where $x_{j+\frac{1}{2}}^{k+1} = \frac{x_j^{k+1} + x_{j+1}^{k+1}}{2}$, $c_k^i = x_i^k\Big(1 - \frac{1}{2\delta} + r\Delta - \frac{1}{2}\Delta\vartheta^2\delta(x_i^k)^{2(\delta-1)}\Big)$ and $m_k^i = \frac{1}{2}\Delta\vartheta^2\delta(x_i^k)^{2\delta-1}$. Then, the expectations and probabilities in (7.3), (7.4) and (7.5) are computed as follows

$$\mathbb{P}\big(\mathcal{U}_k(x_i^k, \varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\big) = F_\varepsilon(x_{j,i+}) - F_\varepsilon(x_{j,i-})$$

and

$$\begin{aligned}\mathbb{E}\left(\mathcal{U}_k(x_i^k, \varepsilon_{k+1})\mathbb{1}_{\{\mathcal{E}_k(x_i^k, \varepsilon_{k+1}) \in C_j(\Gamma_{k+1})\}}\right) &= \big(c_i^k + m_i^k(1 + {\mu_i^k}^2)\big)\Big(F_\varepsilon(x_{j,i+}) - F_\varepsilon(x_{j,i-})\Big) \\ &+ \frac{m_i^k}{\sqrt{2\pi}}\Bigg[(2\mu_i^k + x_{j,i+}^-)e^{-\frac{{x_{j,i+}^-}^2}{2}} - (2\mu_i^k + x_{j,i+}^+)e^{-\frac{{x_{j,i+}^+}^2}{2}} \\ &+ (2\mu_i^k + x_{j,i-}^+)e^{-\frac{{x_{j,i-}^+}^2}{2}} - (2\mu_i^k + x_{j,i-}^-)e^{-\frac{{x_{j,i-}^-}^2}{2}}\Bigg].\end{aligned}$$

From a practical point of view, we consider $n = 25$, build quantizers of size $N = 100$ and consider the same parameters as previously. We compare the results obtained for different values of K between 90 and 120 to the price obtained by a Richardson-Romberg extrapolation with $n = 5$ and $N = 800$. The induced errors are reported in Table 7.5.

K	**RQ**		**GRQ**		**OQ**		**Romberg**
	Value	Error	Value	Error	Value	Error	
90	6.556	0.042	6.545	0.023	6.564	0.004	6.568
95	8.254	0.034	8.246	0.042	8.266	0.022	8.288
100	10.251	0.035	10.234	0.052	10.267	0.019	10.286
105	12.569	0.05	12.546	0.073	12.579	0.04	12.619
110	15.225	0.039	15.202	0.086	15.23	0.05	15.288
120	21.674	0.053	21.652	0.072	21.68	0.044	21.724
Average error		0.037		0.058		0.031	

Table 7.5: Pricing of an American put option in a CEV model discretized according to a Milstein scheme and recursive (RQ), greedy recursive (GRQ) and optimal (OQ) quantization compared to a Romberg extrapolation method for different values of K.

7.4.3 Two-dimensional American put options

Our goal is to approximate the price of a multi-dimensional American geometric put option given by the solution of the RBSDE (6.1) with a driver equal to 0 and $g_t(x^1, \dots, x^d) = h_t(x^1, \dots, x^d) = \max\big(K - |x^1 \cdots x^d|^{\frac{1}{d}}, 0\big)$, K being the strike price. Due to the choice of h and g, this d-dimensional problem can be reduced to a one-dimensional problem case. The forward process evolves following a Black-Scholes dynamics and is discretized by a Euler scheme, i.e. for $i \in \{1, \dots, d\}$

$$\bar{X}_{k+1}^i = \bar{X}_k^i + r\Delta\bar{X}_k^i + \sigma\sqrt{\Delta}\varepsilon_k^i$$

where r is the interest rate, σ the volatility and $(\varepsilon_k^i)_{k,i}$ a sequence of Normally distributed random variables. We carry out simulations for $d = 2$, discretize in $n = 10$ time steps and build quantizers of

size $N_X = 100$ by optimal and hybrid recursive quantization. For the hybrid recursive quantization, we use optimal quantizers of the standard Normal distribution of size $N_\varepsilon = 1450$ and the computations are similar to those in Example 6.6.2 in Chapter 6. The parameters of the example are

$$X_0^i = 100 \; \forall i \, , \quad T = 1 \, , \quad r = 0.05 \, , \quad \sigma = 0.4 \; , \quad \rho = 0 \, .$$

In Table 7.6, we vary K between 100 and 120 and expose the results obtained by the two methods and compare them to the benchmark obtained by a two-dimensional binomial tree which consists on the following: One starts by partitioning the interval $[0, T]$ into n sub-intervals $[t_k, t_{k+1}]$ where $t_k = k\Delta = k\frac{T}{n}$ for $k \in \{1, \ldots, n\}$. At each time t_k, the tree has k^2 nodes representing each the price of the forward process $X_k = (X_k^1, X_k^2)$ and, at time t_{k+1}, this price becomes equal to one of the 4 following possibilities

$$\begin{array}{ll} (X_k^1 u, X_k^2 u) \text{ with probability } p_{uu}, & (X_k^1 d, X_k^2 u) \text{ with probability } p_{du}, \\ (X_k^1 u, X_k^2 d) \text{ with probability } p_{ud}, & (X_k^1 d, X_k^2 d) \text{ with probability } p_{dd}, \end{array}$$

where $u = e^{\sigma\sqrt{\Delta}}$ is the factor corresponding to the price raise and $d = e^{-\sigma\sqrt{\Delta}}$ to the price drop. The probabilities are given by

$$p_{uu} = \frac{1}{4}\left(1 + \rho + \sqrt{\Delta}\left(\frac{2r - \sigma^2}{\sigma}\right)\right) \quad , \quad p_{dd} = \frac{1}{4}\left(1 + \rho - \sqrt{\Delta}\left(\frac{2r - \sigma^2}{\sigma}\right)\right)$$

$$\text{and} \qquad p_{ud} = p_{du} = \frac{1}{4}(1 - \rho)$$

where ρ is the correlation coefficient between the two variables. Hence, we start by a forward simulation to design the tree and then proceed with a backward simulation to compute the price of the American put via

$$P_{k,i} = \max\left(h_k(X_{k,i}^1, X_{k,i}^2), \mathbb{E}(P_{k+1}|P_{k,i})\right) .$$

K	**OQ**		**HRQ**		**Benchmark**
	Value	Error	Value	Error	
100	10.316	0.14	10.481	0.025	10.456
105	13.094	0.186	13.412	0.132	13.28
110	16.303	0.126	16.616	0.187	16.429
115	19.763	0.114	20.154	0.277	19.877
120	23.532	0.066	23.92	0.322	23.598
Average error		0.1264		0.1886	

Table 7.6: Pricing of a two-dimensional American put option in a BS model discretized according to an Euler scheme and hybrid recursive (HRQ) and optimal (OQ) quantization for different values of K.

7.4.4 Multi-dimensional example

We compute the solution of the following RBSDE *(example due to J.F. Chassagneux)*

$$-dY_t = (Z_t^1 + \ldots + Z_t^d)\Big(Y_t - \frac{2+d}{2d}\Big)dt - Z_t dW_t + dK_t$$

$$Y_T = \frac{e_T}{1 + e_T} \qquad \text{and} \qquad Y_t \geq \varphi(t)\frac{e_t}{1 + e_t}$$

where $e_t = e^{t+W_t^1+\ldots+W_t^d}$ and $\varphi(t) = -\frac{1}{2}\left(\frac{t-T}{T}\right)^2 + 1$ for a dimension $d \in \{2; 3\}$. The forward process $(X_t)_{t\in[0,T]}$ in this example is given by

$$dX_t = dW_t$$

where $(W_t)_{t\in[0,T]}$ is a d-dimensional standard Brownian motion. The corresponding discretized forward process is hence given by

$$\bar{X}_{k+1}^i = \bar{X}_k^i + \sqrt{\Delta}\varepsilon_{k+1}^i$$

where $(\varepsilon_k^1, \ldots, \varepsilon_k^d)_{1\leq k\leq n}$ is a sequence of i.i.d random variables with distribution $\mathcal{N}(0, I_q)$ and $\Delta = \frac{T}{n}$ is the time step parameter and n is the number of time steps.

