Optimal Quantization: Limit Theorems, Clustering and Simulation of the McKean-Vlasov Equation

Yating Liu

Optimal Quantization: Limit Theorems, Clustering and Simulation of the McKean-Vlasov Equation

Yating Liu

Laboratoire de Probabilités, Statistique et Modélisation - UMR 8001
Sorbonne Université

Thèse pour l'obtention du grade de:

Docteur de l'université Sorbonne Université

Dirigée par : Gilles Pagès

Rapportée par : Benjamin Jourdain (École des Ponts ParisTech)
Harald Luschgy (University of Trier)

Présentée devant un jury composé de :

Gérard Biau (Sorbonne Université)
François Bolley (Sorbonne Université)
Jean-François Chassagneux (Université Paris Diderot)
Marc Hoffmann (Université Paris-Dauphine)
Benjamin Jourdain (École des Ponts ParisTech)
Harald Luschgy (University of Trier)
Gilles Pagès (Sorbonne Université)
Clémentine Prieur (Université Grenoble Alpes)

École Doctorale de Sciences Mathématiques de
Paris-Centre
Section Mathématiques appliquées

3 Décembre 2019

Acknowledgements

First and foremost I would like to thank my supervisor Gilles Pagès. His course *Théorèmes limites pour les processus stochastiques* for Master 2 *Probabilités et modèles aléatoires* is one of the best courses I have ever attended and I am very grateful for the opportunity I'd been given to be a Ph.D. student under his supervision. Throughout the past three years, his inspiring vision and helpful advices are invaluable to me. His strict and ceaseless supervision on my work have been motivating me until the end of this thesis. It's a so special and unique experience to learn so much from him.

At Sorbonne University, I have benefited from a high-level education and I am grateful to all professors who have taught me for their excellent courses. Specifically, I'd like to thank Cédric Boutillier, who supervised me on a TER project in Master 1. Even though this project now appears only a toy example, this experience was my first motivation to pursue a doctoral degree, which makes me realize that it is not about "being a good student", but about "doing something interesting". I also want to express my gratitude for Yongdao Zhou, who was my professor in Sichuan University (now in Nankai University). His course *Stochastic process* was the first introduction to the Markov chain for me and interests me from then on in this field.

During these three years, I discussed with and learned a lot from people in Laboratoire de Probabilités, Statistique et Modélisation (LPSM). I want to thank Marie-Claire Quenez for a reference that she sent to me, which inspires me to establish the proof in Section 7.1, also Nicolas Fournier (and his co-author Arnaud Guillin, of course) and François Bolley, whose papers about the Wasserstein distance set a theoretical basis of many results in this thesis. I also want to thank other Ph.D. students and young researchers I have met here who are very helpful and kind to me: Babacar, Carlo, Chenguang, Côme, Cyril, Daphné, Éric, Henri, Junchao, Nicolas G., Nicolas T., Othmane, Paul, Rancy, Sarah, Shuo, Yoan, ... Especially to Qiming from whom I learned about the McKean-Vlasov equation for the first time. Every coffee-break discussion with him brings me so much inspiration.

My special thanks go to Armand, Guillermo, Léa, Romain and Robert, they are not only my closest colleagues but also friends that I can really trust. So many I've

discovered with them: lots of music, french expressions for the young people, and RDR2, the best game I've ever had. Everyday in the office, I feel relax, free and joyful when I stay with them - it makes me even a little sad when Guillermo and Romain finish their Ph.D. thesis, though in the same time I am happy for them.

I have been blessed to be surrounded by so many warm-hearted friends here and in China: Bo, François, Hang, Lilia, Shuyi, Yangjunjie... especially Jinhong and Mengjun. They accompany me even in the most difficult times, encourage me and make me feel confident. Thanks to the music of Joe Hisaishi and Macro Stereo, which comforts me in the overworking night. I also want to thank Jérôme Ballif for his professional advice, which helps to stabilize me after facing intractable difficulties at the end of these three years. By this I would also like to suggest foreign students like me to seek for professional help when need.

Special thanks to my family in China, for everything I can imagine. Each time I was asked "why mathematics", I always remembered those summer-evening-"mathematical"-questions-games I played with my father during my childhood. It is still in my mind that my father use our flip-flops to explain me how to move a wolf, a sheep and a box of cabbage to the other side of the river... It is not logical, but I felt like it all starts from that moment.

Most importantly, I want to thank my husband, Yassir, love of my life, for his unconditional love and support. Without him, I could never achieve so many, including this thesis.

Abstract

This thesis contains two parts. The first part addresses two limit theorems related to optimal quantization. The first limit theorem is the characterization of the convergence in the Wasserstein distance of probability measures by the pointwise convergence of L^p-quantization error functions on $\mathbb{R}^d$ and on a separable Hilbert space. The second limit theorem is the convergence rate of the optimal quantizer and the clustering performance for a probability measure sequence $(\mu_n)_{n\in\mathbb{N}^*}$ on $\mathbb{R}^d$ converging in the Wasserstein distance, especially when $(\mu_n)_{n\in\mathbb{N}^*}$ are the empirical measures with finite second moment but possibly unbounded support. The second part of this manuscript is devoted to the approximation and the simulation of the McKean-Vlasov equation, including several quantization based schemes and a hybrid particle-quantization scheme. We first give a proof of the existence and uniqueness of a strong solution of the McKean-Vlasov equation $dX_t = b(t, X_t, \mu_t)dt + \sigma(t, X_t, \mu_t)dB_t$ under the Lipschitz coefficient condition by using Feyel's method (see Bouleau (1988)[Section 7]). Then, we establish the convergence rate of the "theoretical" Euler scheme $\bar{X}_{t_{m+1}} = \bar{X}_{t_m} + hb(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m}) + \sqrt{h}\,\sigma(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m})Z_{m+1}$ and as an application, we establish functional convex order results for scaled McKean-Vlasov equations with an affine drift. In the last chapter, we prove the convergence rate of the particle method, several quantization based schemes and the hybrid scheme. Finally, we simulate two examples: the Burger's equation (Bossy and Talay (1997)) in one dimensional setting and the Network of FitzHugh-Nagumo neurons (Baladron et al. (2012)) in dimension 3.

Keywords: Optimal quantization, Wasserstein convergence characterization, K-means clustering, Simulation of McKean-Vlasov equation, Convex order.

Résumé

Cette thèse contient deux parties. Dans la première partie, on démontre deux théorèmes limites de la quantification optimale. Le premier théorème limite est la caractérisation de la convergence sous la distance de Wasserstein d'une suite de mesures de probabilité par la convergence simple des fonctions d'erreur de la quantification. Ces résultats sont établis en $\mathbb{R}^d$ et également dans un espace de Hilbert séparable. Le second théorème limite montre la vitesse de convergence des grilles optimales et la performance de quantification pour une suite de mesures de probabilité qui convergent sous la distance de Wasserstein, notamment la mesure empirique. La deuxième partie de cette thèse se concentre sur l'approximation et la simulation de l'équation de McKean-Vlasov. On commence cette partie par prouver, par la méthode de Feyel (voir Bouleau (1988)[Section 7]), l'existence et l'unicité d'une solution forte de l'équation de McKean-Vlasov $dX_t = b(t, X_t, \mu_t)dt + \sigma(t, X_t, \mu_t)dB_t$ sous la condition que les fonctions de coefficient b et σ sont lipschitziennes. Ensuite, on établit la vitesse de convergence du schéma d'Euler théorique de l'équation de McKean-Vlasov $\bar{X}_{t_{m+1}} = \bar{X}_{t_m} + hb(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m}) + \sqrt{h}\,\sigma(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m})Z_{m+1}$ et également les résultats de l'ordre convexe fonctionnel pour les équations de McKean-Vlasov avec $b(t, x, \mu) = \alpha x + \beta$, $\alpha, \beta \in \mathbb{R}$. Dans le dernier chapitre, on analyse l'erreur de la méthode de particule, de plusieurs schémas basés sur la quantification et d'un schéma hybride particule-quantification. À la fin, on illustre deux exemples de simulations: l'équation de Burgers (Bossy and Talay (1997)) en dimension 1 et le réseau de neurones de FitzHugh-Nagumo (Baladron et al. (2012)) en dimension 3.

Mots-clés: Quantification optimale, Caractérisation de la convergence Wasserstein, Classification non supervisée, Simulation de l'équation de McKean-Vlasov, Ordre convexe.

Contents

Résumé détaillé

La quantification optimale est originellement développée comme une méthode de transmission et compression des signaux par le Laboratoire bell en 1950s; elle est maintenant un outil largement utilisé dans le domaine de l'apprentissage non-supervisé et de la probabilité numérique. De façon générale, la quantification est une méthode d'approximation d'une mesure de probabilité μ en utilisant un K-uplet $x = (x_1, ..., x_K)$ et son vecteur de poids $w = (w_1, ..., w_K)$. L'estimateur de μ par la méthode de quantification s'écrit comme $\widehat{\mu}^x := \sum_{k=1}^K w_k \cdot \delta_{x_k}$, où δ_a est la masse de Dirac en a. On appelle $x = (x_1, ..., x_K)$ la *grille* de quantification (*quantizer* en anglais). Le poids $w = (w_1, ..., w_K)$ est souvent calculé par $w_k := \mu\big(C_k(x)\big)$, $k = 1, ..., K$, où $\big(C_k(x)\big)_{1\leq k\leq K}$ est la partition de Voronoï de $\mathbb{R}^d$.

Soit $\mathcal{P}_p(\mathbb{R}^d) := \big\{\mu \text{ mesure de probabilité sur } \mathbb{R}^d \mid \int_{\mathbb{R}^d} |\xi|^p \, \mu(d\xi) < +\infty\big\}$ et soit $\mathcal{W}_p$ la distance de Wasserstein d'ordre p sur $\mathcal{P}_p(\mathbb{R}^d)$. La fonction de distorsion de la quantification de $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ au niveau K et de l'ordre p, notée par $\mathcal{D}_{K,p}(\mu, \cdot)$, est définie par

$$x = (x_1, ..., x_K) \in \mathbb{R}^d \mapsto \mathcal{D}_{K,p}(\mu, x) := \int_{\mathbb{R}^d} \min_{1\leq k\leq K} |\xi - x_k|^p \, \mu(d\xi).$$

De plus, la fonction d'erreur de quantification est définie par $e_{K,p}(\mu, \cdot) = \mathcal{D}_{K,p}(\mu, \cdot)^{1/p}$. Si x^* satisfait $x^* \in \operatorname{argmin} e_{K,p}(\mu, \cdot) = \operatorname{argmin} \mathcal{D}_{K,p}(\mu, \cdot)$, on appelle x^* une grille optimale de μ au niveau K et d'ordre p. L'existence d'une telle grille optimale est établie dans Graf and Luschgy (2000)[Theorem 4.12] et Graf et al. (2007).

Parmi un large champ de propriétés et d'applications de la méthode de quantification, cette thèse se concentre sur deux théorèmes limites et l'application de la quantification optimale à la simulation de l'équation de McKean-Vlasov.

Partie I : Théorèmes limites de la quantification optimale (Chapitres 2 et 3)

Le Chapitre 2 présente la caractérisation de mesure de probabilité par la fonction d'erreur de quantification. Dans ce chapitre, on établit l'existence d'un niveau minimal

K^* tel que pour tout $K \geq K^*$,

- pour tout $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, $e_{K,p}(\mu, \cdot) = e_{K,p}(\nu, \cdot) \iff \mu = \nu$,
- pour tout $\mu_n \in \mathcal{P}_p(\mathbb{R}^d)$, $n \in \mathbb{N}^* \cup \{\infty\}$,

$$e_{K,p}(\mu_n, \cdot) \xrightarrow{n\to+\infty} e_{K,p}(\mu_\infty, \cdot) \text{ simplement} \iff \mathcal{W}_p(\mu_n, \mu_\infty) \xrightarrow{n\to+\infty} 0.$$

La preuve de ces équivalences est basée sur une approche géométrique qui est équivalente à l'existence d'une cellule de Voronoï bornée dans un diagramme de Voronoï. Cette existence peut se déduire d'un recouvrement minimal de la sphère d'unité par les boules d'unité fermées centrées sur cette sphère. Cette approche géométrique est vraie pour toutes les normes de $\mathbb{R}^d$. De plus, pour le cas quadratique, on établit le niveau minimal pour les caractérisations $K^* = 2$ par les méthodes d'analyse hilbertienne. Ce résultat de caractérisation peut s'étendre à un espace de Hilbert séparable quelconque. On définit aussi dans ce chapitre pour tout $K \geq K^*$ une distance basée sur la fonction d'erreur de quantification

$$\mathcal{Q}_{K,p} \coloneqq \|e_{K,p}(\mu, \cdot) - e_{K,p}(\nu, \cdot)\|_{\sup}$$

et on démontre que cette distance $\mathcal{Q}_{K,p}$ est équivalente à la distance de Wasserstein $\mathcal{W}_p$. En outre, on montre que $\big(\mathcal{P}_1(\mathbb{R}), \mathcal{Q}_{1,1}\big)$ est un espace complet et on fournit un contre-exemple montrant que $\big(\mathcal{P}_2(\mathbb{R}), \mathcal{Q}_{K,2}\big)$ n'est pas complet pour tout $K \geq 2$.

Dans le Chapitre 3, on établit la vitesse de convergence de la quantification optimale quadratique ($p = 2$) pour une suite de mesures de probabilité qui converge sous la distance de Wassserstein. Ce chapitre généralise deux papiers précédents Pollard (1982a) et Biau et al. (2008). Soient $\mu_n \in \mathcal{P}_2(\mathbb{R}^d), n \in \mathbb{N}^* \cup \{\infty\}$ telles que $\mathcal{W}_2(\mu_n, \mu_\infty) \xrightarrow{n\to+\infty} 0$. On note $x^{(n)}$ la grille optimale quadratique de μ_n pour tout $n \in \mathbb{N}^*$ et on définit

$$G_K(\mu_\infty) \coloneqq \big\{(x_1^*, ..., x_N^*) \mid (x_1^*, ..., x_N^*) \text{ est une grille optimale quadratique de } \mu_\infty\big\}$$

l'ensemble des grilles optimales quadratiques de μ_∞ au niveau K. On démontre la performance de quantification: pour tout $n \in \mathbb{N}^*$,

$$\mathcal{D}_{K,2}(\mu_\infty, x^{(n)}) - \inf_{x \in (\mathbb{R}^d)^K} \mathcal{D}_{K,2}(\mu_\infty, x) \leq 4e^*_{K,\mu_\infty} \mathcal{W}_2(\mu_n, \mu_\infty) + 4\mathcal{W}_2^2(\mu_n, \mu_\infty),$$

où $e^*_{K,\mu_\infty} \coloneqq \inf_{y \in \mathbb{R}^d} e_{K,2}(\mu_\infty, y)$ est l'erreur optimale de la quantification optimale. On démontre également la vitesse de convergence des grilles optimale: à partir d'un certain rang,

$$d\big(x^{(n)}, G_K(\mu_\infty)\big)^2 \leq C^{(1)}_{\mu_\infty} \mathcal{W}_2(\mu_n, \mu_\infty) + C^{(2)}_{\mu_\infty} \mathcal{W}_2^2(\mu_n, \mu_\infty).$$

sous la condition que la matrice hessienne $H_{\mathcal{D}_{K,\mu_\infty}}$ de $\mathcal{D}_{K,2}(\mu_\infty,\cdot)$ existe et soit définie positive. En outre, on donne aussi la formule exacte de la matrice hessienne $H_{\mathcal{D}_{K,\mu_\infty}}$ dans ce chapitre.

Soient $X_1,...,X_n$ variables aléatoires i.i.d qui suivent la mesure de probabilité μ et soit $\mu_n^\omega := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ la mesure empirique de μ. La deuxième partie du Chapitre 3 se concentre sur la valeur $\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x)$, qui est appelée la performance de la classification non supervisée (*la performance de clustering*) (voir Biau et al. (2008)). On établit deux bornes supérieures de la performance de clustering. Si $\mu\in\mathcal{P}_q(\mathbb{R}^d)$ avec $q>2$, le premier résultat qu'on obtient est

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x) \leq C_{d,q,\mu,K}\times\begin{cases} n^{-1/4}+n^{-(q-2)/2q} & \text{si } d<4 \text{ et } q\neq 4\\ n^{-1/4}\big(\log(1+n)\big)^{1/2}+n^{-(q-2)/2q} & \text{si } d=4 \text{ et } q\neq 4\\ n^{-1/d}+n^{-(q-2)/2q} & \text{si } d>4 \text{ et } q\neq d/(d-2)\end{cases},$$

où $C_{d,q,\mu,K}$ est une constante dépendant de d,q,μ et décroît en K d'ordre $K^{-1/d}$. Soit maintenant $\mu\in\mathcal{P}_2(\mathbb{R}^d)$. La deuxième borne qu'on obtient pour la performance de *clustering* est

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x) \leq \frac{2K}{\sqrt{n}}\Big[r_{2n}^2+\rho_K(\mu)^2+2r_1\big(r_{2n}+\rho_K(\mu)\big)\Big],$$

où $r_n := \big\|\max_{1\leq i\leq n}|X_i|\big\|_2$ et $\rho_K(\mu)$ est le rayon maximal des grilles optimales quadratiques i.e. $\rho_K(\mu) := \max\big\{\max_{1\leq k\leq K}|x_k^*| \;\big|\; (x_1^*,...,x_K^*) \text{ est une grille optimale de } \mu\big\}$. Si $\mu=\mathcal{N}(m,\Sigma)$, la loi normale multidimensionnelle, on a

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x) \leq C_\mu\cdot\frac{2K}{\sqrt{n}}\Big[1+\log n+\gamma_K\log K\big(1+\frac{2}{d}\big)\Big],$$

avec $\limsup_K\gamma_K=1$ et $C_\mu = 12\cdot\Big[1\vee\log\big(2\int_{\mathbb{R}^d}\exp(\frac{1}{4}|\xi|^4)\mu(d\xi)\big)\Big]$.