We consider $T = 0.5$ and $X_0 = 0.5$ so that $Y_0 = 0.5$. We discretize in $n = 10$ time steps and build quantizers by optimal (OQ) and hybrid recursive quantization (HRQ). In this example, we compare the different values obtained for various sizes N_X of the optimal quantizers and the hybrid recursive quantizers and different sizes N_ε of the optimal quantizer of $\mathcal{N}(0, I_d)$ used in the hybrid recursive quantization of $\bar{X}_k$. We expose the results in Table 7.7 in the two-dimensional case and in Table 7.8 for the three-dimensional case.

N_X	N_ε	**HRQ**		**OQ**	
		Value	Error	Value	Error
50	750	0.5869	0.0869	0.5225	0.0225
50	1450	0.5864	0.0864	0.5225	0.0225
80	750	0.5856	0.0856	0.5163	0.0163
80	1450	0.58555	0.08555	0.5163	0.0163
100	900	0.58552	0.08552	0.514	0.014
100	1450	0.5854	0.0854	0.514	0.014
Average error			0.0858		0.0176

Table 7.7: Values of Y_0 in the two-dimensional framework based on optimal (OQ) and hybrid recursive quantization (HRQ) for different value of N_X and N_ε.

N_X	N_ε	**HRQ**		**OQ**	
		Value	Error	Value	Error
50	900	0.5959	0.0959	0.5566	0.0566
50	1300	0.5957	0.0957	0.5566	0.0566
80	900	0.5912	0.0912	0.5459	0.0459
80	2000	0.5909	0.0909	0.5459	0.0459
Average error			0.0934		0.0512

Table 7.8: Values of Y_0 in the three-dimensional framework based on optimal (OQ) and hybrid recursive quantization (HRQ) for different value of N_X and N_ε.

7.5 Application to Barrier options

The goal of this section is the pricing of a class of path-dependent payoffs, i.e. options whose payoff at maturity T depending, not only on the value of the underlying asset, but also on the maximum or minimum of its price on the interval $[0, T]$. In other words, the payoff can be written as

$$h_T = \Psi(X_T, \sup_{t\in[0,T]} X_t) \qquad \text{or} \qquad h_T = \Psi(X_T, \inf_{t\in[0,T]} X_t).$$

In particular, we consider Barrier options whose payoff is given by

$$h_T = \varphi(X_T)\mathbb{1}_{\sup_{t\in[0,T]} X_t \in I} \qquad \text{or} \qquad h_T = \varphi(X_T)\mathbb{1}_{\inf_{t\in[0,T]} X_t \in I}$$

where I is a subset of $\mathbb{R}^d$. Among others, one can name the following examples of Barrier options:
$\triangleright$ An Up-and-Out Call option which becomes equal to 0 as soon as the price of the underlying asset becomes higher than a certain barrier L, and whose payoff is given by

$$h_T(X_T) = (X_T - K)_+ \mathbb{1}_{\sup_{t\in[0,T]} X_t \leq L}$$

where T is the maturity, K the strike price and L the barrier.
$\triangleright$ A Down-and-Out Call option which becomes equal to 0 as soon as the price of the underlying asset becomes lower than a certain barrier L, and whose payoff is given by

$$h_T(X_T) = (X_T - K)_+ \, \mathbb{1}_{\inf_{t\in[0,T]} X_t \geq L}.$$

7.5.1 Theoretical approach

There exist formulas aiming to approach the prices of the options in this framework. To do so, it is necessary to handle the distribution of the maximum or the minimum of the Euler scheme associated to the price of the assets between two discretization steps t_k and t_{k+1}, conditioned w.r.t. its values at times t_k, $k \in \{0, \dots, n\}$. This means that one needs to study the diffusion bridges between t_k and t_{k+1}. Let us give some details and refer to [31, 57, 72] for example, for further details.

Let $W = (W_t)_{t\geq 0}$ be a standard Brownian motion. We start by presenting some basic properties of a Brownian diffusion bridge over $[0, T]$. It is a centered Gaussian process, measurable w.r.t. the filtration $(\mathcal{F}_t)_{t\in[0,T]} = \big(\sigma(W_s,\, s \leq t, \mathcal{N}_{\mathbb{P}})\big)_{t\in[0,T]}$, where $\mathcal{N}_{\mathbb{P}}$ is the class of all $\mathbb{P}$-negligible sets of $\mathcal{A}$, and defined by

$$Y_t^{W,T} = W_t - \frac{t}{T} W_T, \qquad t \in [0, T].$$

It is independent from $(W_{T+s})_{s\geq 0}$ and its covariance matrix is given by

$$\mathbb{E}(Y_s^{W,T} Y_t^{W,T}) = s \wedge t - \frac{st}{T} = \frac{(s \wedge t)(T - s \vee t)}{T}.$$

Furthermore, if we consider $0 < T_0 < T_1$, then

$$\mathcal{L}\big((W_t)_{t\in[T_0,T_1]} \,|\, W_s, s \notin (T_0, T_1)\big) = \mathcal{L}\big((W_t)_{t\in[T_0,T_1]} | W_{T_0}, W_{T_1}\big).$$

Hence, $(W_t)_{t\in[T_0,T_1]}$ and $(W_s)_{s\notin(T_0,T_1)}$ are independent conditioned to (W_{T_0}, W_{T_1}) and

$$\mathcal{L}\big((W_t)_{t\in[T_0,T_1]} \,|\, W_{T_0} = x, W_{T_1} = y\big) = \mathcal{L}\left(x + \frac{t - T_0}{T_1 - T_0}(y - x) + (Y_{t-T_0}^{W,T_1-T_0})_{t\in[T_0,T_1]}\right).$$

Going back to our problem, the Brownian diffusion bridge associated to the genuine Euler scheme of the diffusion (7.1) is characterized in the following proposition.

Proposition 7.5.1. (*i*) *The processes* $(X_t)_{t\in[t_k,t_{k+1}]}$, $k \in \{0, \dots, n-1\}$, *are independent, conditioned to* $\sigma(X_{t_k}, k = 0, \dots, n-1)$.
(*ii*)

$$\begin{aligned}\mathcal{L}\big((X_t)_{t\in[t_k,t_{k+1}]} \,|\, X_{t_l} = x_l, l = 0, \dots, n\big) &= \mathcal{L}\big((X_t)_{t\in[t_k,t_{k+1}]} \,|\, X_{t_k} = x_k, X_{t_{k+1}} = x_{k+1}\big) \\ &= \mathcal{L}\left(\left(x_k + \frac{n(t - t_k)}{T}(x_{k+1} - x_k) + \sigma(t_k, x_k) Y_{t-t_k}^{W,\Delta}\right)_{t\in[t_k,t_{k+1}]}\right).\end{aligned}$$

Having in hand the conditional distribution of the Euler scheme between t_k and t_{k+1} w.r.t its values at times t_k, $k \in \{0, \dots, n-1\}$, one can deduce the conditional distribution of its maximum or minimum over $[0, T]$.

Proposition 7.5.2. *Let $(u_k)_{k=0,\dots,n-1}$ be a sequence of i.i.d. random variables with Uniform distribution. Then,*

$$\mathcal{L}\Big(\max_{t\in[0,T]}(X_t) \,|\, X_{t_k} = x_k, k = 0, \dots, n\Big) = \mathcal{L}\Big(\max_{k=0,\dots,n-1} G_{u_k}^{-1}(x_k, x_{k+1})\mathbb{1}^{-1}_{u_k \geq \max(x_k, x_{k+1})}\Big)$$

and

$$\mathcal{L}\Big(\min_{t\in[0,T]}(X_t) \,|\, X_{t_k} = x_k, k = 0, \dots, n\Big) = \mathcal{L}\Big(\min_{k=0,\dots,n-1} F_{u_k}^{-1}(x_k, x_{k+1})\mathbb{1}^{-1}_{u_k \leq \min(x_k, x_{k+1})}\Big)$$

where

$$G_u(x,y) = \left(1 - e^{-2n\frac{(x-u)(y-u)}{T\sigma^2(x)}}\right) \qquad \text{and} \qquad F_u(x,y) = e^{-2n\frac{(x-u)(y-u)}{T\sigma^2(x)}}. \tag{7.17}$$

At this stage, we are able to give general formulas to approximate the price of Barrier options, in other words, to compute $\mathbb{E}\,\Psi(X_T, \max_{t\in[0,T]} X_t)$ or $\mathbb{E}\,\Psi(X_T, \min_{t\in[0,T]} X_t)$.