Partie II : Équation de McKean-Vlasov: méthode de particule, schémas basés sur la quantification et schéma hybride, ordre convexe (Chapitres 4, 5, 6 et 7)

L'équation McKean-Vlasov, qui est premièrement introduite dans McKean (1967), indique dans cette thèse une classe d'équations différentielles stochastiques avec les fonctions de coefficient dépendant non seulement de l'état de (X_t) mais aussi de la loi de (X_t). Plus précisément, l'équation McKean-Vlasov qu'on discute dans cette thèse est

définie comme suit,

$$\begin{cases} dX_t = b(t, X_t, \mu_t)dt + \sigma(t, X_t, \mu_t)dB_t \\ X_0 : (\Omega, \mathcal{F}, \mathbb{P}) \to \big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big) \text{ variable aléatoire} \\ \forall t \geq 0,\ \mu_t \text{ est la mesure de probabilité de } X_t \end{cases} . \tag{0.0.1}$$

Dans le Chapitre 5, on donne une preuve de l'existence et l'unicité de la solution forte de l'équation de McKean-Vlasov (0.0.1) par la méthode de Feyel (e.g. Bouleau (1988)[Section 7]) sous la condition lipschitzienne suivante

$$\begin{aligned} &\forall t \in [0,T], \forall x, y \in \mathbb{R}^d \text{ et } \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}^d), \exists L \ t.q. \\ &\qquad |b(t,x,\mu) - b(t,y,\nu)| \vee |||\sigma(t,x,\mu) - \sigma(t,y,\nu)||| \leq L\big[\,|x-y| + \mathcal{W}_p(\mu,\nu)\big]. \end{aligned} \tag{0.0.2}$$

L'idée de cette preuve est de définir une application Φ_C qui dépend d'une constante $C \in \mathbb{R}_+^*$ sur un espace produit " l'espace des processus $\times$ l'espace des mesures de probabilité de processus" comme suit

$$\begin{aligned} &(Y, P_Y) \mapsto \Phi_C(Y, P_Y) \\ &\qquad := \Big(\underbrace{\big(X_0 + \int_0^t b(s, Y_s, \nu_s)ds + \int_0^t \sigma(s, Y_s, \nu_s)dB_s\big)_{t\in[0,T]}}_{=:\Phi_C^{(1)}(Y,P_Y)}, P_{\Phi_C^{(1)}(Y,P_Y)}\Big) \end{aligned}$$

où pour un processus stochastique X, on note sa mesure de probabilité P_X (voir la Section 5.1 pour la définition détaillée de P_X), puis on montre que cet espace est complet et que Φ_C est une application contractante sur un sous-ensemble fermé si la constante C est assez grande. On en déduit l'existence et l'unicité forte de solution de l'équation de McKean-Vlasov en utilisant le théorème du point fixe.

Une fois qu'on obtient l'existence et l'unicité forte de la solution, on montre dans le Chapitre 5 la vitesse de convergence du schéma d'Euler théorique de l'équation de McKean-Vlasov (0.0.1), qui est défini par

$$\begin{cases} \bar{X}_{t_{m+1}} = \bar{X}_{t_m} + h \cdot b(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m}) + \sqrt{h}\,\sigma(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m})Z_{m+1} \\ \bar{\mu}_{t_m} \text{ est la mesure de probabilité de } \bar{X}_{t_m}, m = 0, ..., M \\ \bar{X}_0 = X_0 \end{cases} , \tag{0.0.3}$$

où $M \in \mathbb{N}^*$ est le nombre de discrétisations en temps et $t_m := \frac{T}{M} \cdot m,\ m = 0, ..., M$. Si b, σ satisfont (0.0.2) et

$$\forall t, s \in [0,T] \text{ t.q. } s < t, \forall x \in \mathbb{R}^d, \forall \mu \in \mathcal{P}(\mathbb{R}^d), \text{ il existe } \tilde{L}, \gamma \in \mathbb{R}_+ \text{ t.q.}$$

$$|b(t,x,\mu)-b(s,x,\mu)| \vee |||\sigma(t,x,\mu)-\sigma(s,x,\mu)||| \leq \tilde{L}\big(1+|x|+\mathcal{W}_p(\mu,\delta_0)\big)(t-s)^{\gamma}, \tag{0.0.4}$$

la vitesse de convergence du schema d'Euler théorique qu'on obtient est

$$\sup_{0\leq m\leq M}\mathcal{W}_p(\bar{\mu}_{t_m},\mu_{t_m}) \leq \left\| \sup_{0\leq m\leq M}\left|X_{t_m}-\bar{X}_{t_m}\right| \right\|_p \leq C_e h^{\frac{1}{2}\wedge\gamma}, \tag{0.0.5}$$

où C_e est une constante qui dépend de $b,\sigma,L,T,\tilde{L}$ et $\|X_0\|_p$.

Le Chapitre 6 établit le résultat de l'ordre convexe pour l'équation de McKean-Vlasov $(X_t)_{t\in[0,T]}$, $(Y_t)_{t\in[0,T]}$ définies par

$$\begin{aligned} dX_t &= (\alpha X_t+\beta)dt+\sigma(t,X_t,\mu_t)dB_t, \quad X_0\in L^p(\mathbb{P}),\\ dY_t &= (\alpha Y_t+\beta)\,dt+\theta(t,\,Y_t,\,\nu_t)\,dB_t, \quad Y_0\in L^p(\mathbb{P}), \end{aligned} \tag{0.0.6}$$

où $\alpha,\beta\in\mathbb{R}$ et pour tout $t\in[0,T]$, $\mu_t=\mathbb{P}\circ X_t^{-1}$, $\nu_t=\mathbb{P}\circ Y_t^{-1}$. Soient X,Y deux variables aléatoires à valeur dans un espace de Banach $(E,\|\cdot\|_E)$. Si on a $\mathbb{E}\,\varphi(X)\leq\mathbb{E}\,\varphi(Y)$ pour toutes les fonctions convexes $\varphi:E\to\mathbb{R}$ telle que $\mathbb{E}\,\varphi(X)$ et $\mathbb{E}\,\varphi(Y)$ soient bien définies, on dit que X est dominée par Y pour l'ordre convexe et on note cette relation d'ordre par $X\preceq_{cv}Y$. On définit respectivement les schémas d'Euler théorique de $(X_t)_{t\in[0,T]}$, $(Y_t)_{t\in[0,T]}$ par (0.0.3), et on les note par $\bar{X}_{t_m},\bar{Y}_{t_m},m=0,...,M$. Dans le Chapitre 6, on montre que le schéma d'Euler théorique de l'équation de McKean-Vlasov diffuse l'ordre de convexe i.e. $\bar{X}_{t_m}\preceq_{cv}\bar{Y}_{t_m},m=0,...,M$, sous les conditions que

- $X_0\preceq_{cv}Y_0$,
- pout tout $t\in[0,T]$, $x\in\mathbb{R}^d$, $\mu\in\mathcal{P}(\mathbb{R}^d)$, $\theta(t,x,\mu)\theta(t,x,\mu)^*-\sigma(t,x,\mu)\sigma(t,x,\mu)^*$ est une matrice définie positive,
- σ est convexe en x et croissante en μ par rapport à l'ordre convexe.

De plus, on en déduit, en utilisant une induction rétrogradé (backward) et la convergence du schéma d'Euler théorique (0.0.5), le résultat de l'ordre convexe fonctionnel pour les processus: pour une fonction convexe quelconque $F:\mathcal{C}([0,T],\mathbb{R}^d)\to\mathbb{R}$ telle que $F(X)$ et $F(Y)$ soient bien définies et

$$\forall\alpha\in\mathcal{C}([0,T],\mathbb{R}^d),\exists C\geq 0 \ \text{ t.q. } \ |F(\alpha)|\leq C(1+\|\alpha\|_{\sup}^r), \ \text{ avec } 1\leq r\leq p,$$

on a $\mathbb{E}\,F(X)\leq\mathbb{E}\,F(Y)$. En outre, ce résultat peut encore se généraliser aux fonctionnel du processus et de la loi du processus sous la forme de

$$G:(\alpha,(\gamma_t)_{t\in[0,T]})\in\mathcal{C}([0,T],\mathbb{R}^d)\times\mathcal{C}([0,T],\mathcal{P}_p(\mathbb{R}^d))\mapsto G(\alpha,(\gamma_t)_{t\in[0,T]})\in\mathbb{R},$$

telle que G est convexe en α, non décroissante en $(\gamma_t)_{t\in[0,T]}$ par rapport à l'ordre convexe et admettant une croissance polynomial d'ordre r, $1 \le r \le p$. On obtient à la fin du Chapitre 6 que $\mathbb{E}\, G(X, (\mu_t)_{t\in[0,T]}) \le \mathbb{E}\, G(Y, (\nu_t)_{t\in[0,T]})$, où pour tout $t \in [0,T]$, $\mu_t = \mathbb{P} \circ X_t^{-1}$, $\nu_t = \mathbb{P} \circ Y_t^{-1}$.

Le Chapitre 7 propose et analyse la méthode de particule, deux schémas basés sur la quantification et un schéma hybride particule-quantification pour l'équation de McKean-Vlasov homogène

$$\begin{cases} dX_t = b(X_t, \mu_t)dt + \sigma(X_t, \mu_t)dB_t \\ X_0 : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right) \text{ variable aléatoire} \\ \forall t \ge 0,\ \mu_t \ \text{ est la mesure de probabilité de } X_t \end{cases} . \tag{0.0.7}$$

On considère principalement le cas homogène dans ce chapitre afin d'alléger les notations mais tous les résultats peuvent se généraliser au cas non-homogène avec les méthodes classiques comme pour une équation différentielle stochastique standard. Le schéma d'Euler théorique dans le cas homogène est

$$\begin{cases} \bar{X}_{t_{m+1}} = \bar{X}_{t_m} + h \cdot b(\bar{X}_{t_m}, \bar{\mu}_{t_m}) + \sqrt{h}\, \sigma(\bar{X}_{t_m}, \bar{\mu}_{t_m}) Z_{m+1} \\ \bar{X}_0 = X_0, \ \ \bar{\mu}_{t_m} = P_{\bar{X}_{t_m}}, \end{cases} \tag{0.0.8}$$

où $M \in \mathbb{N}^*$, $h = \frac{T}{M}$, et $t_m = m \cdot h$, $m \in \{1, ..., M\}$.

La première méthode qu'on étudie est la méthode de particule, qui s'est inspirée du principe de la propagation du chaos et qui peut être considérée comme sa version discrète. Soient $\bar{X}_0^{1,N}, ..., \bar{X}_0^{N,N}$ des i.i.d variables aléatoires qui ont la même loi que X_0 dans (0.0.7). La méthode de particule est définie par

$$\begin{cases} \forall n \in \{1, ..., N\}, \\ \bar{X}_{t_{m+1}}^{n,N} = \bar{X}_{t_m}^{n,N} + hb(\bar{X}_{t_m}^{n,N}, \bar{\mu}_{t_m}^N) + \sqrt{h}\, \sigma(\bar{X}_{t_m}^{n,N}, \bar{\mu}_{t_m}^N) Z_{m+1}^n \\ \bar{\mu}_{t_m}^N := \frac{1}{N} \sum_{n=1}^N \delta_{\bar{X}_{t_m}^{n,N}} \end{cases} , \tag{0.0.9}$$

où $Z_m^n, n = 1, ..., N, m = 0, ..., M \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \mathbf{I}_q)$. L'idée de cette méthode est d'utiliser $\bar{\mu}_{t_m}^N$ comme un estimateur de $\bar{\mu}_{t_m}$ pour chaque étape d'Euler. Dans le cas de dimension 1, la vitesse de convergence de $\bar{\mu}_{t_m}^N$ à $\bar{\mu}_m$ a déjà été démontré dans Bossy and Talay (1997). Pour la vitesse de convergence dans la dimension supérieure, on obtient dans la Section 7.1 que pour toutes les dimensions d,

$$\left\| \sup_{1\le m\le M} \mathcal{W}_p(\bar{\mu}_{t_m}^N, \bar{\mu}_{t_m}) \right\|_p \le C_{d,p,L,T} \left\| \mathbb{W}_p(\bar{\mu}, \nu^N) \right\|_p ,$$

où $\bar{\mu}$ est la mesure de probabilité de $\bar{X} = (\bar{X}_t)_{t\in[0,T]}$, qui le processus défini par le schéma d'Euler continu (voir (5.2.3)) et ν^N est la mesure empirique de $\bar{\mu}$. De plus, si $\|X_0\|_{p+\varepsilon} < +\infty$ pour un $\varepsilon > 0$, on obtient dans la Section 7.1 en utilisant les résultats de Fournier and Guillin (2015) que

$$\left\| \sup_{1\le m\le M} \mathcal{W}_p(\bar{\mu}^N_{t_m}, \bar{\mu}_m) \right\|_p \le \widetilde{C} \times \begin{cases} N^{-\frac{1}{2p}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{si } p > d/2 \text{ et } \varepsilon \neq p \\ N^{-\frac{1}{2p}} \big[\log(1+N)\big]^{\frac{1}{p}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{si } p = d/2 \text{ et } \varepsilon \neq p \\ N^{-\frac{1}{d}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{si } p \in (0, d/2) \text{ et } p+\varepsilon \neq \frac{d}{(d-p)} \end{cases},$$

où $\widetilde{C}$ est une constante qui dépend de $p, \varepsilon, d, b, \sigma, L, T$.

La deuxième méthode afin de simuler l'équation de McKean-Vlasov qu'on présente dans le Chapitre 7 est la méthode de quantification optimale quadratique. Soient $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$, la grille de quantification de $\bar{X}_{t_m}$, $m = 1, ..., M$. Le schéma théorique basé sur la quantification est

$$\begin{cases} \widetilde{X}_0 = X_0, \;\; \widehat{X}_0 = \mathrm{Proj}_{x^{(0)}}(\widetilde{X}_0) \\ \widetilde{X}_{t_{m+1}} = \widehat{X}_{t_m} + h \cdot b(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) + \sqrt{h}\, \sigma(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) Z_{m+1}, \;\; m = 0, ..., M-1 \\ \text{avec } h = \frac{T}{M} \;\text{ et }\; \widehat{\mu}_{t_m} = P_{\widehat{X}_{t_m}} \\ \widehat{X}_{t_{m+1}} = \mathrm{Proj}_{x^{(m+1)}}(\widetilde{X}_{t_{m+1}}). \end{cases}$$

On montre dans la Section 7.2 l'analyse d'erreur de ce schéma et on propose trois façons différentes de simuler explicitement $\widehat{\mu}_{t_m}$.

(1) Dans le cas de Vlasov, i.e. $b(x,\mu) = \int_{\mathbb{R}^d} \beta(x,u)\mu(du)$ et $\sigma(x,\mu) = \int_{\mathbb{R}^d} a(x,u)\mu(du)$, on peut utiliser la méthode de quantification récursive, qui a été introduite dans Pagès and Sagna (2015) pour une équation stochastique régulière. On peut en déduire une transition markovienne de $(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})$. Soient $p^{(m)} = (p_1^{(m)}, ..., p_K^{(m)})$ le poids qui correspondent à $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$, $m = 0, ..., M$ et par conséquent $\widehat{\mu}_{t_m} = \sum_{k=1}^K \delta_{x_k^{(m)}} p_k^{(m)}$. La transition markovienne de $(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})$ qu'on obtient dans la Section 7.3 est

$$\begin{aligned} &\mathbb{P}\big(\widehat{X}_{t_{m+1}} = x_j^{(m+1)} \mid \widehat{X}_{t_m} = x_i^{(m)}, p^{(m)}\big) \\ &= \mathbb{P}\Big[\big(x_i^{(m)} + h\sum_{k=1}^K p_k^{(m)} \beta(x_i^{(m)}, x_k^{(m)}) + \sqrt{h}\sum_{k=1}^K p_k^{(m)} a(x_i^{(m)}, x_k^{(m)}) Z_{m+1}\big) \in C_j(x^{(m+1)})\Big] \end{aligned}$$

et étant donné $p^{(m)}$, on peut calculer pour tout $j = 1, ..., K$,

$$\begin{aligned} p_j^{(m+1)} &= \mathbb{P}\big(\widehat{X}_{t_{m+1}} = x_j^{(m+1)} \mid p^{(m)}\big) \\ &= \sum_{i=1}^{K} \mathbb{P}\big(\widehat{X}_{t_{m+1}} = x_j^{(m+1)} \mid \widehat{X}_{t_m} = x_i^{(m)}, p^{(m)}\big) \cdot \mathbb{P}(\widehat{X}_{t_m} = x_i^{(m)}). \end{aligned}$$

La preuve de ces transitions markoviennes se trouve dans la Section 7.3. De plus, on explique également dans cette section comment utiliser l'algorithme de Lloyd afin d'améliorer l'exactitude de la simulation.

(2) La deuxième façon d'exprimer explicitement $\widehat{\mu}_{t_m}$ est d'utiliser la grille optimale de la distribution normale $\mathcal{N}(0, \mathbf{I}_q)$ et son poids, qui peuvent être téléchargées dans le site `www.quantize.maths-fi.com/gaussian_database` pour les dimension $q = 1, ..., 10$. Soient $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$ une grille de quantification de $\bar{X}_{t_m}$, $m = 0, ..., M$. Soit $z = (z_1, ..., z_J)$ une grille optimale de $\mathcal{N}(0, \mathbf{I}_q)$ avec $J > K$ et soit $w = (w_1, ..., w_J)$ le poids correspondant de z. Le schéma basé sur les deux grilles de quantification x et z est comme suit

$$\begin{cases} \widetilde{X}_0 = X_0, \quad \widehat{X}_0 = \mathrm{Proj}_{x^{(0)}}(\widetilde{X}_0) \\ \widetilde{X}_{t_{m+1}} = \widehat{X}_{t_m} + h \cdot b(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) + \sqrt{h}\,\sigma(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})\widehat{Z}_{m+1}, \quad m = 0, ..., M-1 \\ \text{où } h = \frac{T}{M} \text{ et } \widehat{\mu}_{t_m} = P_{\widehat{X}_{t_m}} \\ \widehat{X}_{t_{m+1}} = \mathrm{Proj}_{x^{(m+1)}}(\widetilde{X}_{t_{m+1}}), \end{cases},$$

où $\widehat{Z}_m \overset{i.i.d}{\sim} \sum_{j=1}^{J} w_j \delta_{z_j}$. On appelle cette méthode le schéma double-quantisé et on montre dans la Section 7.4 l'analyse de l'erreur de ce schéma.