Proposition 7.5.3. *The price of an Up-and-Out option with maturity T and barrier L and whose payoff is given by the bounded function g is*

$$V_{UO} = e^{-rT}\mathbb{E}\left[g(X_n)\mathbb{1}_{\sup_{t\in[0,T]} X_t \leq L}\right] = e^{-rT}\mathbb{E}\left[g(X_n)\prod_{k=1}^{n} G_L(X_{k-1}, X_k)\mathbb{1}_{X_k, X_{k-1} \leq L}\right].$$

The price of a Down-and-Out option with maturity T and barrier L and whose payoff is given by the bounded function g is

$$V_{DO} = e^{-rT}\mathbb{E}\left[g(X_n)\mathbb{1}_{\inf_{t\in[0,T]} X_t \geq L}\right] = e^{-rT}\mathbb{E}\left[g(X_n)\prod_{k=1}^{n}\left(1 - F_L(X_{k-1}, X_k)\mathbb{1}_{X_k, X_{k-1} \geq L}\right)\right]$$

where G_L and F_L are the functions defined by (7.17)

To approximate the price of these options, time and space discretization schemes of the diffusion process $(X_t)_{t\in[0,T]}$ are mandatory. For the time discretization, we consider the Euler scheme $(\bar{X}_{t_k})_{0\leq k\leq n}$, with uniform mesh $t_k = k\Delta$ for $k \in \{0, \dots, n\}$ and $\Delta = \frac{T}{n}$, associated to the process $(X_t)_{t\in[0,T]}$ which is recursively given by

$$\bar{X}_{t_{k+1}} = \bar{X}_{t_k} + \Delta b_k(\bar{X}_{t_k}) + \sigma_k(\bar{X}_{t_k})(W_{t_{k+1}} - W_{t_k}), \qquad \bar{X}_0 = X_0 = x_0, \tag{7.18}$$

where $W_{t_{k+1}} - W_{t_k} = \sqrt{\Delta}\varepsilon_{k+1}$, for every $k \in \{0, \dots, n-1\}$ and $(\varepsilon_k)_{0\leq k\leq n}$ is a sequence of i.i.d. random variables with distribution $\mathcal{N}(0, I_q)$. Its continuous counterpart, the *genuine Euler scheme*, is given by

$$d\bar{X}_t = b(\underline{t}, \bar{X}_{\underline{t}})dt + \sigma(\underline{t}, \bar{X}_{\underline{t}})dW_t \tag{7.19}$$

where $\underline{t} = t_k$ when $t \in [t_k, t_{k+1})$. This process satisfies for every $p \in (0, +\infty)$ and every $n \geq 1$, (see [10])

$$\Big\|\sup_{t\in[0,T]} X_t\Big\|_p + \sup_{n\geq 1}\Big\|\sup_{t\in[0,T]} \bar{X}_t\Big\|_p \leq C_{b,T,\sigma}(1 + |x_0|) \qquad \text{and} \qquad \Big\|\sup_{t\in[0,T]} |X_t - \bar{X}_t|\Big\|_p \leq C_{b,T,\sigma}\sqrt{\Delta}(1 + |x_0|)$$

where $C_{b,T,\sigma}$ is a positive constant depending on p, T, b and σ.

Concerning the space discretization, we rely on vector quantization, more precisely, recursive vector quantization which was introduced in [63] and revisited and developed in Chapter 6. A(n optimized) recursive quantization of $(\bar{X}_k)_{0\le k\le n}$ is defined by the following recursion: $\widehat{X}_0 = \bar{X}_0 = x_0$ and

$$\begin{cases} \widetilde{X}_k &= \mathcal{E}_{k-1}(\widehat{X}_{k-1}^{\Gamma_{k-1}}, \varepsilon_k), \\ \widehat{X}_k^{\Gamma_k} &= \mathrm{Proj}_{\Gamma_k}(\widetilde{X}_k), \qquad \forall k = 1, \dots, n. \end{cases} \tag{7.20}$$

For the high-dimensional framework, the computations by this scheme become very complex so multi-dimensional extensions are necessary. One can cite the recursive product quantization in [28], a massive "embedded" Monte Carlo simulation or a more interesting alternative, introduced in Chapter 6, which is a kind of *hybrid* recursive quantization where the white noise $(\varepsilon_k)_{0\le k\le n}$ is replaced by its (already computed) quantized version $(\widehat{\varepsilon}_k)_{0\le k\le n}$. In other words, we consider, instead of (7.20), the following recursive scheme

$$\begin{cases} \widetilde{X}_k &= \mathcal{E}_{k-1}(\widehat{X}_{k-1}, \widehat{\varepsilon}_k), \\ \widehat{X}_k &= \mathrm{Proj}_{\Gamma_k}(\widetilde{X}_k), \qquad \forall k = 1, \dots, n. \end{cases} \tag{7.21}$$

where $(\widehat{\varepsilon}_k)_k$ is now a sequence of optimal quantizers of the Normal distribution $\mathcal{N}(0, I_q)$, which are already computed and kept off line, they can be found and downloaded from the quantization website www.quantize.maths-fi.com (for non-commercial purposes). A priori error bounds of these two types of quantization have been established in $L^p, p \in (1, 2+d)$, when assuming that $\widehat{X}_k$ is a stationary quadratic optimal quantizer of $\widetilde{X}_k$ for every $k \in \{1, \dots, n\}$.

In the following, we detail the approximation of Up-and-Out options, the study for other types of Barrier options is identical. The recursive quantization scheme allowing the computation of the price of the Barrier Up-and-Out option V_{UO} is given by the following Backward Dynamic Programming principle (BDPP) based on the recursive quantization $(\widehat{X}_k)_{0\le k\le n}$ of $(\bar{X}_k)_{0\le k\le n}$

$$\widehat{L}_n = g(\widehat{X}_n)\mathbb{1}_{\widehat{X}_n \ge L} \qquad \text{and} \qquad \widehat{L}_k = \mathbb{E}\left(G_L(\widehat{X}_k, \widehat{X}_{k+1})\widehat{L}_{k+1}\mathbb{1}_{\widehat{X}_k \ge L}|\mathcal{F}_k\right). \tag{7.22}$$

One can define the same BDPP for the non-quantized process:

$$L_n = g(\bar{X}_n)\mathbb{1}_{\bar{X}_n \ge L} \qquad \text{and} \qquad L_k = \mathbb{E}\left(G_L(\bar{X}_k, \bar{X}_{k+1})L_{k+1}\mathbb{1}_{\bar{X}_k \ge L}|\mathcal{F}_k\right). \tag{7.23}$$

It is clear that $L_0 = V_{UO}$.

Our aim is to establish upper bounds for the error induced by the approximation of L_k by $\widehat{L}_k$. We assume that, for every $k \in \{1, \dots, n\}$, the recursive quantization $\widehat{X}_k$ of $\bar{X}_k$ is computed according to (7.20) or (7.21) where $\widehat{X}_k$ is a quadratic optimal quantization of $\widetilde{X}_k$. It has been shown, in [57] for example, that a quadratic optimal quantizer $\widehat{X}^{\Gamma^N}$ of X is always a stationary quantizer in the following sense

$$\mathbb{E}(X|\widehat{X}^{\Gamma^N}) = \widehat{X}^{\Gamma^N}. \tag{7.24}$$

Before estimating the error bounds, we show that the functions $G_L(x, y)$ and $F_L(x, y)$ are locally Lipschitz continuous.