(3) Soient $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$, $m = 0, 1, ..., M$ une suite de grilles de quantification. Après avoir obtenu la vitesse de convergence de la méthode de particule, on peut aussi appliquer la méthode de quantification optimale sur (0.0.9) comme suit

$$\begin{cases} \forall n \in \{1, ..., N\}, \\ \widetilde{X}_{t_{m+1}}^{n,N} = \widetilde{X}_{t_m}^{n,N} + h \cdot b(\widetilde{X}_{t_m}^{n,N}, \widehat{\mu}_{t_m}^K) + \sqrt{h}\,\sigma(\widetilde{X}_{t_m}^{n,N}, \widehat{\mu}_{t_m}^K) Z_{m+1}^n \\ \widehat{\mu}_{t_m}^K = \big(\frac{1}{N}\sum_{n=1}^{N} \delta_{\widetilde{X}_{t_m}^{n,N}}\big) \circ \mathrm{Proj}_{x^{(m)}}^{-1} = \sum_{k=1}^{K} \big[\delta_{x_k^{(m)}} \cdot \sum_{n=1}^{N} \mathbb{1}_{V_k(x^{(m)})}(\widetilde{X}_{t_m}^{n,N})\big] \\ \bar{X}_0^{n,N} \overset{i.i.d}{\sim} X_0, \quad Z_m^n \overset{i.i.d}{\sim} \mathcal{N}(0, I_q) \end{cases}.$$

On appelle ce schéma le schéma hybride particule-quantification (schéma hybride, à court terme). L'analyse d'erreur de ce schéma se trouve dans la Section 7.5.

À la fin du Chapitre 7, on montre des simulations par les méthodes présentées précédemment à travers deux exemples. Le premier exemple est la simulation de l'équation de Burgers introduite dans Sznitman (1991) et Bossy and Talay (1997).

L'équation de Burgers admet d'une solution explicite, on peut donc comparer le niveau de précision des méthodes différents. Le deuxième exemple, le réseau de neurones de FitzHugh-Nagumo, est en dimension 3 et introduit premièrement dans Baladron et al. (2012) et également dans Reis et al. (2018).

Chapter 1

Introduction

1.1 General background on optimal quantization

Vector quantization was originally developed as an optimal discretization method for the signal transmission and compression by the Bell laboratories in the 1950s. Many seminal and historical contributions on vector quantization and its connections with information theory were gathered and published later in IEEE Transactions on Information Theory (1982). Nowadays, vector quantization becomes an efficient tool widely used in different fields. For example, in unsupervised learning, vector quantization has a close connection with the clustering analysis and the pattern recognition; in numerical probability, vector quantization is used for numerical integration, conditional expectation computation, simulation of stochastic differential equations and also for option pricing in financial mathematics. Among a wide range of properties and applications of the quantization method, this thesis focuses on two limit theorems of the optimal quantization theory and its application to the simulation of the McKean-Vlasov equation.

1.1.1 Principle of optimal quantization[(1)]

Let X be an $\mathbb{R}^d$-valued random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with probability distribution μ having a p-th finite moment, $p \geq 1$. Let $|\cdot|$ denote the norm on $\mathbb{R}^d$. The quantization method consists in discretely estimating μ (or X) by using a K-tuple $x = (x_1, ..., x_K) \in (\mathbb{R}^d)^K$ and its weight $w = (w_1, ..., w_K)$. Here the K-tuple $x = (x_1, ..., x_K)$ is called by a *quantizer* (or *quantization grid*, *cluster center*, *codebook* in the literature). To be more precise, the quantized estimator of μ induced by x, denoted

(1) We allow ourselves a slight relaxation of mathematical rigour (only) in this section to quickly present the basic principles of optimal quantization.

by $\widehat{\mu}^x$, is defined by

$$\widehat{\mu}^x \coloneqq \sum_{k=1}^{K} \underbrace{\mu\big(C_k(x)\big)}_{=:w_k,\text{ the weight of each quantizer point } x_k} \cdot \delta_{x_k}, \tag{1.1.1}$$

where δ_a denote the Dirac mass at a and $\big(C_k(x)\big)_{k=1,\dots,K}$ is the Voronoï partition (see for example Figure 1.1) generated by x, which is a mesurable partition of $\mathbb{R}^d$ satisfying

$$\forall k \in \{1,...,K\}, \quad C_k(x) \subset \Big\{ y \in \mathbb{R}^d \;\Big|\; |y - x_k| = \min_{1 \le j \le K} |y - x_j| \Big\}.$$

Similarly, the estimator of X by the quantization method is defined by

$$\widehat{X}^x \coloneqq \mathrm{Proj}_x(X) \coloneqq \sum_{k=1}^{K} x_k \mathbb{1}_{C_k(x)}(X). \tag{1.1.2}$$

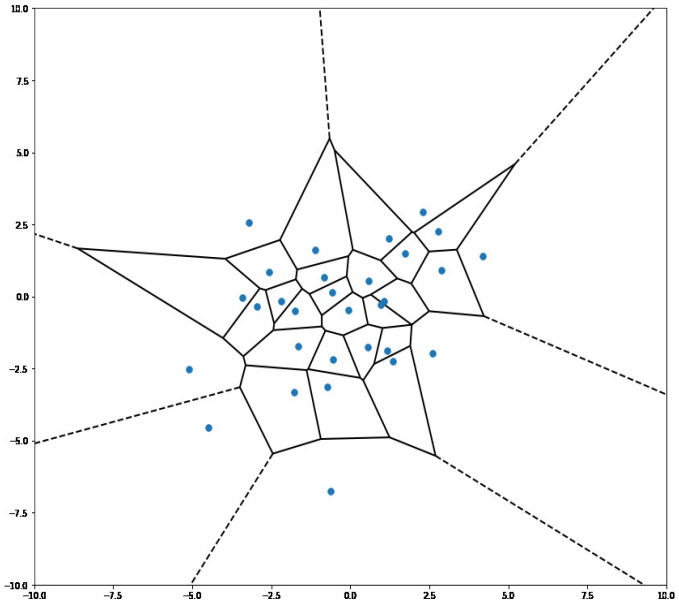


Figure 1.1 An example of the Voronoï diagram on $\mathbb{R}^2$ equipped with the Euclidean norm

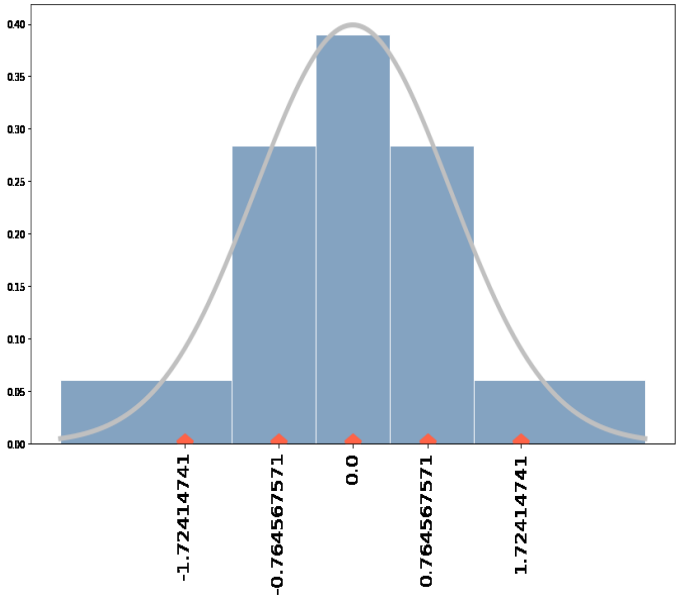


Figure 1.2 An optimal quantizer of $\mathcal{N}(0,1)$ (red point). The vertical axis is the weight divided by the length of the corresponding Voronoï cell.

Let $d(\xi, A) := \min_{a\in A} |\xi - a|$ define the distance between a point $\xi \in \mathbb{R}^d$ and a set $A \subset \mathbb{R}^d$. For $p \geq 1$, the *L^p-quantization error* of the quantizer $x = (x_1, ..., x_K) \in (\mathbb{R}^d)^K$ is defined by the L^p-norm of $d(X, \Gamma^x)$ with $\Gamma^x := \{x_1, ..., x_K\} \subset \mathbb{R}^d$, namely,

$$e_{K,p}(\mu, x) := \left[\mathbb{E}\, d(X, \Gamma^x)^p\right]^{1/p} = \left[\int_{\mathbb{R}^d} \min_{1\leq k\leq K} |\xi - x_k|^p \, \mu(d\xi)\right]^{\frac{1}{p}}.$$

A quantizer $x^* \in (\mathbb{R}^d)^K$ satisfying $e_{K,p}(\mu, x^*) = \inf_{x\in(\mathbb{R}^d)^K} e_{K,p}(\mu, x)$ is called an *L^p-optimal quantizer* of μ at level K. Such a quantizer always exists if μ has a finit p-th moment (see Graf and Luschgy (2000)[Theorem 4.12]).

In the quadratic case ($p = 2$), the optimal quantizer can be numerically computed by using the CLVQ algorithm (stochastic gradient algorithm), the Lloyd I algorithm (randomized or deterministic fixed point algorithm) or some variants. Figure 1.2 shows a quadratic optimal quantizer at level 5

$$x^* = (-1.72414741,\ -0.764567571,\ 0.0,\ 0.764567571,\ 1.72414741)$$

of the normal distribution $\mathcal{N}(0,1)$ on $\mathbb{R}$ computed by the Lloyd I algorithm.

Another classical method to discretely approximate a probability measure μ is the Monte-Carlo method. Let $X_1, ..., X_N$ be an i.i.d sample defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with

probability distribution μ. The estimator of μ by the Monte-Carlo method is

$$\bar{\mu}^{N,\omega} := \frac{1}{N}\sum_{n=1}^{N}\delta_{X_n(\omega)}. \tag{1.1.3}$$

Compared with the Monte-Carlo method, the optimal quantization method has two intuitional advantages

- The optimal quantizer is deterministic, which means the optimal quantizer does not depend on ω in $(\Omega, \mathcal{F}, \mathbb{P})$, so that the estimator $\widehat{\mu}^{x^*}$ defined in (1.1.1) is also deterministic. This means one can achieve a prescribed level of accuracy by enlarging the size of optimal quantizer with the help of an upper bound of the optimal error (see further the non-asymptotic Zador's theorem in Theorem 1.1.1).
- If we consider a K-level optimal quantizer $x^* = (x_1, ..., x_K)$ and an i.i.d sample $X_1, ..., X_K$ with the same size K, we will always get a higher accuracy with respect to the Wasserstein distance by using the quantization estimator $\widehat{\mu}^{x^*}$ defined in (1.1.1) than using the Monte-Carlo estimator $\bar{\mu}^{K,\omega}$ defined in (1.1.3).

However, the shortcoming of the optimal quantization method often occurs on the computing time due to the adding procedure to find the optimal quantizer.

The first advantage is obvious. Here we give a quick explanation to the second advantage. Let $\mathcal{P}(K)$ denote the set of all discrete probabilities ν on $\mathbb{R}^d$ with $\mathrm{Card}\big(\mathrm{supp}(\nu)\big) \leq K$. Let $x^* = (x_1^*, ..., x_K^*)$ denote an L^p-optimal quantizer of $\mu \in \mathcal{P}_p(\mathbb{R}^d)$. It follows from Graf and Luschgy (2000)[Lemma 3.4] that

$$e_{K,p}(\mu, x^*) = \mathcal{W}_p(\mu, \widehat{\mu}^{x^*}) = \inf_{\nu\in\mathcal{P}(K)} \mathcal{W}_p(\mu, \nu).$$

Thus for any K-size i.i.d sample $X_1, ..., X_K$ with probability distribution μ, we have

$$\mathcal{W}_p(\mu, \widehat{\mu}^{x^*}) \leq \mathcal{W}_p(\mu, \bar{\mu}^{K,\omega}) \quad \text{a.s.},$$

where $\widehat{\mu}^{x^*}$ is defined in (1.1.1) and $\bar{\mu}^{K,\omega}$ is defined in (1.1.3).

The optimal quantization method is applied in the following fields, besides the signal transmission and compression as its original purpose.

– In the numerical probability, the optimal quantization is used to compute the numerical integration, conditional expectation (see e.g. Pagès (1998)) and offers a spatial discretization in the simulation of stochastic differential equation (see e.g. Gobet et al. (2006)). Let X be an $\mathbb{R}^d$-valued random variable with probability

distribution μ having a p-th finite moment and let $x = (x_1, ..., x_K)$ be its quantizer. A simple example is that for a Lipschitz continuous function $F : (E, \mathcal{E}) \to \big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ with Lipschitz constant $[F]_{\text{Lip}}$, one can use

$$\mathbb{E}\, F(\widehat{X}^x) = \sum_{k=1}^{K} F(x_k)\mu\big(C_k(x)\big)$$

to approximate $\mathbb{E}\, F(X)$. Note that $\left\|X - \widehat{X}^x\right\|_p = e_{K,p}(\mu, x)$, so the (strong) error of the above simulation can be upper-bounded by

$$\mathbb{E}\left|F(\widehat{X}^x) - F(X)\right| \le [F]_{\text{Lip}} \left\|X - \widehat{X}^x\right\|_1 \le [F]_{\text{Lip}} \left\|X - \widehat{X}^x\right\|_p, \quad p \ge 1.$$

If F is differentiable with a Lipschitz continuous gradient ∇F, then (see Pagès (1998) or Pagès (2018)[Proposition 5.2])

$$\left|\mathbb{E}\, F(X) - \mathbb{E}\, F(\hat{X}^x)\right| \le \frac{1}{2}[\nabla F]_{\text{Lip}} \left\|X - \widehat{X}^x\right\|_2^2.$$

– In the field of the unsupervised learning, the optimal quantization is also called the *K-means* clustering. It is used to solve the problem of automatic classification. In this context, the quantizer is also called *cluster center* in the literature. The main idea is to consider a vector data set $\{y_1, ..., y_N\}$ as a empirical measure $\frac{1}{N}\sum_{n=1}^{N} \delta_{y_n}$ and to compute/train the optimal quantizer of this data set. The following figure shows an example of the optimal quantization of a data set.

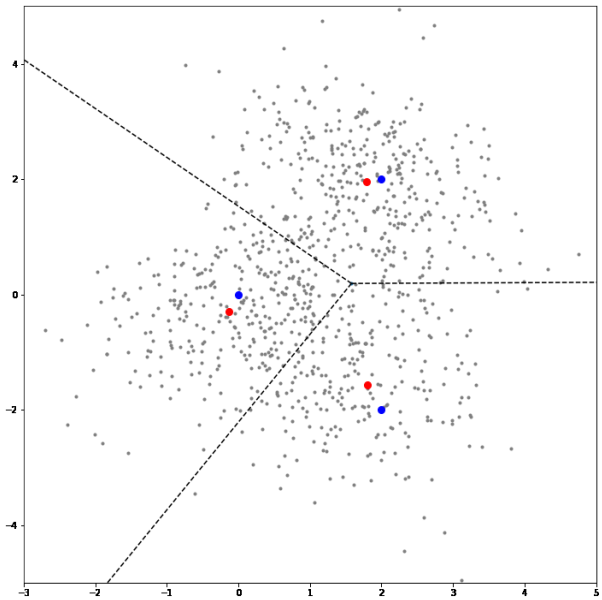


Figure 1.3 The optimal quantizer of a data set.

The data set in Figure 1.3 is a mix of three i.i.d sample of size $N = 300$ of respective probability distribution $\mathcal{N}\left(\left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right]\right)$, $\mathcal{N}\left(\left[\begin{smallmatrix} 2 \\ 2 \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix}\right]\right)$, and $\mathcal{N}(\left[\begin{smallmatrix} 1.5 \\ -1.5 \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 & -0.2 \\ -0.2 & 1 \end{smallmatrix}\right])$. The red points are an optimal quantizer of size 3 of this data set. The blue points are the three centers of the normal distributions.

1.1.2 Frequently used definitions and basic properties

Now we present several frequently used definitions to mathematically formalize the optimal quantization and some of its basic properties. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space and let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, |\cdot|_E)$ be a random variable valued in a separable Banach space E with norm $|\cdot|_E$. Let

$$\mathcal{P}_p(E) \coloneqq \left\{\mu \text{ probability distribution on } E \text{ s.t. } \int_E |\xi|_E^p \, \mu(d\xi) < +\infty\right\}$$

and let $\mathcal{W}_p$ denote the *L^p-Wasserstein distance* on $\mathcal{P}_p(E)$, defined by

$$\mathcal{W}_p(\mu, \nu) \coloneqq \Big(\inf_{\pi \in \Pi(\mu,\nu)} \int_{E\times E} |x - y|_E^p \, \pi(dx, dy)\Big)^{\frac{1}{p}}$$
$$= \inf\left\{\Big[\mathbb{E}\,|X - Y|_E^p\Big]^{\frac{1}{p}},\ X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, |\cdot|_E) \text{ with } \mathbb{P}\circ X^{-1} = \mu, \mathbb{P}\circ Y^{-1} = \nu\right\},$$

where in the first line of the above definition, $\Pi(\mu, \nu)$ denotes the set of all probability measures on $(E^2, \mathcal{E}^{\otimes 2})$ with respective marginals μ and ν and $\mathcal{E}$ denotes the σ-algebra generated by $|\cdot|_E$. For two random variables $X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, |\cdot|_E)$ with respective probability distributions μ and ν, we write $\mathcal{W}_p(X, Y) \coloneqq \mathcal{W}_p(\mu, \nu)$.