Lemme 7.5.4. *Let $L \in \mathbb{R}^d$, $d, p \in (0, +\infty)$ and $r \in \left(1, 1 + \frac{d}{p}\right)$. Assume that σ is uniformly elliptic, i.e. there exists $\sigma_0 > 0$ such that $\sigma(x) > \sigma_0$ for every $x \in \mathbb{R}^d$. Then, for every $x, y \in \mathbb{R}^d$, $G_L(x, y)$ and $F_L(x, y)$ are L^p-locally Lipschitz, i.e.*

$$\|G_L(x, y) - G_L(x', y')\|_p \le K_{\mathrm{Lip}}(\|x - x'\|_{rp} + \|y - y'\|_{rp})(1 + \|x\|^2_{2p\frac{r}{r-1}} + \|y'\|^2_{2p\frac{r}{r-1}})$$

and

$$\|F_L(x, y) - F_L(x', y')\|_p \le K_{\mathrm{Lip}}(\|x - x'\|_{rp} + \|y - y'\|_{rp})(1 + \|x\|^2_{2p\frac{r}{r-1}} + \|y'\|^2_{2p\frac{r}{r-1}}),$$

where $K_{\mathrm{Lip}} = \dfrac{n}{T}\max\left(\dfrac{1}{\sigma_0^2}, \dfrac{2[\sigma]_{\mathrm{Lip}}}{\sigma_0^3}\right)$.

Proof. The proof is identical for G_L and F_L. For every $x, y, x', y' \in \mathbb{R}^d$, the fact that $|e^u - e^v| \leq |u - v|$ for $u, v < 0$ yields

$$\begin{aligned}
|G_L(x,y) - G_L(x',y')| &= \left| e^{-2n\frac{(x-L)(y-L)}{\sigma^2(x)}} - e^{-2n\frac{(x'-L)(y'-L)}{\sigma^2(x')}} \right| \\
&\leq \frac{2n}{T} \left| \frac{(x-L)(y-L)}{\sigma^2(x)} - \frac{(x'-L)(y'-L)}{\sigma^2(x')} \right| \\
&\leq \frac{2n}{T} \left| \frac{xy}{\sigma^2(x+L)} - \frac{x'y'}{\sigma^2(x'+L)} \right| \qquad \text{(simple change of variables)} \\
&\leq \frac{2n}{T} \left| \frac{|x||y-y'|}{\sigma^2(x+L)} + |y'| \left(\frac{x}{\sigma^2(x+L)} - \frac{x'}{\sigma^2(x'+L)} \right) \right| \\
&\leq \frac{2n}{T} \left| \frac{|x||y-y'|}{\sigma^2(x+L)} + |y'| \left(x \left(\frac{1}{\sigma^2(x+L)} - \frac{1}{\sigma^2(x'+L)} \right) + \frac{|x-x'|}{\sigma^2(x'+L)} \right) \right| \\
&\leq \frac{2n}{T} \left| \frac{|x||y-y'|}{\sigma^2(x+L)} + |y'||x| \left| \frac{1}{\sigma^2(x+L)} - \frac{1}{\sigma^2(x'+L)} \right| + |y'| \frac{|x-x'|}{\sigma^2(x'+L)} \right|.
\end{aligned}$$

The assumption made on σ yields that $\max\left(\frac{1}{\sigma^2(x+L)}, \frac{1}{\sigma^2(x'+L)} \right) \leq \frac{1}{\sigma_0^2}$ so that

$$\frac{1}{\sigma^2(x+L)} - \frac{1}{\sigma^2(x'+L)} \leq 2[\sigma]_{\mathrm{Lip}} |x - x'| \frac{\sigma(x+L) + \sigma(x'+L)}{\sigma^4(x+L) + \sigma^4(x'+L)} \leq \frac{2[\sigma]_{\mathrm{Lip}}}{\sigma_0^3} |x - x'|.$$

Hence

$$\begin{aligned}
|G_L(x,y) - G_L(x',y')| &\leq \frac{2n}{T} \left| \frac{|x||y-y'|}{\sigma_0^2} + \frac{2[\sigma]_{\mathrm{Lip}}}{\sigma_0^3} |x-x'||y'||x| + |y'| \frac{|x-x'|}{\sigma_0^2} \right| \\
&\leq \frac{2n}{T} \max\left(\frac{1}{\sigma_0^2}, \frac{2[\sigma]_{\mathrm{Lip}}}{\sigma_0^3} \right) (|x-x'| + |y-y'|) (|x| + |y'| + |x||y'|) \\
&\leq \frac{n}{T} \max\left(\frac{1}{\sigma_0^2}, \frac{2[\sigma]_{\mathrm{Lip}}}{\sigma_0^3} \right) (|x-x'| + |y-y'|) \left(1 + |x|^2 + |y'|^2\right).
\end{aligned}$$

By taking the expectation, applying Hölder inequality with the conjugate exponents r an $\frac{r}{r-1}$ and then Minkowski's inequality, one obtains the result. □

Theorem 7.5.5. *Let $(\bar{X}_k)_{0 \leq k \leq n}$ be the process defined by (7.18) and let $(\widehat{X}_k)_{0 \leq k \leq n}$ be its recursive quantization. Assume that $\widehat{\widetilde{X}}_k$ is a stationary quadratic optimal quantization of $\widetilde{X}_k$ (in the sense of (7.24)), for every $k \in 0, \ldots, n$. Then, for $p \in (1, +\infty)$,*

$$\|L_k - \widehat{L}_k\|_p \leq \max\left(\kappa, C_0 K_{\mathrm{Lip}}\right) \|g_n\|_{\sup} \sum_{l=k}^{n} \|\bar{X}_l - \widehat{X}_l\|_{p+\frac{d}{2}} + \|\bar{X}_l - \widehat{X}_l\|_p^{\frac{1}{pp'}}$$

where C_0 and κ are finite positive constants depending on p, p' is a finite number larger than 1 and $K_{\mathrm{Lip}} = \frac{n}{T} \max\left(\frac{1}{\sigma_0^2}, \frac{2[\sigma]_{\mathrm{Lip}}}{\sigma_0^3} \right)$.

Proof. We use the previous Lemma to write

$$\|G_L(\bar{X}_k, \bar{X}_{k+1}) - G_L(\widehat{X}_k, \widehat{X}_{k+1})\|_p \leq K_{\mathrm{Lip}} \left(\|\bar{X}_k - \widehat{X}_k\|_{rp} + \|\bar{X}_{k+1} - \widehat{X}_{k+1}\|_{rp} \right) \left(1 + \|\bar{X}_k\|_{2ps}^2 + \|\widehat{X}_{k+1}\|_{2ps}^2\right)$$

where $s = \frac{r}{r-1}$ and $r \in (1, 1 + \frac{d}{p})$. Since $\widehat{X}_{k+1}$ is a quadratic optimal quantization of $\widetilde{X}_{k+1}$, it is also stationary. This property, combined with Jensen's inequality, yields

$$\|\widehat{X}_{k+1}\|_{2ps}^2 = \|\mathbb{E}(\widetilde{X}_{k+1} | \widehat{X}_{k+1}\|_{2ps}^2 \leq \|\widetilde{X}_{k+1}\|_{2ps}^2.$$

And, using inequality (6.26), one obtains

$$\|\widehat{X}_{k+1}\|_{2ps}^2 \leq \|\bar{X}_0\|_{2ps} e^{t_k(C_1+C_2)/2ps} + \left(\frac{C_3}{C_1+C_2} e^{t_{k-1}(C_1+C_2)}\right)^{\frac{1}{ps}}$$

where C_1, C_2 and C_3 are constants defined in Lemma 6.2.4. This, combined with the fact that $\bar{X}_k \in L^r$ for every $r \in (0, +\infty)$ (by a property of the Euler scheme), yields the existence of a finite positive constant C_0 such that $1 + \|\bar{X}_k\|_{2ps}^2 + \|\widehat{X}_{k+1}\|_{2ps}^2 \leq C_0$. Then,

$$\|G_L(\bar{X}_k, \bar{X}_{k+1}) - G_L(\widehat{X}_k, \widehat{X}_{k+1})\|_p \leq C_0 K_{\text{Lip}} \big(\|\bar{X}_k - \widehat{X}_k\|_{rp} + \|\bar{X}_{k+1} - \widehat{X}_{k+1}\|_{rp}\big). \tag{7.25}$$