Let μ denote the probability distribution of X and assume that $\mu \in \mathcal{P}_p(E)$. The *quantizer* (also called *codebook* in signal compression or *cluster center* in unsupervised learning theory) is originally denoted by a finite point set $\Gamma = \{x_1, ..., x_K\} \subset E$. The L^p-mean quantization error of Γ, which describes the accuracy of representing the probability measure μ by Γ, is defined by

$$e_p(\mu, \Gamma) \coloneqq \|d(X, \Gamma)\|_p = \Big[\int_E \min_{a\in\Gamma} |\xi - a|_E^p \, \mu(d\xi)\Big]^{\frac{1}{p}},$$

where $d(\xi, A) = \min_{a\in A} |\xi - a|_E$ defines the distance between a point $\xi \in E$ and a set $A \subset E$. A quantizer $\Gamma^{*,(K)}$ satisfying

$$e_p(\mu, \Gamma^{*,(K)}) = \inf_{\substack{\Gamma\subset E,\\ \mathrm{card}(\Gamma)\le K}} \Big[\mathbb{E}\, d(X, \Gamma)^p\Big]^{\frac{1}{p}} = \inf_{\substack{\Gamma\subset E,\\ \mathrm{card}(\Gamma)\le K}} \Big[\int_E \min_{a\in\Gamma} |\xi - a|_E^p \, \mu(d\xi)\Big]^{\frac{1}{p}} \tag{1.1.4}$$

is called an *L^p-optimal quantizer* (or *optimal quantizer* in short) at *level K*. Such an opti-

mal L^p-quantizer always exists when $X \in L^p(\mathbb{P})$ (see Graf and Luschgy (2000)[Theorem 4.12] and Graf et al. (2007)).

Now we define the *L^p-mean quantization error function* and the *L^p-distortion function.*

Definition 1.1.1 (Quantization error function and distortion function)**.** *Let $\mu \in \mathcal{P}_p(E)$, $p \in [1, +\infty)$. The L^p-mean quantization error function of μ at level K, denoted by $e_{K,p}(\mu, \cdot)$, is defined by:*

$$\begin{array}{rccc} e_{K,p}(\mu,\cdot): & E^K & \longrightarrow & \mathbb{R}_+ \\ & x=(x_1,\ldots,x_K) & \longmapsto & e_{K,p}(\mu,x)=\Big[\displaystyle\int_{\mathbb{R}^d}\min_{1\le i\le K}|\xi-x_i|_E^p\,\mu(d\xi)\Big]^{\frac{1}{p}}. \end{array} \tag{1.1.5}$$

Moreover, the L^p-distortion function of μ at level K is defined by $\mathcal{D}_{K,p}(\mu,\cdot) := e^p_{K,p}(\mu,\cdot)$.

When $p = 2$, E is a Hilbert space and $|\cdot|_E$ is induced by an inner product, we call $e_{K,2}(\mu,\cdot)$ the *quadratic* quantization error function and $\mathcal{D}_{K,2}(\mu,\cdot)$ the *quadratic* distortion function. In this case, we remove sometimes the subscript 2.

Let $\mathrm{card}(\Gamma)$ denote the cardinality of the point set $\Gamma \subset E$. The generic variable of the function $e_{K,p}(\mu,\cdot)$ and $\mathcal{D}_{K,p}(\mu,\cdot)$ is a priori a K-tuple in E^K. However, for a finite quantizer $\Gamma \subset E$, if the level $K \geq \mathrm{card}(\Gamma)$, then for any K-tuple $x^\Gamma = (x_1^\Gamma, \ldots, x_K^\Gamma) \in E^K$ such that $\Gamma = \{x_1^\Gamma, \ldots, x_K^\Gamma\}$, we have $e_p(\mu,\Gamma) = e_{K,p}(\mu, x^\Gamma)$. For example,

$$e_p\big(\mu,\{x_1,x_2\}\big) = e_{2,p}\big(\mu,(x_1,x_2)\big) = e_{3,p}\big(\mu,(x_1,x_1,x_2)\big), \text{ etc.}$$

Note that $e_{K,p}(\mu,\cdot)$ and $\mathcal{D}_{K,p}(\mu,\cdot)$ are symmetric functions on E^K and that, owing to the above definition,

$$\inf_{\Gamma\subset E,\mathrm{card}(\Gamma)\le K} e_p(\mu,\Gamma) = \inf_{x\in E^K} e_{K,p}(\mu,x). \tag{1.1.6}$$

Therefore, with a slight abuse of notation, we will use for convenience either a K-tuple $x = (x_1, \ldots, x_K) \in E^K$ or a point set $\Gamma = \{x_1, ..., x_K\} \subset E$ to represent a quantizer and we will denote by $x^* \in \mathrm{argmin} e_{K,p}(\mu,\cdot)$ the L^p-optimal quantizer of μ at level K. Furthermore, we denote

$$e^*_{K,p}(\mu) := \inf_{y=(y_1,\ldots,y_K)\in E^K}\Big[\int_{\mathbb{R}^d}\min_{1\le i\le K}|\xi-y_i|^2\,\mu(d\xi)\Big]^{\frac{1}{2}}. \tag{1.1.7}$$

the *L^p-optimal quantization error* of μ at level K.

There exist other terminologies in the literature which play a similar role as the quantizer. For example, in Graf and Luschgy (2000)[Section 3], the authors define the

quantizer by an application $f_K : E \to E$ such that $\text{card}\big(\text{supp}(f_K)\big) \leq K$, where supp denotes the support of a function and card denotes the cardinality of a set. Another example is in Pollard (1982b). The author uses a probability distribution μ_K on E, where the subscript K means $\text{card}\big(\text{supp}(\mu_K)\big) \leq K$, to represent the quantizer. The equivalence of these two representations and our definition has been proved in Pollard (1982b)[Theorem 3] and Graf and Luschgy (2000)[Lemma 3.1, 3.4 and 4.4].

Quantization theory has a close connection with Voronoï partitions. Let $x = (x_1, ..., x_K) \in E^K$ be a quantizer at level K, where $x_i \neq x_j$ if $i \neq j$. The *Voronoï cell* (or *Voronoï region*) generated by x_k is defined by

$$V_{x_k}(x) = \big\{\xi \in E : |\xi - x_k|_E = \min_{1\leq j\leq K} |\xi - x_j|_E \big\} \tag{1.1.8}$$

and $\big(V_{x_k}(x)\big)_{1\leq k\leq K}$ is called the *Voronoï diagram* of x. On a Hilbert or a Euclidean space, the Voronoï cells are intersections of half-spaces defined by the median hyperplanes, i.e.

$$V_{x_k}(x) = \cap_{j\neq k} E_{kj},$$

where E_{kj} is the half-space defined by the median hyperplane of x_k and x_j that contains x_k.

A mesurable partition $\big(C_{x_k}(x)\big)_{1\leq k\leq K}$ is called a *Voronoï partition* of E induced by x if

$$\forall k \in \{1, ..., K\}, \qquad C_{x_k}(x) \subset V_{x_k}(x). \tag{1.1.9}$$

When there is no ambiguity, we write $C_k(x)$ and $V_k(x)$ instead of $C_{x_k}(x)$ and $V_{x_k}(x)$. We also define the *open Voronoï cell* generated by x_k by

$$V^o_{x_k}(x) = \big\{\xi \in E : |\xi - x_k|_E < \min_{1\leq j\leq K, j\neq k} |\xi - x_j|_E \big\}. \tag{1.1.10}$$

One quantizer $x = (x_1, ..., x_K)$ may generate different Voronoï partitions, this depends on the choice between $V^o_{x_i}(x)$ and $V^o_{x_j}(x)$ with which we put together $V_{x_i}(x) \cap V_{x_j}(x)$. Figure 1.2 in Graf and Luschgy (2000) emphasizes that when the norm is not Euclidean then $\text{int}V_{x_i}(x)$ and $V^o_{x_i}(x)$ may be different. However, on a Hilbert or a Euclidean space, there is always equality.

Based on a Voronoi partition $\big(C_{x_k}(x)\big)_{1\leq k\leq K}$, one can rewrite the L^p-distortion function $\mathcal{D}_{K,p}(\mu, \cdot)$ (also the quantization error function) by

$$\mathcal{D}_{K,p}(\mu, x) = \sum_{k=1}^{K} \int_{C_{x_k}(x)} |\xi - x_k|^p \, \mu(d\xi), \tag{1.1.11}$$

but the value of $\mathcal{D}_{K,p}(\mu, x)$ is independent of the choice of Voronoi partition. For the

properties of Voronoï cell, we refer to Graf and Luschgy (2000)[Chapter I] among many other references.

In fact, both the definition of Voronoï region and the quantization error function strongly depend on the chosen norm on E. For example, Figure 1.1 in Graf and Luschgy (2000) shows two different Voronoï diagrams of the same finite point set in $\mathbb{R}^2$ with respect to l_1-norm and l_2-norm. When $E = \mathbb{R}^d$ and the underlying norm is strictly convex or l_p-norm with $1 \leq p \leq +\infty$, we have $\lambda_d\big(\partial V_{x_k}(x)\big) = 0$, where λ_d denotes the Lebesgue measure on $\mathbb{R}^d$ and ∂A denotes the boundary of A (see Graf and Luschgy (2000)[Theorem 1.5]). In particular, if $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and x^* is a quadratic optimal quantizer of μ at level K with respect to the Euclidean norm, even if μ is not absolutely continuous with respect to λ_d, we have $\mu\big(\partial V_{x_k}(x^*)\big) = 0$ for all $k \in \{1, ..., K\}$ (see Graf and Luschgy (2000)[Theorem 4.2]).

Furthermore, based on a Voronoi partition $\big(C_{x_k}(x)\big)_{1\leq k\leq K}$ generated by a quantizer $x = (x_1, ..., x_K)$ satisfying $x_i \neq x_j, i \neq j$, we can define a projection function $\mathrm{Proj}_x : E \to \{x_1, ..., x_K\}$ by

$$\xi \in E \mapsto \mathrm{Proj}_x(\xi) \coloneqq \sum_{k=1}^{K} x_k \mathbb{1}_{C_{x_k}(x)}(\xi). \tag{1.1.12}$$

Thus, for a random variable X with probability distribution μ, we define

$$\widehat{X}^x \coloneqq \mathrm{Proj}_x(X). \tag{1.1.13}$$

Then $\left\|\widehat{X}^x - X\right\|_p = e_{K,p}(\mu, x)$. When there is no ambiguity, we denote by $\widehat{X}$ instead of $\widehat{X}^x$. The variable $\widehat{X}^x$ and its probability distribution

$$\widehat{\mu}^x = \sum_{k=1}^{K} \delta_{x_k} \mu\big(C_{x_k}(x)\big) \tag{1.1.14}$$

are often considered as quantization based estimators of X and μ. Moreover, it follows from Graf and Luschgy (2000)[Lemma 3.4](1) that

$$\mathcal{W}_p(\widehat{\mu}^x, \mu) = \left\|\widehat{X}^x - X\right\|_p = e_{K,p}(\mu, x). \tag{1.1.15}$$

(1) The statement of Graf and Luschgy (2000)[Lemma 3.4] is established for the optimal quantizer. However, the third inequality of its proof is also valid for an arbitrary quantizer from where we derive (1.1.15).

1.1.3 A brief review of the literature and motivations

Most work in the field of the optimal quantization addresses the following three questions around which we organize this section:

- *Question 1: Why does the optimal quantization provide a good discrete representation of the probability distribution?*
- *Question 2: How to find the (quadratic) optimal quantizer?*
- *Question 3: How to apply the optimal quantization in numerical probability or in unsupervised learning?*

Moreover, for a first mathematically rigorous monograph of various aspects of vector quantization theory, we refer to Graf and Luschgy (2000) (and the references therein). See also Pagès (2015) for numerical applications. For more engineering applications to signal compression, see e.g. Gersho and Gray (2012) among an extensive literature.

1.1.3.1 Why does the optimal quantization provide a *good* representation of the probability distribution?

We start with some basic properties of the optimal quantizer and the optimal quantization error to answer *Question* 1. First, the existence of optimal quantizer is proved in Pagès (1998) and Graf and Luschgy (2000)[Theorem 4.12] for $E = \mathbb{R}^d$ and in Graf et al. (2007) for any Banach space. Generally, there does not exist a unique optimal quantizer for a probability distribution μ. If $x^* = (x_1, ..., x_K)$ is an optimal quantizer of μ, it is obvious that any permutation of $x_1, ..., x_K$ such as $x' = (x_K, ..., x_1)$ is also an optimal quantizer of μ. However, if $E = \mathbb{R}$ and we set an order for $x = (x_1, ..., x_K)$ by letting $x_1 \leq x_2 \leq ... \leq x_K$, the uniqueness of optimal quantizer is proved in Kieffer (1983) if μ is absolutly continuous with respect to the Lebesgue measure λ and has a log-concave density function.

Moreover, the optimal quantizer and the optimal quantization error provide the following properties for a fixed quantization level $K \in \mathbb{N}^*$. We present later their asymptotic properties when the quantization level $K \to +\infty$.

Theorem 1.1.1. *[Properties of the optimal quantization error]*

(i) (Strictly decreasing of $K \mapsto e^*_{K,p}(\mu)$)
For every $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ with $\mathrm{card}\big(\mathrm{supp}(\mu)\big) \geq K$, one has $e^*_{K,p}(\mu) < e^*_{K-1,p}(\mu)$, for $K \geq 2$.

(ii) (Upper bound of the optimal quantization error: Non-asymptotic Zador's theorem) Let $\eta > 0$. For every $\mu \in \mathcal{P}_{p+\eta}(\mathbb{R}^d)$ and for every quantization level K, there exists a constant $C_{d,p,\eta} \in (0, +\infty)$ which depends only on d, p and η such that

$$e^*_{K,p}(\mu) \leq C_{d,p,\eta} \cdot \sigma_{p+\eta}(\mu) K^{-1/d}, \tag{1.1.16}$$

where for $r \in (0, +\infty)$, $\sigma_r(\mu) = \min_{a \in \mathbb{R}^d} \left[\int_{\mathbb{R}^d} |\xi - a|^r \, \mu(d\xi) \right]^{1/r}$.

Theorem 1.1.2. *[Properties of optimal quantizers]*

(i) (Boundedness and cardinality of optimal quantizers)
Let $\mu \in \mathcal{P}_p(\mathbb{R}^d)$. Assume that $\operatorname{card}\big(\operatorname{supp}(\mu)\big) \geq K$. Let

$$\mathcal{G}_K(\mu) \coloneqq \Big\{ \{x^*_1, ..., x^*_K\} \;\Big|\; (x^*_1, ..., x^*_K) \in \operatorname{argmin} e_{K,p}(\mu, \cdot) \Big\}$$

contains the points which compose an L^p-optimal quantizer of μ at level K. Then $\mathcal{G}_K(\mu)$ is a nonempty compact set so that

$$\rho_{K,p}(\mu) \coloneqq \max \Big\{ \max_{1 \leq k \leq K} |x^*_k| \,,\; (x^*_1, ..., x^*_K) \text{ is an optimal quantizer of } \mu \Big\} \tag{1.1.17}$$

is finite for a fixed level K. Moreover, if $\Gamma^* \subset \mathbb{R}^d$ is an L^p-optimal quantizer of μ, then $\operatorname{card}(\Gamma^*) = K$. In particular, if $\Gamma^* = \{x_1, ..., x_K\}$, then $x^{\Gamma^*} \coloneqq (x_1, ..., x_K) \in \operatorname{argmin} e_{K,p}(\mu, \cdot)$ and vice versa.

(ii) (Stationary of quadratic optimal quantizers)
Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big)$ be a random variable with probability distribution $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\operatorname{card}(\operatorname{supp}(\mu)) \geq K$. If the norm on $\mathbb{R}^d$ is the Euclidean norm, then any quadratic optimal quantizer $x^* = (x^*_1, ..., x^*_K)$ of level K is stationary in the sense that

$$\mathbb{E}\left[X \mid \widehat{X}^{x^*} \right] = \widehat{X}^{x^*}, \tag{1.1.18}$$

where $\widehat{X}^{x^*}$ is defined in (1.1.13) and the equality of (1.1.18) is valid for every Voronoï partition generated by x^*.

We refer to Graf and Luschgy (2000)[Theorem 4.12] for the proof of Theorem 1.1.1-(i) and Theorem 1.1.2-(i), to Luschgy and Pagès (2008) and Pagès (2018)[Theorem 5.2] for the proof of Theorem 1.1.1-(ii) and to Pagès (2008) and Pagès (2018)[Proposition 5.1] for the proof of Theorem 1.1.2-(ii).

The quantization error function $e_{K,p}(\mu, \cdot)$ and the distortion function $\mathcal{D}_{K,p}(\mu, \cdot)$ are two efficient tools to study the optimal quantization as the optimal quantizer $x^* = (x^*_1, ..., x^*_K) \in \operatorname{argmin} e_{K,p}(\mu, \cdot) = \operatorname{argmin} \mathcal{D}_{K,p}(\mu, \cdot)$. In fact, the quantization error function is entirely characterized by the targeted probability distribution μ in

the following sense. Let E denote a separable Banach space equipped with a norm $|\cdot|$. Let $X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, |\cdot|)$ be two random variables with respective probability distributions $\mu, \nu \in \mathcal{P}_p(E)$. For every K-tuple $x = (x_1, ..., x_K) \in E^K$, we have

$$\begin{aligned}
\big|e_{K,p}(\mu, x) - e_{K,p}(\nu, x)\big| &= \Big| \ \Big\| \min_{i=1,...,K} |X - x_i| \Big\|_p - \Big\| \min_{i=1,...,K} |Y - x_i| \Big\|_p \ \Big| \\
&\le \Big\| \min_{i=1,...,K} |X - x_i| - \min_{i=1,...,K} |Y - x_i| \Big\|_p \text{ (by the Minkowski inequality)} \\
&\le \Big\| \max_{i=1,...,K} | \ |X - x_i| - |Y - x_i| \ | \Big\|_p \le \|X - Y\|_p \,. \qquad (1.1.19)
\end{aligned}$$

As this inequality holds for every couple (X, Y) with marginal distributions μ and ν, it follows that for every level $K \ge 1$,

$$\|e_{K,p}(\mu, \cdot) - e_{K,p}(\nu, \cdot)\|_{\sup} := \sup_{x \in E^K} |e_{K,p}(\mu, x) - e_{K,p}(\nu, x)| \le \mathcal{W}_p(\mu, \nu). \qquad (1.1.20)$$

Hence, if $(\mu_n)_{n \ge 1}$ is a sequence in $\mathcal{P}_p(E)$ converging for the $\mathcal{W}_p$-distance to $\mu_\infty \in \mathcal{P}_p(E)$, then

$$\|e_{K,p}(\mu_n, \cdot) - e_{K,p}(\mu_\infty, \cdot)\|_{\sup} \le \mathcal{W}_p(\mu_n, \mu_\infty) \xrightarrow{n \to +\infty} 0. \qquad (1.1.21)$$

Moreover, for any $\mu \in \mathcal{P}_p(E)$, the function $e_{K,p}(\mu, \cdot)$ defined in (1.1.5) is 1-Lipschitz continuous for every $K \ge 1$ since for any $x = (x_1, \dots, x_K), y = (y_1, \dots, y_K) \in E^K$,

$$\begin{aligned}
|e_{K,p}(\mu, x) - e_{K,p}(\mu, y)| &= \Big| \Big[\int_E \min_{1 \le i \le K} |\xi - x_i|^p \, \mu(d\xi) \Big]^{\frac{1}{p}} - \Big[\int_E \min_{1 \le j \le K} |\xi - y_j|^p \, \mu(d\xi) \Big]^{\frac{1}{p}} \Big| \\
&\le \Big[\int_E \Big| \min_{1 \le i \le K} |\xi - x_i| - \min_{1 \le j \le K} |\xi - y_j| \Big|^p \mu(d\xi) \Big]^{\frac{1}{p}} \quad \text{(by the Minkowski inequality)} \\
&\le \Big[\int_E \max_{1 \le i \le K} |x_i - y_i|^p \, \mu(d\xi) \Big]^{\frac{1}{p}} = \max_{1 \le i \le K} |x_i - y_i| \,. \qquad (1.1.22)
\end{aligned}$$

Now we show the asymptotic properties of the optimal quantization on $\mathbb{R}^d$ when the quantization level $K \to +\infty$.

Theorem 1.1.3. *Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big)$ be a random variable with probability distribution μ. Let $\mu = \mu_a + \mu_s = h \cdot \lambda_d + \mu_s$ denote the Lebesgue decomposition of μ with respect to the Lebesgue measure λ_d, where μ_a is the absolutely continuous part with density function h and μ_s is the singular part of μ.*

(i) Let $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, For every $K \in \mathbb{N}^*$, let $\widehat{\mu}^{x^{*,(K)}}$ and $\widehat{X}^{x^{*,(K)}}$ denote the quantization estimator of μ and X defined in (1.1.14) and (1.1.13) with respect to an optimal quantizer $x^{*,(K)} = (x_1^{*,(K)}, ..., x_K^{*,(K)})$ at level K. Then

$$\Big\| X - \widehat{X}^{x^{*,(K)}} \Big\|_p = \mathcal{W}_p(\mu, \widehat{\mu}^{x^{*,(K)}}) = e^*_{K,p}(\mu) \to 0 \ \text{ as } \ K \to +\infty. \qquad (1.1.23)$$

(*ii*) (Zador's Theorem) *Let $\mu \in \mathcal{P}_{p+\eta}(\mathbb{R}^d)$ for some $\eta > 0$. Then there exists a constant $C_{p,d}$ depending on p and d such that*

$$\lim_{K\to+\infty} K^{1/d}\, e^*_{K,p}(\mu) = C_{p,d}\Big[\int_{\mathbb{R}^d} h^{\frac{d}{d+p}} d\lambda_d\Big]^{\frac{1}{p}+\frac{1}{d}}.$$

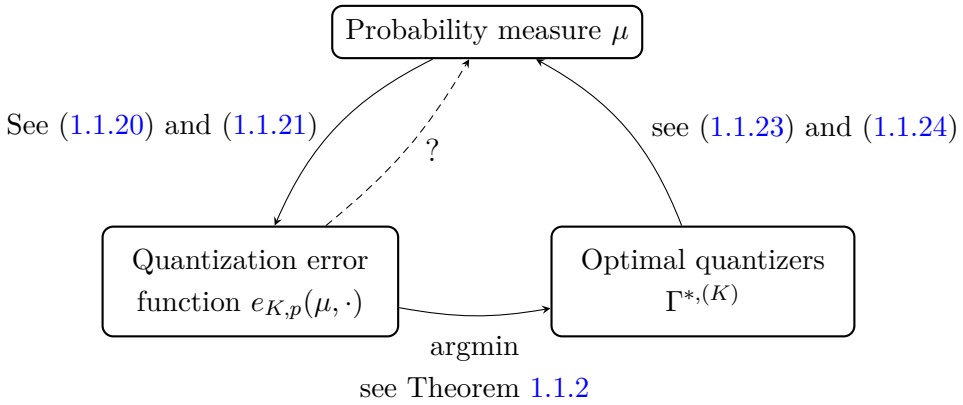
(*iii*) (Empirical measure theorem) *If $h \neq 0$ and $h \in L^{d/(d+r)}(\lambda_d)$, then*

$$\frac{1}{K}\sum_{1\le k\le K} \delta_{x_k^{*,(K)}} \overset{(\mathbb{R}^d)}{\Longrightarrow} \widetilde{\mu} = \frac{h^{d/(d+p)}(\xi)}{\int h^{d/(d+p)} d\lambda_d}\lambda_d(d\xi), \quad as \ \ K \to +\infty, \tag{1.1.24}$$

where for every $K \in \mathbb{N}^*$, $x^{*,(K)} = (x_1^{*,(K)}, ..., x_K^{*,(K)})$ denotes an optimal quantizer of μ at level K and $\overset{(S)}{\Longrightarrow}$ denotes the weak convergence of probability measures on a Polish space S.

We refer to Graf and Luschgy (2000)[Lemma 6.1, Theorem 6.2 and Theorem 7.5] for the proof of Theorem 1.1.3.

The answer to *Question 1* is composed by not only the above three convergences in Theorem 1.1.3 when $K \to +\infty$, but also by a close connection among the probability distribution, the quantization error function and the optimal quantizer (eventually the weight of optimal quantizer) when K is finite. First, the optimal quantizer is entirely characterized by the quantization error function since $x^* = (x_1^*, ..., x_K^*) \in \operatorname{argmin} e_{K,p}(\mu, \cdot)$. Moreover, Inequalities (1.1.20) and (1.1.21) show that for every $K \geq 1$ and $p \in [1, +\infty)$, the quantization error function $e_{K,p}(\mu, \cdot)$ is characterized by the probability distribution μ. Hence, the characterization relations between a probability measure μ, its L^p-quantization error function and its optimal quantizers can be synthesized by the following scheme:



Our motivation of Chapter 2 is to investigate more deeply the relations between these

three elements. In Chapter 2, we consider the "reverse" questions of (1.1.20) and (1.1.21): *When and how is a probability measure $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ characterized by its L^p-quantization error functions $e_{K,p}(\mu,\cdot)$? And if so, does the convergence in an appropriate sense of the L^p-quantization error functions characterizes the convergence of their probability distributions for the $\mathcal{W}_p$-distance?*

1.1.3.2 How to find the (quadratic) optimal quantizer?

(A) If the target probability distribution μ is known...

As far as we know, there does not exist a general method to find the L^p-optimal quantizers of $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ for every $p \geq 1$. However, if $p = 2$ and if the underlying norm on $\mathbb{R}^d$ is the Euclidean norm, there exist several numerical methods to find the quadratic optimal quantizer which correspond to the properties of the optimal quantizer in Theorem 1.1.2-(i) and (ii).

(A.1) Zero search algorithm and CLVQ algorithm. Let X be an $\mathbb{R}^d$-valued random variable with probability distribution μ satisfying $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Assume that μ is absolutely continuous with respect to the Lebesgue measure, i.e. $\mu = f \cdot \lambda_d$ with f its density function. For a fixed quantization level K, its quadratic distortion function $\mathcal{D}_{K,2}(\mu,\cdot)$ is differentiable at all point $x = (x_1, ..., x_K)$ s.t. $x_i \neq x_j, i \neq j$,

$$\frac{\partial \mathcal{D}_{K,2}(\mu,\cdot)}{\partial x_k}(x) = 2\int_{V_k(x)} (x_k - \xi) f(\xi)\lambda_d(d\xi) = 2\mathbb{E}\left[\mathbb{1}_{\{X \in V_k(x)\}}(x_k - X)\right], \quad \text{for } k = 1, ..., K. \tag{1.1.25}$$

As the quadratic optimal quantizer $x^* = (x_1^*, ..., x_K^*) \in \operatorname{argmin} \mathcal{D}_{K,2}(\mu,\cdot)$, one can use a zero search algorithm of the gradient $\nabla \mathcal{D}_{K,2}(\mu,\cdot)$, namely,

$$x^{[l+1]} = x^{[l]} - \gamma_{l+1} \nabla \mathcal{D}_{K,2}(\mu, x^{[l]}), \qquad \text{with } x^{[0]} \in \big(\text{Hull}(\text{supp}(\mu))\big)^K, \tag{1.1.26}$$

where $x^{[0]}$ has pairwise distinct components and $\big(\text{Hull}(\text{supp}(\mu))\big)^K$ denotes the closed convex hull of the support of μ. Furthermore, we obtain in Chapter 3 a detailed formula for the Hessian matrix $H_{\mathcal{D}_{K,2}(\mu,\cdot)}$ by applying Fort and Pagès (1995)[Lemma 11]. Consequently, when $d = 1$, one can replace γ_{l+1} by the inverse of the Hessian matrix $H_{\mathcal{D}_{K,2}(\mu,\cdot)}$, which leads to the classical Newton-Raphson procedure as follows,

$$x^{[l+1]} = x^{[l]} - H_{\mathcal{D}_{K,2}(\mu,\cdot)}(x^{[l]})^{-1} \nabla \mathcal{D}_{K,2}(\mu, x^{[l]}). \tag{1.1.27}$$

Furthermore, one can improve (1.1.27) by using the Levenberg-Marquardt algorithm

with an appropriate choice of λ_l as follows

$$x^{[l+1]} = x^{[l]} - \left[H_{\mathcal{D}_{K,2}(\mu,\,\cdot\,)}(x^{[l]}) + \lambda_l I_d\right]^{-1} \nabla \mathcal{D}_{K,2}(\mu, x^{[l]}). \tag{1.1.28}$$

Taking advantage of the representation of $\nabla \mathcal{D}_{K,2}(\mu, \cdot)$ as an expectation (see (1.1.25)), the above gradient descent has a stochastic counterpart called the CLVQ algorithm (Competitive Learning Vector Quantization), which works also in higher dimension ($d \geq 2$)

$$x^{[l+1]} = x^{[l]} - \gamma_{l+1}\left[\mathbb{1}_{\{X_{l+1} \in V_k(x)\}}(x_k^{[l]} - X_{l+1})\right]_{1 \leq k \leq K}, \quad \text{with } x^{[0]} \in \big(\mathrm{Hull}(\mathrm{supp}(\mu))\big)^K, \tag{1.1.29}$$

where $x^{[0]}$ has pairwise distinct components and $(X_l)_{l \geq 1}$ are independent copies of X. We refer to Pagès (2015)[Section 3.2] for more details of the CLVQ algorithm.

(A.2) Lloyd I algorithm. Lloyd I algorithm, firstly introduced in Lloyd (1982), is a fixed point search procedure which comes from the stationary property described in Theorem 1.1.2-(ii). Let $x^{[0]} = (x_1^{[0]}, ..., x_K^{[0]}) \in \mathrm{supp}(\mu)^K$, having pairwise distinct components, the Lloyd I algorithm computes the following iteration

$$x_k^{[l+1]} = \frac{\int_{C_k(x^{[l]})} \xi \mu(d\xi)}{\mu\big(C_k(x^{[l]})\big)}, \; k = 1, ..., K, \tag{1.1.30}$$

until some stopping criterions, for example, $x^{[l+1]} = (x_1^{[l+1]}, ..., x_K^{[l+1]}) = x^{[l]}$. In dimension 1, if μ is absolutely continuous with respect to the Lebesgue measure and its density function ρ is log-concave and $\log \rho$ is not piecewise affine, the Lloyd I algorithm has an exponential convergence rate (see Kieffer (1982)). The convergence of the Lloyd I algorithm in higher dimension is proved in Pagès and Yu (2016).

The integral over a Voronoï cell in (1.1.30) can be computed by using cubature formulas for numerical integration on convex set. For example, in low dimension ($d \leq 3$), we refer to the libraries available at the website *www.qhull.org*. In higher dimension, the computing time of such integral becomes intractable and we are led to switch to the *Randomized Lloyd I algorithm*, which relies on a Monte-Carlo method and can be written as follows,

- Let $N > K$. Simulate $X_1, ..., X_N \overset{i.i.d}{\sim} \mu$.
- Set $x^{[0]} = (x_1^{[0]}, ..., x_K^{[0]})$.

– Compute $x^{[l+1]} = \left(x_1^{[l+1]}, ..., x_K^{[l+1]}\right)$ by

$$x_k^{[l+1]} = \frac{\sum_{n=1}^N X_n \mathbb{1}_{\{X_n \in C_k(x^{[l]})\}}}{\sum_{n=1}^N \mathbb{1}_{\{X_n \in C_k(x^{[l]})\}}}, \quad k = 1, ..., K. \tag{1.1.31}$$

– Repeat the above iteration until some stopping criterion occurs.

(B) If the target probability distribution μ is unknown but there exists a known probability distribution sequence μ_n converging to μ in the Wasserstein distance...

This is a common situation in applications that μ_n is the empirical measure or μ is the stationary measure of a diffusion process $dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t$. This leads us to consider the consistency and the convergence rate of optimal quantizers for a $\mathcal{W}_p$-converging sequence of probability distributions.

Let $\mu_n \in \mathcal{P}_p(\mathbb{R}^d)$, $n \in \mathbb{N} \cup \{\infty\}$. For every $n \in \mathbb{N}$, let $x^{(n)}$ denote the optimal quantizer of μ_n at level K and order p. There are two ways to consider the consistency and the convergence rate of the optimal quantization. The first way is to directly study the convergence of optimal quantizers:

– Will $(x^{(n)})_{n\in\mathbb{N}}$ converge to an optimal quantizer of μ_∞?
This question is solved in Pollard (1982b)[Theorem 9] for $p = 2$ and we will prove it for every $p \geq 1$ in Chapter 3.

– Let $G_K(\mu_\infty) \coloneqq \left\{(x_1, ..., x_K) \in (\mathbb{R}^d)^K \mid (x_1, ..., x_K) \text{ is an optimal quantizer of } \mu_\infty\right\}$.
Can we obtain the convergence rate of $d\left(x^{(n)}, G_K(\mu_\infty)\right)$?
This question is solved in Chapter 3.

The second way is to study the convergence of the quantization errors, that is, we consider $x^{(n)}$ as a quantizer of μ_∞ and study the convergence of the quantization error $e_{K,p}(\mu_\infty, x^{(n)})$ $\left(\text{or equivalently } \mathcal{D}_{K,p}(\mu_\infty, x^{(n)})\right)$ to the optimal quantization error $e^*_{K,p}(\mu_\infty)$ of μ_∞ (or $\inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,p}(\mu_\infty, x)$).

– Does $\mathcal{D}_{K,p}(\mu_\infty, x^{(n)})$ converge to $\inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,p}(\mu_\infty, x)$?

– Can we obtain an estimation (e.g. an upper bound) of the convergence rate of

$$\left|\mathcal{D}_{K,p}(\mu_\infty, x^{(n)}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,p}(\mu_\infty, x)\right|?$$

When $\mu_n, n \in \mathbb{N}$, are the empirical measures and μ_∞ has a bounded support, a result is established in Biau et al. (2008). For a more general setting, e.g. for any $\mathcal{W}_p$-converged

probability distribution sequence or for the empirical measure with non-bounded support, the convergence rate results are established in Chapter 3.

1.1.3.3 How to apply the optimal quantization in numerical probability or in unsupervised learning?

In the unsupervised learning area, vector quantization has a close connection with the automatic classification (clustering analysis) through the *K-means* algorithm. The term *K-means* originates from the paper MacQueen (1967), which aims at finding an optimal partition $S = \{S_1, ..., S_K\}$ of a given set of observations $(\xi_1, ..., \xi_N) \in (\mathbb{R}^d)^N$ in order to minimize

$$\frac{1}{N}\sum_{n=1}^{N} \min_{k=1,...,K} d(\xi_n, m_k)^2, \quad \text{with } m_k \text{ the mean (or the centroid) of points in } S_k,$$

where d is a distance function or other functions to represent the similarity. If d is the l_p-distance, we recognise the common thread between the K-means algorithm and the optimal quantization method if we consider a probability measure μ defined by $\mu = \frac{1}{N}\sum_{n=1}^{N} \delta_{\xi_n}$. However, in the clustering analysis, d can also be other functions such as an inner product or the Jaccard distance according to the features we want to extract from the observations. For more details on the K-means algorithm, we refer to Duda et al. (2001) and Linder (2002) among many other references.

In the numerical probability, vector quantization is an efficient tool to compute regular and conditional expectations (see Pagès (1998), Bally and Pagès (2003) and Pagès and Printems (2003)). Thus, the quantization based numerical scheme has been developed for the simulation of the solution of the stochastic differential equation (see Pagès and Sagna (2015)) and for the Backward Stochastic Differential Equation or nonlinear filtering (see Pagès and Sagna (2018)). Moreover, the functional quantization technique can be used for the variance reduction in the simulation of diffusion process (see Lejay and Reutenauer (2012)) or solving stochastic inversion problems (see El Amri et al. (2019)). In financial mathematics, the quantization based scheme can be used in the option pricing, see Bally et al. (2005), Callegaro et al. (2017), Callegaro et al. (2015) and Bormetti et al. (2018).