At this stage, we can proceed with the estimation of the upper bound. For every $k \in \{1, \dots, n\}$, one has

$$\begin{aligned}
\|L_k - \widehat{L}_k\|_p &\leq \Big\| \mathbb{E}\Big(G_L(\bar{X}_k, \bar{X}_{k+1}) L_{k+1} \mathbb{1}_{\bar{X}_k \geq L} - G_L(\widehat{X}_k, \widehat{X}_{k+1}) \widehat{L}_{k+1} \mathbb{1}_{\widehat{X}_k \geq L} \,\big|\, \mathcal{F}_k \Big) \Big\|_p \\
&\leq \Big\| \, G_L(\bar{X}_k, \bar{X}_{k+1}) L_{k+1} \mathbb{1}_{\bar{X}_k \geq L} - G_L(\widehat{X}_k, \widehat{X}_{k+1}) \widehat{L}_{k+1} \mathbb{1}_{\widehat{X}_k \geq L} \Big\|_p \\
&\leq \Big\| G_L(\bar{X}_k, \bar{X}_{k+1}) L_{k+1} (\mathbb{1}_{\bar{X}_k \geq L} - \mathbb{1}_{\widehat{X}_k \geq L}) + \mathbb{1}_{\widehat{X}_k \geq L} \Big(G_L(\bar{X}_k, \bar{X}_{k+1}) L_{k+1} - G_L(\widehat{X}_k, \widehat{X}_{k+1}) \widehat{L}_{k+1} \Big) \Big\|_p \\
&\leq \|G_L\|_{\sup} \|L_{k+1}\|_p \Big\| \mathbb{1}_{\bar{X}_k \geq L} - \mathbb{1}_{\widehat{X}_k \geq L} \Big\|_p + \|\mathbb{1}_{\widehat{X}_k \geq L}\|_p \Big\| G_L(\bar{X}_k, \bar{X}_{k+1}) L_{k+1} - G_L(\widehat{X}_k, \widehat{X}_{k+1}) \widehat{L}_{k+1} \Big\|_p .
\end{aligned} \tag{7.26}$$

It is clear that $\|G_L\|_{\sup} < 1$ and $\|L_{k+1}\|_p \leq \|g_n\|_{\sup}$. Moreover,

$$\mathbb{1}_{\bar{X}_k \geq L} - \mathbb{1}_{\widehat{X}_k \geq L} = \mathbb{1}_{\min(\bar{X}_k, \widehat{X}_k) \leq L \leq \max(\bar{X}_k, \widehat{X}_k)}$$

so that, by applying Holder's inequality with the conjugate coefficients p' and q',

$$\begin{aligned}
\big\| \mathbb{1}_{\bar{X}_k \geq L} - \mathbb{1}_{\widehat{X}_k \geq L} \big\|_p^p &= \int_{\mathbb{R}^d} \mathbb{1}_{\min(\bar{X}_k, \widehat{X}_k) \leq L \leq \max(\bar{X}_k, \widehat{X}_k)} dP \\
&\leq \left(\int_{\mathbb{R}^d} \mathbb{1}_{\min(\bar{X}_k, \widehat{X}_k) \leq L \leq \max(\bar{X}_k, \widehat{X}_k)} d\lambda_d \right)^{\frac{1}{p'}} \left(\int_{\mathbb{R}^d} f^{q'} d\lambda_d \right)^{\frac{1}{q'}} \\
&\leq \|\bar{X}_k - \widehat{X}_k\|_1^{\frac{1}{p'}} \times C \\
&\leq \kappa' \|\bar{X}_k - \widehat{X}_k\|_p^{\frac{1}{p'}}
\end{aligned}$$

for a finite constant C and $\kappa' > 0$ where the previous-to-last inequality is due to the fact that $f \in L^p(\lambda_d), p > 1$ and that one can choose q' as close to 1 as possible. For convenience, we denote $T = G_L(\bar{X}_k, \bar{X}_{k+1}) L_{k+1} - G_L(\widehat{X}_k, \widehat{X}_{k+1}) \widehat{L}_{k+1}$. We have

$$\begin{aligned}
\|T\|_p \leq & \Big\| G_L(\bar{X}_k, \bar{X}_{k+1})(L_{k+1} - \widehat{L}_{k+1}) + \widehat{L}_{k+1}(G_L(\bar{X}_k, \bar{X}_{k+1}) - G_L(\widehat{X}_k, \widehat{X}_{k+1})) \Big\|_p \\
\leq & \|G_L\|_{\sup} \|L_{k+1} - \widehat{L}_{k+1}\|_p + \|\widehat{L}_{k+1}\|_p \Big\| G_L(\bar{X}_k, \bar{X}_{k+1}) - G_L(\widehat{X}_k, \widehat{X}_{k+1}) \Big\|_p \\
\leq & \|L_{k+1} - \widehat{L}_{k+1}\|_p + \|g_n\|_{\sup} \Big\| G_L(\bar{X}_k, \bar{X}_{k+1}) - G_L(\widehat{X}_k, \widehat{X}_{k+1}) \Big\|_p .
\end{aligned} \tag{7.27}$$

Finally, we combine (7.25), (7.26) and (7.27) to obtain

$$\|L_k - \widehat{L}_k\|_p \leq \kappa \|g_n\|_{\sup} \|\bar{X}_k - \widehat{X}_k\|_p^{\frac{1}{p}} + \|L_{k+1} - \widehat{L}_{k+1}\|_p + \|g_n\|_{\sup} C_0 K_{\text{Lip}} \left(\|\bar{X}_k - \widehat{X}_k\|_{pr} + \|\bar{X}_{k+1} - \widehat{X}_{k+1}\|_{pr} \right).$$

To control the errors $\|\bar{X}_k - \widehat{X}_k\|_{pr}$, we rely on the distortion mismatch Theorem 3.3. For this, we chose $r = 1 + \frac{d}{2p} \in \left(1, 1 + \frac{d}{p}\right)$ so that $pr = p + \frac{d}{2} \in (p, p + d)$ and one can handle these error terms.
Finally, a backward induction yields the result. □

Remark 7.5.6. *The errors $\|\bar{X}_k - \widehat{X}_k\|_p$ and $\|\bar{X}_k - \widehat{X}_k\|_{p+\frac{d}{2}}$ appearing in Theorem 7.5.5 are quantization errors. It has been shown that the L^p-quantization error $\|\bar{X}_k - \widehat{X}_k\|_p$ induced by the L^p-optimal and L^p-recursive quantization of a X_k of size N_k is of $\mathcal{O}(N_k^{-\frac{1}{d}})$ where d is the dimension. However, for what concerns the errors $\|\bar{X}_k - \widehat{X}_k\|_{p+\frac{d}{2}}$, their rate of convergence is given by the distortion mismatch property, established first in [33] and then developed in [65]. In fact, since $p + \frac{d}{2} \in (p, p+d)$, this means that one can apply Theorem 6.2.2 and deduce that the quantization erros are of $\mathcal{O}(N^{-\frac{1}{d}})$ as well. Consequently, having in mind that K_{Lip} depends on n, one has that*

$$\|L_k - \widehat{L}_k\|_p = \mathcal{O}\left(\frac{n}{N^{\frac{1}{d}} + nN^{\frac{1}{pd}}}\right).$$

Hence, to obtain acceptable converging upper bounds and error margins, it suffices to choose n small enough and N large enough.

7.5.2 Algorithmics

In this section, we expose the numerical technique for the approximation of the price of a Down-and-Out Barrier option. The computation of the prices of the other Barrier options is identical, with a trivial change of the functions appearing in the payoff. We consider a Down-and-Out option with maturity T, strike price K, barrier L and whose payoff is given by

$$h_T(X_T) = g(X_T)\mathbb{1}_{\inf_{t\in[0,T]} X_t \geq L}$$

where g is a bounded function. Its price is given by

$$V_{DO} = e^{-rT}\mathbb{E}\left[g(X_n)\mathbb{1}_{\inf_{t\in[0,T]} X_t \geq L}\right] = e^{-rT}\mathbb{E}\left[g(\bar{X}_n)\prod_{k=1}^{n}\left(1 - F_L(\bar{X}_{k-1}, \bar{X}_k)\right)\right]$$

where $(\bar{X}_k)_{1\leq k\leq n}$ is the Euler scheme corresponding to $(X_t)_{0\leq t\leq T}$. As already mentioned, the second term is based on the theory of Brownian diffusion bridges.

In this chapter, our aim is to approximate these prices by recursive quantization. So, after computing the recursive quantization sequences $(\widehat{X}_k)_{0\leq k\leq n}$ of $(\bar{X}_k)_{0\leq k\leq n}$ as detailed in the previous Chapter 6, the price $\widehat{V}_{DO}$ is equal the initial value L_0 of the following backward dynamic programming principle

$$\begin{cases} \widehat{L}_n &= g(\widehat{X}_n) \\ \widehat{L}_k &= \mathbb{E}_k\left(1 - F_L(\widehat{X}_k, \widehat{X}_{k+1})\right), \qquad k = 0, \dots, n-1 \end{cases} \tag{7.28}$$

If we denote $\left(x_1^k, \dots, x_{N_k}^k\right)$ the recursive quantizer of size N_k of $\bar{X}_k$ at time t_k, it is clear that there exists a sequence of functions $(\hat{l}_k)_{0\leq k\leq n}$ such that $\widehat{L}_k = \hat{l}_k(X_k)$ for every $k \in \{0, \dots, n\}$ and defined by the following Backward Dynamic Programming Principle (BDPP)

$$\begin{cases} \hat{l}_n(x_i^n) &= g(x_i^n)\mathbb{1}_{x_i^n \geq L}, \quad i = 1, \dots, N_n, \\ \hat{l}_k(x_i^k) &= \sum_{j=1}^{N_{k+1}} p_{ij}^k\, \hat{l}_{k+1}(x_j^{k+1})\Big(1 - F_L(x_i^k, x_j^{k+1})\Big), \quad i = 1, \dots, N_k,\ k = 1, \dots, n. \end{cases} \tag{7.29}$$

where $(p_{ij}^k)_{i,j}$ is the transition weight from x_i^k at time t_k to x_j^{k+1} at time t_{k+1}.

Likewise, the BDPP corresponding to the computation of the price of an Up-and-Out option is given by the following (with the same notations as for a Down-and-Out option)

$$\begin{cases} \hat{l}_n(x_i^n) &= g(x_i^n)\mathbb{1}_{x_i^n \leq L}, \quad i = 1, \dots, N_n, \\ \hat{l}_k(x_i^k) &= \sum_{j=1}^{N_{k+1}} p_{ij}^k\, \hat{l}_{k+1}(x_j^{k+1}) G_L(x_i^k, x_j^{k+1}), \quad i = 1, \dots, N_k,\ k = 1, \dots, n. \end{cases} \tag{7.30}$$

7.5.3 Numerical examples

In this section, we give two examples: the pricing of a Down-and-Out call option in a Black-Scholes model and, the pricing of an Up-and-Out call option in a CEV model. In both cases, the time discretization is established following an Euler scheme with n time steps and, for the space discretization, we build quantizers of $(\bar{X}_k)_{0\leq k\leq n}$ by standard recursive quantization (RQ), greedy recursive quantization (GRQ) and optimal quantization (OQ). We also compute the prices by a Monte Carlo simulation with control variate (MC-VC). The different quantization techniques are already explained in details in the previous chapter and the MC-VC technique is detailed in the following in the case of a Down-and-Out Call option, it is the same principle for an Up-and-Out option.

The goal is to compute

$$C_{do} = \mathbb{E}[Y] = \mathbb{E}\Big[e^{-rT}(X_T - K)_+ \prod_{k=1}^{n} \big(1 - F_L(X_{k-1}, X_k)\big)\Big].$$

where we denote $Y = e^{-rT}(X_T - K)_+ \prod_{k=1}^{n} \big(1 - F_L(X_{k-1}, X_k)\big)$. We start by noticing that the sum of the price of a Down-and-Out Call option and the price of a Down-and-In Call option is the price of a standard European Call option. In fact,

$$\begin{aligned}
&\mathbb{E}\Big[e^{-rT}(X_T - K)_+ \prod_{k=1}^{n} \big(1 - F_L(X_{k-1}, X_k)\big)\Big] + \mathbb{E}\Big[e^{-rT}(X_T - K)_+ \Big(1 - \prod_{k=1}^{n} \big(1 - F_L(X_{k-1}, X_k)\big)\Big)\Big] \\
&= \mathbb{E}\left[e^{-rT}(X_T - K)_+\right] \\
&= \mathbb{E}\left[(X_0 - Ke^{-rT})_+\right].
\end{aligned}$$

This yields that

$$C_{do} = \mathbb{E}[Y] = \mathbb{E}[Y']$$

where $Y' = (X_0 - Ke^{-rT})_+ - e^{-rT}(X_T - K)_+ \Big(1 - \prod_{k=1}^{n} \big(1 - F_L(X_{k-1}, X_k)\big)\Big)$.
At this stage, we introduce the variable

$$\Xi = Y - Y' = e^{-rT}(X_T - K)_+ - (X_0 - Ke^{-rT})_+$$

satisfying $\mathbb{E}[\Xi] = 0$ and $\mathrm{Var}(\Xi) > 0$. Then, we introduce, for every $\lambda \in \mathbb{R}$, $Y^\lambda = Y - \lambda\Xi$ and notice that

$$\mathrm{Var}(Y^\lambda) = \lambda^2 \mathrm{Var}(\Xi) - 2\lambda \mathrm{Cov}(Y, \Xi) + \mathrm{Var}(Y)$$

attains its minimum at

$$\lambda_{\min} = \frac{\mathrm{Cov}(Y, \Xi)}{\mathrm{Var}(\Xi)} = 1 + \frac{\mathrm{Cov}(Y', \Xi)}{\mathrm{Var}(\Xi)}.$$

Hence, $\mathbb{E}[Y^{\lambda_{\min}}] = \mathbb{E}[Y]$ and $\mathrm{Var}(Y^{\lambda_{\min}}) < \mathrm{Var}(Y)$. Consequently, we approximate the price of the Down-and-Out call option, with more precision, by

$$C_{do} = \mathbb{E}[Y^{\lambda_{\min}}].$$

From a practical point of view, this computation is realized via the following steps

- Start by generating M independent copies $(Y_m, \Xi_m)_{1\leq m\leq M}$ of (Y, Ξ),
- Compute

$$V_M = \frac{1}{M}\sum_{m=1}^{M} \Xi_m^2 \quad \text{and} \quad C_M = \frac{1}{M}\sum_{m=1}^{M} Y_m \Xi_m$$

and, then,

$$\overline{Y}_M = \frac{1}{M}\sum_{m=1}^{M} Y_m \quad \text{and} \quad \overline{\Xi}_M = \frac{1}{M}\sum_{m=1}^{M} \Xi_m$$

• Finally, one has

$$\lambda_M = \frac{C_M}{V_M} \quad \text{and} \quad \overline{Y}_M^{\lambda_M} = \overline{Y}_M - \lambda_M \overline{\Xi}_M.$$

Thus, by the strong law of large numbers, one has $\lambda_M \to \lambda_{\min}$ and $\overline{Y}_M^{\lambda_M} \to \mathbb{E}[Y]$ yielding the desired estimator of the price of the Down-and-Out Call option.