In the second part of this thesis, we are interested in the application of the optimal quantization method to the simulation of the McKean-Vlasov equation. The terminology *McKean-Vlasov equation* originates from the paper McKean (1967) in which H.P. McKean

studies a partial differential equation on $\mathbb{R}^d$ having the following form

$$\begin{cases} \dfrac{\partial p}{\partial t} = \dfrac{1}{2} \displaystyle\sum_{1\leq i,j\leq d} \frac{\partial^2}{\partial x_i \partial x_j} e_{ij} p - \sum_{i\leq d} \frac{\partial}{\partial x_i} f_i p, \quad t > 0,\ x \in \mathbb{R}^d \\ \lim\limits_{t\downarrow 0} p = q \end{cases} \tag{1.1.32}$$

and whose solution p is the density of a stochastic process X. By now, the terminology *McKean-Vlasov equation* refers to the whole family of stochastic differential equations in which the coefficient functions depend not only on the position of process X_t but also on its probability distribution, namely,

$$\begin{cases} dX_t = b(t, X_t, \mu_t)dt + \sigma(t, X_t, \mu_t)dB_t \\ X_0 : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right) \text{ random variable} \\ \forall t \geq 0,\ \mu_t \text{ denotes the probability distribution of } X_t \end{cases} . \tag{1.1.33}$$

One important property of the McKean-Vlasov equation which attracts many studies in the literature is the *propagation of chaos*. Let $X_0^{1,N}, ..., X_0^{N,N}$ be i.i.d copies of X_0 and the N-particle system of the McKean-Vlasov equation is defined by

$$\begin{cases} \forall n \in \{1, ..., N\}, \\ dX_t^{n,N} = b(X_t^{n,N}, \mu_t^N)dt + \sigma(X_t^{n,N}, \mu_t^N)dB_t^n, \\ \text{for any } t \in [0, T],\ \mu_t^N := \frac{1}{N}\sum_{n=1}^N \delta_{X_t^{n,N}}, \end{cases} \tag{1.1.34}$$

Generally speaking, the *propagation of chaos* means that under some appropriate conditions, the empirical measure $\frac{1}{N}\sum_{n=1}^N \delta_{X^{n,N}}$ composed by the N particles $(X^{1,N}, ..., X^{N,N})$ converges to the distribution μ of the solution X of the McKean-Vlasov equation (1.1.33) as the number of particles $N \to +\infty$ and in this case, the N particles $X^{1,N}, \ldots, X^{N,N}$ tend to become independent. We refer to Gärtner (1988) for a detailed proof of the propagation of chaos among many other references.

There are many studies of the existence and uniqueness of solution of (1.1.33) under various conditions on b, σ among which we refer to Sznitman (1991) for a systematic presentation of the McKean-Vlasov equation and propagation of chaos in dimension 1, to Funaki (1984) and Jourdain (2000) for the weak uniqueness, the associated martingale problem and connection to the Boltzmann equation, to Jourdain et al. (2008) for the uniqueness of solution of the McKean-Vlasov equation driven by a Lévy processes and to Lacker (2018) for a recent idea of proof under the Lipschitz condition of coefficient function b and σ. A rigorous proof of the existence and uniqueness of a strong solution also interests us as it is a theoretical basis to devise and analyse the numerical scheme.

Let $M \in \mathbb{N}^*$ and $t_m := \frac{T}{M} \cdot m,\ m = 0, ..., M$. The "theoretical" Euler scheme of the McKean-Vlasov equation is defined by

$$\begin{cases} \bar{X}_{t_{m+1}} = \bar{X}_{t_m} + h \cdot b(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m}) + \sqrt{h}\,\sigma(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m}) Z_{m+1} \\ \bar{\mu}_{t_m} \text{ is the probability distribution of } \bar{X}_{t_m}, m = 0, ..., M \\ \bar{X}_0 = X_0 \end{cases} . \tag{1.1.35}$$

We first prove in Chapter 5 the convergence rate of (1.1.35) to the unique solution of (1.1.33) under appropriate conditions. However, unlike for regular stochastic differentials equation $dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t$, the Euler scheme (1.1.35) does not indicate how to simulate $\bar{\mu}_{t_m}$. That is why we call the scheme (1.1.35) the "theoretical" Euler scheme and this problem leads us to consider the possibility of using a quantization estimated distribution $\widehat{\mu}^x_{t_m}$ instead of $\bar{\mu}_{t_m}$.

Even though the theoretical Euler scheme cannot be directly simulated, the convergence result of the theoretical Euler scheme offers us a way to compare the functional convex order of two McKean-Vlasov processes. The comparison of the functional convex order between two stochastic processes was introduced in Pagès (2016) for the one dimensional martingale diffusions, i.e. solutions of

$$\begin{aligned} dX_t &= \sigma(t, X_t)dB_t, \quad X_0 = x \in \mathbb{R}, \\ dY_t &= \theta(t, Y_t)dB_t, \quad Y_0 = x \in \mathbb{R}. \end{aligned} \tag{1.1.36}$$

In Pagès (2016), the author obtains

$$\mathbb{E}\, F(X) \leq \mathbb{E}\, F(Y) \tag{1.1.37}$$

for any convex function $F : \mathbb{R} \to \mathbb{R}$ with r-polynomial growth under conditions that σ is convex in x and $\sigma \leq \theta$ by applying the convergence result of Euler scheme of (1.1.36). Moreover, such convex order result can be applied in the Optimal Stopping Theory and in the comparison of American option prices (see e.g. Pagès (2016) and Alfonsi et al. (2019)). We are interested in how to extend this functional convex order result to the McKean-Vlasov equation. In Chapter 6, we obtain the similar inequality as (1.1.37) for two processes $X := (X_t)_{t\in[0,T]}$ and $Y := (Y_t)_{t\in[0,T]}$ defined by the scaled McKean-Vlasov equations

$$\begin{aligned} dX_t &= (\alpha X_t + \beta)dt + \sigma(t, X_t, \mu_t)dB_t, \quad X_0 \in L^p(\mathbb{R}^d), \\ dY_t &= (\alpha Y_t + \beta)\, dt + \theta(t, Y_t, \nu_t)\, dB_t, \quad Y_0 \in L^p(\mathbb{R}^d), \\ &\alpha, \beta \in \mathbb{R} \text{ and } \forall t \in [0, T],\ \mu_t = \mathbb{P} \circ X_t^{-1},\ \nu_t = \mathbb{P} \circ Y_t^{-1} \end{aligned}$$

under appropriate conditions. Moreover, since the distribution of the solution process is

an important element for the analysis of the McKean-Vlasov equation, we will generalize the functional convex result to the functional of both process path and distribution of process, i.e.

$$\mathbb{E}\, G(X, (\mu_t)_{t\in[0,T]}) \leq \mathbb{E}\, G(Y, (\mu_t)_{t\in[0,T]}).$$

Now we go back to the numerical aspect of the McKean-Vlasov equation. The reason why we cannot directly simulate $\bar{\mu}_{t_m}$ in (1.1.35) is that we need a spatial discretization in ordre to approximate $\bar{\mu}_{t_m}$. In Bossy and Talay (1997), the authors show the particle method inspired by the principle of propagation of chaos and prove the convergence of this method in dimension 1. The principle of this method is to use an "empirical" mesure on the N-particle instead of $\bar{\mu}_{t_m}$ in the theoretical Euler scheme. We extend their result to dimension $d \geq 2$ in Chapter 7. Moreover, we develop several quantization based schemes and a hybrid particle-quantization scheme, analyse the error of each method and give the corresponding simulation examples in Chapter 7.

1.2 Contributions to the literature

This thesis is divided into two parts: Part I contains Chapter 2 and Chapter 3, which investigate two limit theorems for the optimal quantization. The first one is the characterization of the convergence in L^p-Wasserstein distance of a probability distribution sequence by its quantization error function sequence and the second limit theorem is the consistency and the convergence rate of optimal quantizers and the optimal error. Part II contains Chapter 4, 5, 6 and 7, in which are devised and analyzed several discretization schemes for the McKean-Vlasov equation. It includes the proof of existence and uniqueness of a strong solution, the functional convex order problem, the convergence rate of the particle method and various quantization based schemes.

1.2.1 Part I: Some limit theorems for the optimal quantization

Chapter 2 corresponds to the paper Liu and Pagès (2019) to appear in *Bernoulli* journal. This chapter studies the characterization of probability measure by the quantization error function. We establish the existence of a minimal level $K^* \in \mathbb{N}^*$ such that for any $K \geq K^*$,

– for any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$,

$$e_{K,p}(\mu, \cdot) = e_{K,p}(\nu, \cdot) \iff \mu = \nu,$$

– for any $\mu_n \in \mathcal{P}_p(\mathbb{R}^d),\ n \in \mathbb{N}^* \cup \{\infty\}$,

$$e_{K,p}(\mu_n,\ \cdot\ ) \xrightarrow{n\to+\infty} e_{K,p}(\mu_\infty,\ \cdot\ ) \text{ pointwise } \iff \mathcal{W}_p(\mu_n, \mu_\infty) \xrightarrow{n\to+\infty} 0.$$

The proof relies on a geometrical approach which is equivalent to the existence of a bounded open Voronoï cell in a Voronoï diagram and the above existence can be in turn derived from a minimal covering of the unit sphere by unit closed balls centered on the sphere. This geometrical approach is valid for any norm on $\mathbb{R}^d$. Moreover, in the quadratic Euclidean case, we establish by standard Hilbert analysis arguments that the minimal characterization level $K^* = 2$. This characterization result can be extended to any infinite dimensional separable Hilbert space.

Moreover, we define for $K \geq K^*$ a quantization based distance

$$\mathcal{Q}_{K,p} := \|e_{K,p}(\mu, \cdot) - e_{K,p}(\nu, \cdot)\|_{\sup}$$

and we prove that this distance is topologically equivalent to the Wasserstein distance $\mathcal{W}_p$ on $\mathcal{P}_p(\mathbb{R}^d)$. Furthermore, we prove that $\mathcal{Q}_{1,1}$ is a complete distance on $\mathcal{P}_1(\mathbb{R})$ and give a counterexample to show that the distances $\mathcal{Q}_{K,2}$, $K \geq 2$ are not complete on $\mathcal{P}_2(\mathbb{R})$ at the end of this chapter.

In Chapter 3, we establish the convergence rate of the quadratic optimal quantization for a probability sequence converging in the Wasserstein distance, which generalizes two former papers Pollard (1982a) and Biau et al. (2008). Let $\mu_n \in \mathcal{P}_2(\mathbb{R}^d), n \in \mathbb{N}^* \cup \{\infty\}$ be such that $\mathcal{W}_2(\mu_n, \mu_\infty) \to 0$ as $n \to +\infty$. For every $n \in \mathbb{N}^*$, let $x^{(n)}$ denote a quadratic optimal quantizer of μ_n and let

$$G_K(\mu_\infty) := \left\{(x_1^*, ..., x_N^*) \mid (x_1^*, ..., x_N^*) \text{ is an optimal quantizer of } \mu_\infty\right\}$$

denote the set of quadratic optimal quantizers of μ_∞ at level K. In Chapter 3, we denote the distortion function defined in Definition 1.1.1 of μ_n by $\mathcal{D}_{K,\mu_n}$, $n \in \mathbb{N} \cup \{\infty\}$, since we fix $p = 2$. One first result of Chapter 3 is the non-asymptotic upper bound of the quantization performance: for every $n \in \mathbb{N}^*$,

$$\mathcal{D}_{K,\mu_\infty}(x^{(n)}) - \inf_{x \in (\mathbb{R}^d)^K} \mathcal{D}_{K,\mu_\infty}(x) \leq 4e^*_{K,\mu_\infty} \mathcal{W}_2(\mu_n, \mu_\infty) + 4\mathcal{W}_2^2(\mu_n, \mu_\infty),$$

where e^*_{K,μ_∞} is the quadratic optimal error of μ_∞ at level K (defined in (1.1.7) with $p = 2$). Furthermore, under several appropriate conditions on the differentiability of the distortion function $\mathcal{D}_{K,\mu_\infty}$ and the positive definiteness of the Hessian matrix $H_{\mathcal{D}_{K,\mu_\infty}}$ of $\mathcal{D}_{K,\mu_\infty}$, we obtain the convergence rate of optimal quantizers: for n large enough, there

exist two positive constant $C^{(1)}_{\mu_\infty}$ and $C^{(2)}_{\mu_\infty}$ depending on μ_∞ such that

$$d\big(x^{(n)}, G_K(\mu_\infty)\big)^2 \leq C^{(1)}_{\mu_\infty}\mathcal{W}_2(\mu_n, \mu_\infty) + C^{(2)}_{\mu_\infty}\mathcal{W}_2^2(\mu_n, \mu_\infty).$$

The second part of Chapter 3 is devoted to the convergence rate of optimal quantization error of the empirical measure, which is also called the *clustering performance* in the field of unsupervised learning. We generalize the upper bound in Biau et al. (2008) for the probability distribution with a bounded support to any probability distribution with appropriate finite moments, hence including the normal distribution.

Let $X_1, ..., X_n, ...$ be i.i.d random variables with probability distribution μ and let $\mu_n^\omega := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical measure of μ. Let $x^{(n),\omega}$ denote the optimal quantizer of μ_n^ω. We establish two results about the *clustering performance* $\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x)$. If $\mu \in \mathcal{P}_q(\mathbb{R}^d)$ for some $q > 2$, the first result (see below), which is an application of Fournier and Guillin (2015), is sharp in K but suffers from the curse of dimensionality:

$$\begin{aligned}\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x)& \\ \leq C_{d,q,\mu,K} \times \begin{cases} n^{-1/4} + n^{-(q-2)/2q} & \text{if } d < 4 \text{ and } q \neq 4 \\ n^{-1/4}\big(\log(1+n)\big)^{1/2} + n^{-(q-2)/2q} & \text{if } d = 4 \text{ and } q \neq 4 \\ n^{-1/d} + n^{-(q-2)/2q} & \text{if } d > 4 \text{ and } q \neq d/(d-2) \end{cases}&,\end{aligned} \tag{1.2.1}$$

where $C_{d,q,\mu,K}$ is a constant depending on d, q, μ and roughly decreasing as $K^{-1/d}$.

Meanwhile, we establish another upper bound for the *clustering performance*

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x),$$

which is sharper in n, free from the curse of dimensionality but increasing faster than linearly in K. This second result generalizes the mean performance result for the empirical measure of a distribution μ with bounded support established in Biau et al. (2008) to any distributions μ having simply a finite second moment. We obtain

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x) \leq \frac{2K}{\sqrt{n}}\Big[r_{2n}^2 + \rho_K(\mu)^2 + 2r_1\big(r_{2n} + \rho_K(\mu)\big)\Big],$$

where $r_n := \big\|\max_{1\leq i\leq n}|X_i|\big\|_2$ and $\rho_K(\mu)$ is the maximum radius of $L^2(\mu)$-optimal

quantizers, defined by

$$\rho_K(\mu) \coloneqq \max\Big\{ \max_{1\le k\le K} |x_k^*|\,,\ \ (x_1^*,...,x_K^*) \text{ is an optimal quantizer of } \mu\Big\}.$$

Especially, we provide a precise upper bound for $\mu = \mathcal{N}(m,\Sigma)$, the multidimensionnal normal distribution by applying results in Pagès and Sagna (2012) as follows,

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu}(x) \le C_\mu \cdot \frac{2K}{\sqrt{n}}\Big[1+\log n + \gamma_K \log K\big(1+\frac{2}{d}\big)\Big],$$

where $\limsup_K \gamma_K = 1$ and $C_\mu = 12\cdot\Big[1\vee\log\big(2\int_{\mathbb{R}^d}\exp(\frac{1}{4}|\xi|^4)\mu(d\xi)\big)\Big]$.

1.2.2 Part II: Particle method, quantization based and hybrid schemes of the McKean-Vlasov equation, application to the convex ordering

Chapter 4 introduces Chapter 5, Chapter 6 and Chapter 7. In Chapter 5, we give a proof[(1)] based on Feyel's approach (see e.g. Bouleau (1988)[Section 7]) for the existence and uniqueness of a strong solution of the McKean-Vlasov equation (1.1.33) under the following Lipschitz assumption on b and σ

$$\begin{aligned}&\forall t\in[0,T], \forall x,y\in\mathbb{R}^d \text{ and } \forall \mu,\nu\in\mathcal{P}_p(\mathbb{R}^d), \exists L \ s.t.\\ &\qquad |b(t,x,\mu)-b(t,y,\nu)| \vee |||\sigma(t,x,\mu)-\sigma(t,y,\nu)||| \le L\big[\,|x-y| + \mathcal{W}_p(\mu,\nu)\big].\end{aligned} \tag{1.2.2}$$

The strategy is to define an application Φ_C depending on some constant $C\in\mathbb{R}_+^*$ on the product space "path space $\times$ the space of path distribution" as follows

$$\begin{aligned}&(Y,P_Y)\mapsto \Phi_C(Y,P_Y)\\ &\qquad\coloneqq \Big(\underbrace{\big(X_0+\int_0^t b(s,Y_s,\nu_s)ds+\int_0^t\sigma(s,Y_s,\nu_s)dB_s\big)_{t\in[0,T]}}_{=:\Phi_C^{(1)}(Y,P_Y)}, P_{\Phi_C^{(1)}(Y,P_Y)}\Big)\end{aligned}$$

where for a stochastic process X, P_X denotes its probability distribution (see further Section 5.1 for the detailed definition of P_X), then to prove that an appropriate restriction of Φ_C on a closed subset is a contraction mapping by controlling the value of C. Thus, the existence and uniqueness of a strong solution of the McKean-Vlasov equation is a direct result by applying the fixed-point theorem for contractions on a complete space.

(1) This proof is obvious not the first proof of the existence and uniqueness of a strong solution of the McKean-Vlasov equation under Lipschitz coefficient conditions, but we find the application of Feyel's approach in the McKean-Vlasov framework is mathematically elegant.

Throughout the proof, we also fix the definitions of "path space" and "the space of distribution of process" and respectively define the distances on both spaces. The proof of the existence and uniqueness of a strong solution and the definition of "path space" and "the space of distribution of process" are also the theoretical bases for the further quantization based schemes.