Remark 7.5.7. *Another interesting alternative to compute λ_M is optimal quantization. Based on the optimal quantizers $\Gamma_k^N = (x_i^k)_{1\le i\le N}$, $k \in \{1,\dots,n\}$, of $(\bar{X}_k)_{1\le k\le n}$ and the corresponding Voronoï weights p_i^k, the idea is the following: We start by computing*

$$y_i = e^{-rT}(x_i^n - K)_+ \prod_{k=1}^{n}(1 - F_L(x_i^{k-1}, x_i^k)) \quad \textit{and} \quad \xi_i = e^{-rT}(x_i^n - K)_+ - (X_0 - Ke^{-rT})_+.$$

Then, we compute

$$C_M = \sum_{i=1}^{N}\left(y_i\xi_i \prod_{k=1}^{n} p_i^k\right) \quad \textit{and} \quad V_M = \sum_{i=1}^{N}\left(\xi_i^2 \prod_{k=1}^{n} p_i^k\right).$$

Finally,

$$\lambda_M = \frac{C_M}{V_M}.$$

Down-and-Out Call option in a Black-Scholes model Let $(\bar{X}_k)_{0\le k\le n}$ be a time-discretized diffusion process defined by (7.6). We consider $n = 15$ time steps and build quantizers of size $N_k = N = 100$ for every $k \in \{1,\dots,n\}$ by the different methods mentioned previously. Then, we estimate the price of the Down-and-Out Call option via (7.29). The parameters of this example are

$$K = 130\,, \quad T = 1\,, \quad X_0 = 130\,, \quad r = 0.15\,, \quad \sigma = 0.07.$$

For the Monte Carlo simulations with control variate, we use samples of size $M = 2.10^5$. The benchmark is given by the exact price of a Down-and-Out Call option, given in [27] by the following closed formula

$$C_{do} = X_0\Phi_0(d_1(X_0)) - Ke^{-rT}\Phi_0(d_2(X_0)) - \left(\frac{L}{X_0}\right)^{\frac{2\nu}{\sigma^2}}\left[\frac{L^2}{X_0}\Phi_0\left(d_1\left(\frac{L^2}{X_0}\right)\right) - Ke^{-rT}\Phi_0\left(d_2\left(\frac{L^2}{X_0}\right)\right)\right] \tag{7.31}$$

where $\nu = r - \frac{\sigma^2}{2}$, Φ_0 is the c.d.f. of $\mathcal{N}(0,1)$ and, for every x,

$$d_1(x) = \frac{\log\left(\frac{x}{K}\right) + T\left(r + \frac{\sigma^2}{2}\right)}{\sigma\sqrt{T}} \qquad \text{and} \qquad d_2(x) = d_1(x) - \sigma\sqrt{T}.$$

In Table 7.9, we expose the values obtained by the different methods, for barriers L varying between 115 and 130, and the errors induced by the comparison to the benchmark. Furthermore, we compare the results obtained by optimal quantization when we compute the transition weight matrices exactly by Gaussian quadrature formulas and the results when we approximate them by a certain function, see Section 6.5 of Chapter 6 for further details on this topic. The errors are exposed in Table 7.10 where we deduce that the difference is not in the precision, but rather in the cost of time.

L	**QR**		**GRQ**		**QO**		**MC-VC**		**Exact**
	Value	Error	Value	Error	Value	Error	Value	Error	
115	18.069	0.087	18.06	0.096	18.066	0.09	18.091	0.065	18.156
120	18.058	0.078	18.039	0.097	18.044	10.092	18.096	0.04	18.136
125	17.226	0.087	17.209	0.0104	17.215	0.098	17.271	0.042	17.313
129	7.929	0.05	7.937	0.042	7.941	0.038	8.004	0.025	7.979
Average		0.085		0.09		0.087		0.043	

Table 7.9: Pricing of a Down-and-Out call option in a Black-Scholes model discretized according to an Euler scheme and recursive (RQ), greedy recursive (GRQ), optimal (OQ) quantization and Monte Carlo with control Variate (MC-VC) for different values of L.

L	**Approximated value**	**Exact value**
115	0.093	0.09
120	0.094	0.092
125	0.1	0.098
129	0.039	0.038
Average error	0.088	0.086

Table 7.10: Pricing of a Down-and-Out Call option in a Black-Scholes model by optimal quantization with transition weights computed exactly and approximately for different values of L.

Up-and-Out Call option in a CEV model We consider a process $(X_t)_{0\leq t\leq T}$ following a CEV model and discretize it following a Euler scheme, i.e. $(\bar{X}_k)_{0\leq k\leq n}$ is given by (7.11). We consider $n = 15$ time steps and build quantizers of size $N = 100$ by the different methods mentioned above. The price of the option is computed via the recursion (7.30). The parameters of this example are

$$X_0 = 100\,, \quad T = 1\,, \quad \delta = 0.5\,, \quad r = 0.15\,, \quad K = 100\,, \quad \vartheta = 1.$$

Note that we are aware that such a level for the interest rate is not realistic but we made this choice for numerical purposes in order to check the robustness of the method. In this case, the benchmark is given by a Monte Carlo simulation with control variate (MC-VC) of size 2.10^5. These results and the corresponding errors are exposed in Table 7.11. We recall that, in a CEV model, the transition weight matrices of the optimal quantization tree are obtained by Monte Carlo simulations coupled with a nearest neighbor search.

L	**RQ**		**GRQ**		**OQ**		**MC-VC**
	Exact	Error	Exact	Error	Exact	Error	
110	0.433	0.008	0.435	0.006	0.438	0.003	0.441
115	1.782	0.018	1.79	0.01	1.803	0.016	1.8
120	4.218	0.021	4.216	0.023	4.229	0.01	4.239
125	7.17	0.031	7.188	0.03	7.174	0.011	7.201
130	9.943	0.033	9.947	0.03	9.899	0.009	9.910
Average Error		0.019		0.014		0.009	

Table 7.11: Pricing of an Up-and-Out Call option in a CEV model discretized according to an Euler scheme and recursive (RQ), greedy recursive (GRQ), optimal (OQ) quantization and Monte Carlo with control Variate (MC-VC) for different values of L.

Bibliography

[1] Baldi P. (1995). Exact asymptotics for the probability of exit from a domain and applications to simulations, *The Annals of Applied Probability,*, 23(4): 1644-1670.

[2] Bally V. (1997). Approximation scheme for solutions of BSDE. *Backward Stochastic Differential Equations* (N. El Karoui and L. Mazliak, eds.) 177–191. Pitman, London.

[3] Bally V. & Pagès G. (2003). Error analysis of the quantization algorithm for obstacle problems, *Stochastic Processes & Their Applications,*, **1**: 1-40.

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

[5] Bally V., Pagès G. & Printemps J. (2001). A stochastic quantization method for non-linear problems, *Monte Carlo Methods and Appl.,*, 1: 21-34.

[6] 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, 359 pp.

[7] Bender C. & Denk R. (2007). A forward scheme for backward SDEs, *Stochastic Processes & Their Applications,*, **117**(12): 1793-1812.

[8] Benveniste A., Métivier M. & Priouret P. (1987). *Algorithmes adaptatifs and approximations stochastiques*. Masson, Paris, 367 pp. English updated translation by Wilson S.S. (2012). *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag Berlin and Heidelberg, xi+364 pp.

[9] Bouchard B. & Touzi N. (2004). Discrete-time approximation and Monte-Carlo simulation of backward stochastic differential equations, *Stochastic Processes & Their Applications,*, **111**(2): 175-206.

[10] Bouleau N. & Lépingle D. (1993). Numerical Methods for Stochastic Processes, Wiley-Interscience.

[11] Bouton C. & Pagès G. (1993). Self-organization and *a.s.* convergence of the one-dimensional Kohonen algorithm with non-uniformly distributed stimuli. *Stochastic Process. Appl.*, **47**(2):249-274.

[12] Brancolini A., Buttazzo G., Santambrogio F., & Stepanov E. (2009). Long- term planning versus short-term planning in the asymptotical location problem, *ESAIM: Control Optim. Calc. Var.,* **15**(3):509–524.

[13] Briand P. & Labart C. (2014). Simulation of BSDEs by Wiener chaos expansion, *Ann. Appl. Probab.,*, **24**(3): 1129-1171.

[14] Bundschuh P. & Zhu Y.C. (1993). A method for exact calculation of the discrepancy of low-dimensional point sets I, Abh. Math. Sem. Univ. Hamburg 63, 115-133.

[15] Callegaro G., Fiorin L. & Grasselli M. (2017). Pricing via recursive quantization in stochastic volatility models, *Quantitative Finance*, **17** (6):855-872.

[16] Conze A. & Viswanathan (1991). Path Dependent Options: The Case of Lookback Options, *The Journal of Finance*, 46(5): 1893-1907.

[17] Crisan D., Manolarakis K. & Touzi N. (2010). On the Monte Carlo simulation of BSDE's: an improvement on the malliavin weights., *Stochastic Processes and their Applications,*, 120: 1133-1158.

[18] Cvitanic J, & Ma J.(2001) Reflected forward–backward SDEs and obstacle problems with boundary conditions, *J. Appl. Math. Stochastic Anal.,* **14** (2):113-138.