Once we obtained the existence and uniqueness of a strong solution, we show in Chapter 5 the convergence rate of the *theoretical* Euler scheme (1.1.35) of the McKean-Vlasov equation (1.1.33). If b, σ satisfy (1.2.2) and

$$\begin{aligned}&\forall t, s \in [0,T] \text{ with } s<t, \forall x \in \mathbb{R}^d, \forall \mu \in \mathcal{P}(\mathbb{R}^d), \text{ there exist } \tilde{L}, \gamma \in \mathbb{R}_+ \text{ s.t.}\\ &\quad |b(t,x,\mu)-b(s,x,\mu)| \vee |||\sigma(t,x,\mu)-\sigma(s,x,\mu)||| \leq \tilde{L}\big(1+|x|+\mathcal{W}_p(\mu,\delta_0)\big)(t-s)^\gamma,\end{aligned} \tag{1.2.3}$$

the convergence rate of the theoretical Euler scheme is the following

$$\sup_{0\leq m\leq M} \mathcal{W}_p(\bar{\mu}_{t_m}, \mu_{t_m}) \leq \left\| \sup_{0\leq m\leq M} \left|X_{t_m} - \bar{X}_{t_m}\right| \right\|_p \leq C_e h^{\frac{1}{2}\wedge\gamma}, \tag{1.2.4}$$

where C_e is a constant depending on $b, \sigma, L, T, \tilde{L}$ and $\|X_0\|_p$.

Chapter 6 establishes the convex order results for the scaled[(1)] McKean-Vlasov equation. Let $(X_t)_{t\in[0,T]}$, $(Y_t)_{t\in[0,T]}$ be two processes respectively defined by

$$\begin{aligned} dX_t &= (\alpha X_t + \beta)dt + \sigma(t, X_t, \mu_t)dB_t, \quad X_0 \in L^p(\mathbb{P}),\\ dY_t &= (\alpha Y_t + \beta)\,dt + \theta(t,\, Y_t,\, \nu_t)\,dB_t, \quad Y_0 \in L^p(\mathbb{P}), \end{aligned} \tag{1.2.5}$$

where $\alpha, \beta \in \mathbb{R}$ and for any $t \in [0,T]$, $\mu_t = \mathbb{P}\circ X_t^{-1}$, $\nu_t = \mathbb{P}\circ Y_t^{-1}$. For any two random variables X, Y valued in a Banach space $(E, \|\cdot\|_E)$, if for any convex function $\varphi : E \to \mathbb{R}$ such that

$$\mathbb{E}\,\varphi(X) \leq \mathbb{E}\,\varphi(Y) \text{ as soon as these two expectations make sense,}$$

then we call X is dominated by Y for the *convex order* and denote by $X \preceq_{cv} Y$. In Chapter 6, we prove that the Euler scheme (1.1.35) of the McKean-Vlasov equation propagates the convex order of random variables. Let $\bar{X}_{t_m}, \bar{Y}_{t_m}, m = 0, ..., M$ respectively denote the theoretical Euler scheme (1.1.35) of $(X_t)_{t\in[0,T]}$, $(Y_t)_{t\in[0,T]}$ with step $\frac{T}{M}$. If $X_0 \preceq_{cv} Y_0$ and the coefficient functions σ, θ are ordered for a matrix order in the sense

(1) By *scaled*, we mean that the drift b is an affine function.

that

$$\forall t \in [0,T],\ \forall x \in \mathbb{R}^d,\ \forall \mu \in \mathcal{P}(\mathbb{R}^d),$$
$$\theta(t,x,\mu)\theta(t,x,\mu)^* - \sigma(t,x,\mu)\sigma(t,x,\mu)^* \text{ is a positive semi-definite matrix,}$$

and σ is convex in x and non-decreasing in μ with respect to the convex order, then for any $m = 0, ..., M$, $\bar{X}_{t_m} \preceq_{cv} \bar{Y}_{t_m}$. Moreover, using a backward induction and taking advantage of the convergence result of Euler scheme (1.2.4), we obtain a functional convex order result for the processes, i.e. for any convex function $F : \mathcal{C}([0,T], \mathbb{R}^d) \to \mathbb{R}$ having an r-polynomial growth, $1 \leq r \leq p$, in the sense that

$$\forall \alpha \in \mathcal{C}([0,T], \mathbb{R}^d), \exists C \geq 0 \ \ s.t. \ \ |F(\alpha)| \leq C(1 + \|\alpha\|^r_{\sup}),$$

we have

$$\mathbb{E}\, F(X) \leq \mathbb{E}\, F(Y). \tag{1.2.6}$$

Finally, we generalize the above functional convex result (1.2.6) to functionals of the form

$$G : (\alpha, (\gamma_t)_{t\in[0,T]}) \in \mathcal{C}([0,T], \mathbb{R}^d) \times \mathcal{C}([0,T], \mathcal{P}_p(\mathbb{R}^d)) \mapsto G(\alpha, (\gamma_t)_{t\in[0,T]}) \in \mathbb{R},$$

where G is convex in α, non-decreasing in $(\gamma_t)_{t\in[0,T]}$ with respect to the convex order and has an r-polynomial growth, $1 \leq r \leq p$ and obtain a new convex order result for X, Y, $(\mu_t)_{t\in[0,T]}$ and $(\nu_t)_{t\in[0,T]}$ defined in (1.2.5) as follows,

$$\mathbb{E}\, G(X, (\mu_t)_{t\in[0,T]}) \leq \mathbb{E}\, G(Y, (\nu_t)_{t\in[0,T]}).$$

Chapter 7 analyzes the particle method and several quantization based schemes for the McKean-Vlasov equation

$$\begin{cases} dX_t = b(X_t, \mu_t)dt + \sigma(X_t, \mu_t)dB_t \\ X_0 : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right) \text{ random variable} \\ \forall t \geq 0,\ \mu_t \text{ denotes the probability distribution of } X_t \end{cases} \tag{1.2.7}$$

and the organization of Chapter 7 is detailed further on in Figure 4.1. We mainly consider the homogeneous equation to alleviate the notation but the extension of our results to the general case is standard and can be performed like in the regular SDE framework. The theoretical Euler scheme in the homogeneous case is

$$\begin{cases} \bar{X}_{t_{m+1}} = \bar{X}_{t_m} + h \cdot b(\bar{X}_{t_m}, \bar{\mu}_{t_m}) + \sqrt{h}\,\sigma(\bar{X}_{t_m}, \bar{\mu}_{t_m})Z_{m+1} \\ \bar{X}_0 = X_0, \ \ \bar{\mu}_{t_m} = P_{\bar{X}_{t_m}}, \end{cases} \tag{1.2.8}$$

where $M \in \mathbb{N}^*$, $h = \frac{T}{M}$, and $t_m = m \cdot h$, $m \in \{1, ..., M\}$.

The first method we studied is the *particle method*, which is inspired by the principle of *propagation of chaos* and can be considered as its discretion version. Let $\bar{X}_0^{1,N}, ..., \bar{X}_0^{N,N}$ be i.i.d copies of X_0 in (1.2.7). The *particle method* is defined by

$$\begin{cases} \forall n \in \{1, ..., N\}, \\ \bar{X}_{t_{m+1}}^{n,N} = \bar{X}_{t_m}^{n,N} + hb(\bar{X}_{t_m}^{n,N}, \bar{\mu}_{t_m}^N) + \sqrt{h}\,\sigma(\bar{X}_{t_m}^{n,N}, \bar{\mu}_{t_m}^N) Z_{m+1}^n \\ \bar{\mu}_{t_m}^N \coloneqq \frac{1}{N} \sum_{n=1}^N \delta_{\bar{X}_{t_m}^{n,N}} \end{cases}, \tag{1.2.9}$$

where $Z_m^n, n = 1, ..., N, m = 0, ..., M \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \mathbf{I}_q)$. The particle method is to use $\bar{\mu}_{t_m}^N$ as an estimator of $\bar{\mu}_{t_m}$ for each Euler step. In the case of dimension 1, the convergence rate of $\bar{\mu}_{t_m}^N$ to $\bar{\mu}_m$ as $N \to +\infty$ has been established in Bossy and Talay (1997). For the convergence rate in higher dimension ($d \geq 2$), we obtain in Section 7.1 that

$$\left\| \sup_{1 \leq m \leq M} \mathcal{W}_p(\bar{\mu}_{t_m}^N, \bar{\mu}_{t_m}) \right\|_p \leq C_{d,p,L,T} \left\| \mathbb{W}_p(\bar{\mu}, \nu^N) \right\|_p,$$

where $\bar{\mu}$ denotes the probability distribution of $\bar{X} = (\bar{X}_t)_{t \in [0,T]}$ defined further in (5.2.3) and ν^N denotes the empirical measure of $\bar{\mu}$. Moreover, if $\|X_0\|_{p+\varepsilon} < +\infty$ for some $\varepsilon > 0$, we also obtain in Section 7.1 by using results in Fournier and Guillin (2015) that

$$\left\| \sup_{1 \leq m \leq M} \mathcal{W}_p(\bar{\mu}_{t_m}^N, \bar{\mu}_m) \right\|_p \leq \widetilde{C} \times \begin{cases} N^{-\frac{1}{2p}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p > d/2 \text{ and } \varepsilon \neq p \\ N^{-\frac{1}{2p}} \left[\log(1+N) \right]^{\frac{1}{p}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p = d/2 \text{ and } \varepsilon \neq p \\ N^{-\frac{1}{d}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p \in (0, d/2) \text{ and } p + \varepsilon \neq \frac{d}{(d-p)} \end{cases},$$

where $\widetilde{C}$ is a constant depending on $p, \varepsilon, d, b, \sigma, L, T$.

The second studied method is the *quadratic optimal quantization method*. The idea of devising quantization based scheme for the simulation of the McKean-Vlasov equation first appears in Gobet et al. (2005)[Section 4] in a slightly different framework. Let $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$, $m = 1, ..., M$ be the quantizer of $\bar{X}_{t_m}$ in the m-th Euler step.

The theoretical quantization based scheme is to compute

$$\begin{cases} \widetilde{X}_0 = X_0, \ \ \widehat{X}_0 = \mathrm{Proj}_{x^{(0)}}(\widetilde{X}_0) \\ \widetilde{X}_{t_{m+1}} = \widehat{X}_{t_m} + h \cdot b(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) + \sqrt{h}\,\sigma(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) Z_{m+1}, \ \ m = 0, ..., M-1 \\ \text{where } h = \frac{T}{M} \ \text{ and } \ \widehat{\mu}_{t_m} = P_{\widehat{X}_{t_m}} \\ \widehat{X}_{t_{m+1}} = \mathrm{Proj}_{x^{(m+1)}}(\widetilde{X}_{t_{m+1}}). \end{cases} \tag{1.2.10}$$

We propose in Chapter 7 the error analysis of the above quantization procedure and three different ways of practically implementing the quantization based method to explicitly express $\widehat{\mu}_{t_m}$.

(1) In the Vlasov case, we can use the recursive quantization method, which is firstly introduced in Pagès and Sagna (2015) for regular stochastic differential equations. By the recursive quantization method, we derive a Markovian transition of $(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})$ based on the quantized scheme (1.2.10). Let $p^{(m)} = (p_1^{(m)}, ..., p_K^{(m)})$ denote the corresponding weight of the quantizer $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$. Thus $\widehat{\mu}_{t_m} = \sum_{k=1}^K \delta_{x_k^{(m)}} p_k^{(m)}$. The Markovian transition of $(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})$ by the *recursive quantization method* that we propose in Section 7.3 is

$$\begin{aligned} &\mathbb{P}\big(\widehat{X}_{t_{m+1}} = x_j^{(m+1)} \mid \widehat{X}_{t_m} = x_i^{(m)}, p^{(m)}\big) \\ &\quad = \mathbb{P}\Big[\big(x_i^{(m)} + h \sum_{k=1}^K p_k^{(m)} \beta(x_i^{(m)}, x_k^{(m)}) + \sqrt{h} \sum_{k=1}^K p_k^{(m)} a(x_i^{(m)}, x_k^{(m)}) Z_{m+1}\big) \in C_j(x^{(m+1)})\Big] \end{aligned}$$

and given $p^{(m)}$, we can compute for every $j = 1, ..., K$ by

$$\begin{aligned} p_j^{(m+1)} &= \mathbb{P}\big(\widehat{X}_{t_{m+1}} = x_j^{(m+1)} \mid p^{(m)}\big) \\ &= \sum_{i=1}^K \mathbb{P}\big(\widehat{X}_{t_{m+1}} = x_j^{(m+1)} \mid \widehat{X}_{t_m} = x_i^{(m)}, p^{(m)}\big) \cdot \mathbb{P}(\widehat{X}_{t_m} = x_i^{(m)}). \end{aligned}$$

We provide the proof of the above equalities in Section 7.3 and will explain in the same section how to apply the Lloyd I algorithm to improve the simulation accuracy.

(2) The second way to explicitly express $\widehat{\mu}_{t_m}$ is to use the optimal quantizer of the normal distribution $\mathcal{N}(0, \mathbf{I}_q)$ and its weight, which can be downloaded from the website

$$\mathrm{www.quantize.maths - fi.com/gaussian\_database}$$

for dimension $q = 1, ..., 10$. Let $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$ denote the quantizer of $\bar{X}_{t_m}$ in the m-th Euler step. Let $z = (z_1, ..., z_J)$ be an optimal quantizer of $\mathcal{N}(0, \mathbf{I}_q)$ with $J > K$ and let $w = (w_1, ..., w_J)$ be the corresponding weight of z. The scheme based

on such optimal quantizers of $\mathcal{N}(0, \mathbf{I}_q)$[(1)] can be written by

$$\begin{cases} \widetilde{X}_0 = X_0, \ \ \widehat{X}_0 = \mathrm{Proj}_{x^{(0)}}(\widetilde{X}_0) \\ \widetilde{X}_{t_{m+1}} = \widehat{X}_{t_m} + h \cdot b(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) + \sqrt{h}\, \sigma(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) \widehat{Z}_{m+1}, \ \ m = 0, ..., M-1 \\ \text{where } h = \frac{T}{M} \ \text{ and } \ \widehat{\mu}_{t_m} = P_{\widehat{X}_{t_m}} \\ \widehat{X}_{t_{m+1}} = \mathrm{Proj}_{x^{(m+1)}}(\widetilde{X}_{t_{m+1}}), \end{cases},$$

where $\widehat{Z}_m \overset{i.i.d}{\sim} \sum_{j=1}^{J} w_j \delta_{z_j}$. We call this method the *doubly quantized scheme* and we establish in Section 7.4 the error analysis of this method.

(3) Let $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$, $m = 0, 1, ..., M$, be a sequence of quantizers. As we prove the convergence rate of particle method, one can also implement the optimal quantization method on (1.2.9) as follows:

$$\begin{cases} \forall n \in \{1, ..., N\}, \\ \widetilde{X}_{t_{m+1}}^{n,N} = \widetilde{X}_{t_m}^{n,N} + h \cdot b(\widetilde{X}_{t_m}^{n,N}, \widehat{\mu}_{t_m}^{K}) + \sqrt{h}\, \sigma(\widetilde{X}_{t_m}^{n,N}, \widehat{\mu}_{t_m}^{K}) Z_{m+1}^{n} \\ \widehat{\mu}_{t_m}^{K} = \left(\frac{1}{N} \sum_{n=1}^{N} \delta_{\widetilde{X}_{t_m}^{n,N}}\right) \circ \mathrm{Proj}_{x^{(m)}}^{-1} = \sum_{k=1}^{K} \left[\delta_{x_k^{(m)}} \cdot \sum_{n=1}^{N} \mathbb{1}_{V_k(x^{(m)})}(\widetilde{X}_{t_m}^{n,N})\right] \\ \bar{X}_0^{n,N} \overset{i.i.d}{\sim} X_0, \ \ Z_m^n \overset{i.i.d}{\sim} \mathcal{N}(0, I_q) \end{cases}.$$

We call the above scheme the *hybrid particle-quantization scheme* (*hybrid scheme* for short). The error analysis of this scheme will be shown in Section 7.5.

At the end of Chapter 7, we give two examples simulated by the above numerical methods. The first one is the simulation of the Burgers equation introduced in Sznitman (1991) and Bossy and Talay (1997). The Burgers equation provide an explicit solution so we can compare the accuracy of different methods. The second example is 3-dimensional which was firstly introduced in Baladron et al. (2012) and also simulated in Reis et al. (2018).

(1) By a slight abus of notation, we use here the same notation as in (1.2.10).

Part I :
Limit Theorems for the Optimal Quantization

Chapter 2

Characterization of probability distribution convergence in Wasserstein distance by L^p-quantization error function

This chapter corresponds to the paper Liu and Pagès (2019) to appear in *Bernoulli* journal, which is a joint work with Gilles Pagès.

Abstract: We establish conditions to characterize probability measures by their L^p-quantization error functions in both $\mathbb{R}^d$ and Hilbert settings. This characterization is two-fold: static (identity of two distributions) and dynamic (convergence for the L^p-Wasserstein distance). We first propose a criterion on the quantization level N, valid for any norm on $\mathbb{R}^d$ and any order p based on a geometrical approach involving the Voronoï diagram. Then, we prove that in the L^2-case on a (separable) Hilbert space, the condition on the level N can be reduced to $N = 2$, which is optimal. More quantization based characterization cases in dimension 1 and a discussion of the completeness of a distance defined by the quantization error function can be found at the end of this paper.