[19] Delattre S., Graf S., Luschgy H. & Pagès G. (2004). Quantization of probability distributions under norm-based distortion measures, *Statist. Decisions,* **22**(4):261-282.

[20] Doerr C., Gnewuch M. & Wahlström M. (2014). Calculation of discrepancy measures and applications, *A Panorama of Discrepancy Theory,* 621:678.

[21] Du Q., Faber V. & Gunzburger M. (1999): Centroidal Voronoi tessellations: Applications and algorithms, *SIAM Review*, **41**:637-676.

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

[23] El Nmeir R. & Pagès G. (work in progress). Quantization-based approximation of reflected BSDEs with extended upper bounds for recursive quantization.

[24] El Nmeir R., Lushgy H. & Pagès G. (2020). New approach to greedy vector quantization, ArXiv (available at http://arxiv.org/abs/2003.14145).

[25] El Karoui N., Kapoudjan C., Pardoux E., Peng S. & Quenez M.C. (1997). Reflected solutions of Backward Stochastic Differential Equations and related obstacle problems for PDEs., *Ann. Probab.*, **25**(2): 702-737.

[26] El Karoui N., Peng S. & Quenez M.C. (1997). Backward stochastic differential equations in Finance, *Math. Finance*, **7**(1): 1-71.

[27] Epps T.W.(2000). Pricing Derivative Securities (English Edition), *World Scientific*, Singapore.

[28] Fiorin L., Pagès G. & Sagna A. (2019). Product Markovian quantization of a diffusion process with applications to finance, *Methodol. Comput. Appl. Probab.*, **21**(4): 1087-1118.

[29] Fort J.C. & Pagès G. (2002). Asymptotics of optimal quantizers for some scalar distributions, *Journal of Computational and Applied Mathematics*, **146**: 253-275.

[30] Gersho A. & Gray R.M. (1988). Special issue on Quantization, I-II (A. Gersho and R.M. Gray eds.), *IEEE Trans. Inform. Theory*, **28**.

[31] Gobet E. (2000). Weak approximation of killed diffusion using Euler schemes, *Stoch. Proc. and their Appl.*87:167-197.

[32] Graf S. & Lushgy H. (2000). Foundations of Quantization for Probability Distributions, Lectures Notes in Math. 1730. Springer, Berlin.

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

[34] Henry-Labordère P., Tan X. & Touzi N. (2014). A numerical algorithm for a class of BSDEs via the branching process, *Stochastic Process. Appl.* 124(2):1112-1140.

[35] Hu Y., Nualart T. & Song X. (2011). Malliavin calculus for backward stochastic differential equations and applications to numerical solutions, *The Annals of Applied Probability,* 21(6):2379–2423.

[36] Huang J., Subrahmanyam M. & Yu G. (1996). Pricing and Hedging American Options: A Recursive Integration Method., *The Review of Financial Studies,* 9(1):277-300.

[37] Illand C. (2012). Contrôle stochastique par quantification et applications à la finance, PhD thesis, UPMC.

[38] Kelly J. (1955). General Topology. Van Nostrand, Princeton.

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

[40] Kuipers L. & Niederreiter H. (1974). Uniform distribution of sequences, Wiley.

[41] Lemaire V. & Pagès G. (2010). Unconstrained Recursive Importance Sampling, *Annals of Applied Probability,* **20**:1029-1067427-469.

[42] Lemaire V., Montes T. & Pagès G. (2019). New weak error bounds and expansions for optimal quantization, *Journal of Computational and Applied Mathematics,* p. 112670. issn: 0377-0427.

[43] Luschgy, H. (2012). *Martingale in diskreter Zeit*, Theorie und Anwendungen Reihe: Springer-Lehrbuch Masterclass, Springer, Berlin, 452 pp.

[44] 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.

[45] Lushgy H. & Pagès G. (2015). Greedy vector quantization, *Journal of Approximation Theory*, **198**: 111-131.

[46] Lushgy H. & Pagès G. (2015). Greedy vector quantization (extended version), *ArXiv.* (Available at https://arxiv.org/abs/1409.0732)

[47] Ma J. & Zhang J. (2004). Representation theorems for Backward stochastic differential equations, *Ann. Appl. Probab.*, 12(4):1390-1418.

[48] Ma J. & Zhang J. (2005). Representation and regularities for solutions to BSDEs with reflections, *Stochastic Processes and their applications*, 115:539-569.

[49] Ma J. & Wang Y. (2009). On Variant Reflected Backward SDEs, with Applications, *J. Appl. Math. Stochastic Anal.*, Vol. Art. ID 854768, pp. 26.

[50] Ma J. & Wang Y. (1994). Solving forward–backward stochastic differential equations explicitly—a four step scheme,*Probab. Theory Relat/Fields*, **98**:339-359.

[51] McWalter A., Rudd R., Kienitz J. & Platen E. (2018). Recursive marginal quantization of higher-order schemes,*Quantitative Finance*, **18**(4):693-706.

[52] Montes T. (2020). Numerical methods by optimal quantization in finance, PhD thesis, Sorbonne Université.

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

[54] 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: M. Verleysen (Ed.), Proceedings of the ESANN' 93, Bruxelles, Quorum Editions, (1993), 221-228).

[55] Pagés G. (2007). Quadratic optimal functional quantization methods and numerical applications. *Proceedings of MCQMC, Ulm'06, Springer, Berlin*, 101-142.

[56] Pagès G. (2015). Introduction to optimal vector quantization and its applications for numerics. CEMRACS 2013-modelling and simulation of complex systems : Stochastic and deterministic approaches. *ESAIM.*

[57] Pagès G. (2018). Numerical probability: An introduction with applications to finance, Springer-Verlag, xvi+579p.

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

[59] Pagès G., Pham H. & Printemps J. (2004). Optimal quantization methods and applications to numerical problems in finance, Rachev S.T. (eds) Handbook of Computational and Numerical Methods in Finance. Birkhäuser, Boston, MA.

[60] Pagès G., Pham H. & Printemps J. (2004). An Optimal markovian quantization algorithm for multidimensional stochastic control problems, *Stochastics and Dynamics*, 4:501-545.

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

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

[63] Pagès G. & Sagna A. (2014). Recursive marginal quantization of the Euler scheme of a diffusion. *Appl. Math. Finance*, **22** (5), 463–498.

[64] Pagès G. & Sagna A. (2016). Improved error bounds for quantization based numerical schemes for BSDE and nonlinear filtering, (extended version). Available at: http://Arxiv.Org/Abs/1510.01048.

[65] Pagès G. & Sagna A. (2018). Improved error bounds for quantization based numerical schemes for BSDE and nonlinear filtering, *Stochastic Processes and their Applications*, **128** 847-883.

[66] Pardoux E. & Peng S.G. (1990). Adapted solutions of backward stochastic differential equation, *Systems Control Lett.,* **14**(4)55-61.

[67] Pagès G. & Sellami A. (2011). Convergence of Multi-Dimensional Quantized SDE's, In: Donati-Martin C., Lejay A., Rouault A. (eds) Séminaire de Probabilités XLIII. Lecture Notes in Mathematics, vol 2006. Springer, Berlin, Heidelberg.

[68] Pagès G. & Yu J. (2013). Pointwise convergence of the Lloyd algorithm in higher dimension, Technical report PMA 1604.

[69] Papanicolopulos S.(2016). New fully symmetric and rotationally symmetric cubature rules on the triangle using minimal orthonormal bases. *Journal of Computational and Applied Mathematics* 39-48.

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

[71] Sagna A. (2008). Universal L^s-rate-optimality of L^r-optimal quantizers by dilatation and contraction, *ESAIM: Probability and Statistics*, 13:218-246.

[72] Sagna A. (2011). Pricing of barrier options by marginal functional quantization, *Monte Carlo Methods Appl.*17(4):371398

[73] Trushkin A.V. (1984). Monotony of Lloyd's method II for log-concave density and convex error weighting function, *IEEE Trans. Inform. Theory*, 30, 380-383.

[74] Villeneuve S. & Zanette A. (2002). Parabolic A.D.I. methods for pricing American option on two stocks, *Math. Oper. Res.,* **27**121-149.

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

[76] Zhang J. (2004). A numerical scheme for BSDEs, *Ann. Appl. Probab.*, 14(1):459-488.