Keyword: Probability distribution characterization, Vector quantization, Voronoï diagram, Wasserstein convergence

2.1 Introduction

Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote a probability space and let X be a random variable defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and valued in $(E, |\cdot|_E)$, where E is $\mathbb{R}^d$ or a separable Hilbert space H and $|\cdot|_E$ denotes respectively the norm on $\mathbb{R}^d$ or the norm on H induced by the inner product $(\cdot\,|\,\cdot)_H$. Let μ denote the probability distribution of X, denoted by $\mathbb{P}_X = \mu$ or $\text{Law}(X) = \mu$ and assume that μ has a finite p-th moment, $p \in [1, +\infty)$. The quantizer (also called *codebook* in signal compression or *cluster center* in machine learning theory) is a finite set of points in E, denoted by $\Gamma = \{x_1, ..., x_N\} \subset E$. Let us define the distance between a point ξ and a set A in E by $d(\xi, A) = \min_{a\in A} |\xi - a|_E$. The L^p-mean quantization error of Γ, defined by

$$e_p(\mu, \Gamma) := \|d(X, \Gamma)\|_p = \Big[\int_E \min_{a\in\Gamma} |\xi - a|_E^p \, \mu(d\xi)\Big]^{\frac{1}{p}},$$

is used to describe the accuracy level of representing the probability measure μ by Γ. Let $N \geq 1$. A quantizer $\Gamma^{*,(N)}$ satisfying

$$e_p(\mu, \Gamma^{*,(N)}) = \inf_{\substack{\Gamma\subset E,\\ \text{card}(\Gamma)\leq N}} \Big[\mathbb{E}\, d(X, \Gamma)^p\Big]^{\frac{1}{p}} = \inf_{\substack{\Gamma\subset E,\\ \text{card}(\Gamma)\leq N}} \Big[\int_E \min_{a\in\Gamma} |\xi - a|_E^p \, \mu(d\xi)\Big]^{\frac{1}{p}} \quad (2.1.1)$$

is called an *L^p-optimal quantizer* (or *optimal quantizer* in short) at *level N*. We refer to Graf and Luschgy (2000)[Theorem 4.12] for the existence of such an optimal quantizer on $\mathbb{R}^d$ and to Luschgy and Pagès (2002)[Proposition 2.1] or Cuesta and Matrán (1988) on (separable) Hilbert spaces. There is usually no closed form for optimal quantizers, however, in the quadratic case ($p = 2$), it can be computed by the stochastic optimization methods such as the CLVQ algorithm or the randomized Lloyd algorithm (see Pagès (2015)[Section 3], Kieffer (1982) and Pagès and Yu (2016)).

Optimal quantizers $\Gamma^{*,(N)}$ "carries" the information of the initial measure. For example, let $\mu \in \mathcal{P}_{p+\varepsilon}(\mathbb{R}^d)$ for some $\varepsilon > 0$, where

$$\mathcal{P}_p(E) := \{\mu \text{ probability distribution on } E \text{ s.t. } \int_E |\xi|_E^p \, \mu(d\xi) < +\infty\}.$$

Let $\mu = h \cdot \lambda_d$ be an absolutely continuous distribution (λ_d denotes Lebesgue measure). If for every level $N \geq 1$, $\Gamma^{*,(N)}$ is an optimal quantizer of μ at level N, then

$$\frac{1}{N} \sum_{x\in\Gamma^{*,(N)}} \delta_x \overset{(\mathbb{R}^d)}{\Longrightarrow} \widetilde{\mu} = \frac{h^{d/(d+p)}(\xi)}{\int h^{d/(d+p)} d\lambda_d} \lambda_d(d\xi), \quad \text{as} \quad N \to +\infty, \quad (2.1.2)$$

where, for a Polish space S, $\overset{(S)}{\Longrightarrow}$ denotes the weak convergence of probability measures on S. We refer to Graf and Luschgy (2000)[Theorem 7.5] for a proof of this result. This

weak convergence (2.1.2) emphasizes that, an absolutely continuous probability measure μ is entirely characterized by the sequence of L^p-optimal quantizers $\Gamma^{*,(N)}$ at levels N, $N \geq 1$.

We consider now the L^p-mean quantization error function as follows.

Definition 2.1.1 (Quantization error function)**.** *Let $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, $p \in [1, +\infty)$. The L^p-mean quantization error function of μ at level N, denoted by $e_{N,p}(\mu, \cdot)$, is defined by:*

$$\begin{array}{rccc} e_{N,p}(\mu,\cdot): & (\mathbb{R}^d)^N & \longrightarrow & \mathbb{R}_+ \\ & x=(x_1,\ldots,x_N) & \longmapsto & e_{N,p}(\mu,x)=\Big[\displaystyle\int_{\mathbb{R}^d}\min_{1\leq i\leq N}|\xi-x_i|^p\,\mu(d\xi)\Big]^{\frac{1}{p}}. \end{array} \tag{2.1.3}$$

The definition of $e_{N,p}(\mu, \cdot)$ obviously depends on the associated norm on $\mathbb{R}^d$ and the variable of $e_{N,p}(\mu, \cdot)$ is a priori an N-tuple in $(\mathbb{R}^d)^N$. However, for a finite quantizer $\Gamma \subset \mathbb{R}^d$, if the level $N \geq \mathrm{card}(\Gamma)$, then for any N-tuple $x^\Gamma = (x_1^\Gamma, \ldots, x_N^\Gamma) \in (\mathbb{R}^d)^N$ such that $\Gamma = \{x_1^\Gamma, \ldots, x_N^\Gamma\}$, we have $e_p(\mu, \Gamma) = e_{N,p}(\mu, x^\Gamma)$. For example, $e_p\big(\mu, \{x_1, x_2\}\big) = e_{2,p}\big(\mu, (x_1, x_2)\big) = e_{3,p}\big(\mu, (x_1, x_1, x_2)\big)$, etc. Note that $e_{N,p}$ is a symmetric function on $(\mathbb{R}^d)^N$ and that, owing to the above definition,

$$\inf_{\Gamma\subset\mathbb{R}^d,\mathrm{card}(\Gamma)\leq N} e_p(\mu,\Gamma) = \inf_{x\in(\mathbb{R}^d)^N} e_{N,p}(\mu,x). \tag{2.1.4}$$

Therefore, throughout this paper, with a slight abuse of notation, we will also denote the L^p-quantization error at level N for a quantizer Γ of size at most N by $e_{N,p}(\mu, \Gamma)$.

The equality (2.1.4) directly shows that the optimal quantizers are characterized by the L^p-mean quantization error functions. Next, we show that the quantization error function $e_{N,p}(\mu, \cdot)$ is entirely characterized by the probability distribution μ.

Notice that for any $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, the function $e_{N,p}(\mu, \cdot)$ defined in (2.1.3) is 1-Lipschitz continuous for every $N \geq 1$ since for any $x = (x_1, \ldots, x_N), y = (y_1, \ldots, y_N) \in (\mathbb{R}^d)^N$,

$$\begin{aligned} |e_{N,p}(\mu,x)-e_{N,p}(\mu,y)| &= \Big|\Big[\int_{\mathbb{R}^d}\min_{1\leq i\leq N}|\xi-x_i|^p\,\mu(d\xi)\Big]^{\frac{1}{p}} - \Big[\int_{\mathbb{R}^d}\min_{1\leq j\leq N}|\xi-y_j|^p\,\mu(d\xi)\Big]^{\frac{1}{p}}\Big| \\ &\leq \Big[\int_{\mathbb{R}^d}\Big|\min_{1\leq i\leq N}|\xi-x_i| - \min_{1\leq j\leq N}|\xi-y_j|\Big|^p\mu(d\xi)\Big]^{\frac{1}{p}} \quad \text{(by the Minkowski inequality)} \\ &\leq \Big[\int_{\mathbb{R}^d}\max_{1\leq i\leq N}|x_i-y_i|^p\,\mu(d\xi)\Big]^{\frac{1}{p}} = \max_{1\leq i\leq N}|x_i-y_i|\,. \end{aligned} \tag{2.1.5}$$

We recall now the definition of the L^p-Wasserstein distance.

Definition 2.1.2 (L^p-Wasserstein distance)**.** *Let (S,d) be a Polish space and $\mathcal{S} = Bor(S,d)$ be its Borel σ-field. For $p \in [1,+\infty)$, let $\mathcal{P}_p(S)$ denote the set of probability measures on $(S,\mathcal{S})$ with a finite p^{th}-moment. The L^p-Wasserstein distance $\mathcal{W}_p(\mu,\nu)$ between $\mu,\nu \in \mathcal{P}_p(S)$, denoted by $\mathcal{W}_p(\mu,\nu)$, is defined by*

$$\begin{aligned}\mathcal{W}_p(\mu,\nu) &= \Big(\inf_{\pi\in\Pi(\mu,\nu)} \int_{S\times S} d(x,y)^p \pi(dx,dy)\Big)^{\frac{1}{p}} \\ &= \inf\Big\{ \Big[\mathbb{E}\, d(X,Y)^p\Big]^{\frac{1}{p}},\ X,Y : (\Omega,\mathcal{A},\mathbb{P}) \to (S,\mathcal{S}) \text{ with } \mathbb{P}_X = \mu, \mathbb{P}_Y = \nu \Big\},\end{aligned} \tag{2.1.6}$$

where in the first line of (2.1.6), $\Pi(\mu,\nu)$ denotes the set of all probability measures on $(S^2,\mathcal{S}^{\otimes 2})$ with respective marginals μ and ν.

If we consider $e_{N,p}(\mu,x)$ as a function of $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, then $e_{N,p}$ is also 1-Lipschitz in μ. In fact, let X,Y be two random variables with probability distributions $\mathbb{P}_X = \mu$ and $\mathbb{P}_Y = \nu$. For every N-tuple $x = (x_1,\dots,x_N) \in (\mathbb{R}^d)^N$, we have

$$\begin{aligned}\big|e_{N,p}(\mu,x) - e_{N,p}(\nu,x)\big| &= \Big|\ \Big\| \min_{i=1,\dots,N} |X - x_i| \Big\|_p - \Big\| \min_{i=1,\dots,N} |Y - x_i| \Big\|_p \ \Big| \\ &\le \Big\| \min_{i=1,\dots,N} |X - x_i| - \min_{i=1,\dots,N} |Y - x_i| \Big\|_p \text{(by the Minkowski inequality)} \\ &\le \Big\| \max_{i=1,\dots,N} |\ |X - x_i| - |Y - x_i|\ | \Big\|_p \le \|X - Y\|_p .\end{aligned} \tag{2.1.7}$$

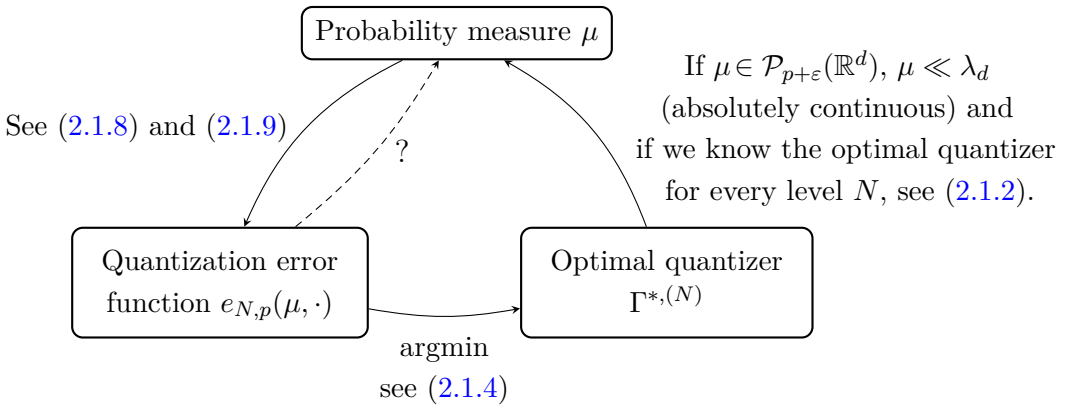
As this inequality holds for every couple (X,Y) of random variables with marginal distributions μ and ν, it follows that for every level $N \ge 1$,

$$\|e_{N,p}(\mu,\cdot) - e_{N,p}(\nu,\cdot)\|_{\sup} := \sup_{x\in(\mathbb{R}^d)^N} |e_{N,p}(\mu,x) - e_{N,p}(\nu,x)| \le \mathcal{W}_p(\mu,\nu). \tag{2.1.8}$$

Hence, if $(\mu_n)_{n\ge1}$ is a sequence in $\mathcal{P}_p(\mathbb{R}^d)$ converging for the $\mathcal{W}_p$-distance to $\mu_\infty \in \mathcal{P}_p(\mathbb{R}^d)$, then

$$\|e_{N,p}(\mu_n,\cdot) - e_{N,p}(\mu_\infty,\cdot)\|_{\sup} \le \mathcal{W}_p(\mu_n,\mu_\infty) \xrightarrow{n\to+\infty} 0. \tag{2.1.9}$$

Definition 2.1.1, and the inequalities (2.1.5), (2.1.7), (2.1.8), (2.1.9) can be directly extended to any separable Hilbert space H. Inequalities (2.1.8) and (2.1.9) show that for every $N \ge 1$, and $p \in [1,+\infty)$, the quantization error function $e_{N,p}(\mu,\cdot)$ is characterized by the probability distribution μ. Hence, the characterization relations between a probability measure μ, its L^p-quantization error function and its optimal quantizers can be synthesized by the following scheme:



The characterization of a probability measure μ by its L^p-optimal quantizers suggests to consider the "reverse" questions of (2.1.8) and (2.1.9): *When is a probability measure $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ characterized by its L^p-quantization error function $e_{N,p}(\mu,\cdot)$? And if so, does the convergence in an appropriate sense of the L^p-quantization error functions characterizes the convergence of their probability distributions for the $\mathcal{W}_p$-distance?*

These questions can be formalized as follows (the first one in a slightly extended sense):

- **Question 1 - Static characterization**:
  If for $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, $e_{N,p}(\mu,\cdot) = e_{N,p}(\nu,\cdot) + C$ for some real constant C, then do we have $\mu = \nu$ (and $C = 0$)?
- **Question 2 - Characterization of $\mathcal{W}_p$-convergence**:
  If for $\mu_n, n \geq 1$, $\mu_\infty \in \mathcal{P}_p(\mathbb{R}^d)$, $e_{N,p}(\mu_n,\cdot)$ converges pointwise to $e_{N,p}(\mu_\infty,\cdot)$, then do we have $\mathcal{W}_p(\mu_n,\mu_\infty) \xrightarrow{n\to+\infty} 0$?

For any $N_1, N_2 \in \mathbb{N}^*$ with $N_1 \leq N_2$, it is clear that

$$e_{N_2,p}(\mu,\cdot) = e_{N_2,p}(\nu,\cdot)\big(\text{resp.}e_{N_2,p}(\mu_n,\cdot) \xrightarrow{n\to+\infty} e_{N_2,p}(\mu_\infty,\cdot)\big)$$

implies

$$e_{N_1,p}(\mu,\cdot) = e_{N_1,p}(\nu,\cdot)\big(\text{resp.}e_{N_1,p}(\mu_n,\cdot) \xrightarrow{n\to+\infty} e_{N_1,p}(\mu_\infty,\cdot)\big).$$

Hence, beyond these two above questions, we need to determine an as low as possible level N for which both answers are positive. For this purpose, we define

$$N_{d,p,|\cdot|} := \min\{N \in \mathbb{N}^* \text{ such that answers to } \textit{Questions 1} \text{ and } \textit{2} \text{ for } e_{N,p} \text{ are positive}\}. \tag{2.1.10}$$

The paper is organized as follows. We first recall in Section 2.1.1 some properties of the Wasserstein distance $\mathcal{W}_p$. Then in Section 2.2, we begin to analyze the problem of probability distribution characterization in a general finite dimensional framework by considering any dimension d, any order p and any norm on $\mathbb{R}^d$. We show that a positive answer to *Question 1 and 2* follows from the existence of a bounded open Voronoï cell in a Voronoï diagram of size N, which in turn can be derived from a minimal covering of the unit sphere by unit closed balls centered on the sphere. As a consequence, we define for $N \geq N_{d,p,|\cdot|}$ a quantization based distance

$$\mathcal{Q}_{N,p} \coloneqq \|e_{N,p}(\mu,\cdot) - e_{N,p}(\nu,\cdot)\|_{\sup}$$

which we will prove to be topologically equivalent to the Wasserstein distance $\mathcal{W}_p$. The results in this section are established for $p \geq 1$, but several results can be extended to the case $0 < p < 1$ by the usual adaptations of the proofs.

In Section 2.3, we consider the quadratic case (*i.e.* the order p=2) and extend the characterization result to probability distributions on a separable Hilbert space H with the norm $|\cdot|_H$ induced by the inner product $(\cdot \mid \cdot)_H$. In this section, we will prove by a purely analytical method that $N_{H,2,|\cdot|_H} = 2$ [(1)] and the topological equivalence of Wasserstein distance $\mathcal{W}_2$ and the distance $\mathcal{Q}^H_{2,2}(\mu,\nu) \coloneqq \|e_{2,2}(\mu,\cdot) - e_{2,2}(\nu,\cdot)\|_{\sup}$ on $\mathcal{P}_2(H)$.

Section 2.4 is devoted to the one-dimensional setting. Quantization based characterization not yet covered by the discussion in Section 2.2 and Section 2.3 are established. Furthermore, we prove that $\mathcal{Q}_{1,1}$ is a complete distance on $\mathcal{P}_1(\mathbb{R})$ and give a counterexample to show that the distances $\mathcal{Q}_{N,2}$, $N \geq 2$ are not complete on $\mathcal{P}_2(\mathbb{R})$ in Section 2.4.2.

2.1.1 Preliminaries on the Wasserstein distance

Let (S,d) be a general Polish metric space. The relation between weak convergence and convergence for the Wasserstein distance $\mathcal{W}_p$ (see Definition 2.1.2) is recalled in Theorem 2.1.1. We recall below some useful facts about the L^p-Wasserstein distance that will be called upon further on. The first one is that, for every $p \in [1,+\infty)$, $\mathcal{W}_p$ is a distance on $\mathcal{P}_p(S)$ $\big(\mathcal{W}_p^p$ if $p \in (0,1)\big)$, see e.g. Villani (2003)[Theorem 7.3] for the proof and Berti et al. (2015) for a recent reference. Next, the metric space $\big(\mathcal{P}_p(S), \mathcal{W}_p\big)$ is separable and complete, see e.g. Bolley (2008) for the proof. More generally, we refer to Villani (2009)[Chapter 6] for an in depth presentation of Wasserstein distance and

(1) Since the dimension of the Hilbert space that we discuss in this section can be finite or infinite, we write directly H instead of d in the subscript of $N_{d,p,|\cdot|}$.

its properties.

Theorem 2.1.1. *(see Villani (2003)[Theorem 7.12]) Let $\mu_n \in \mathcal{P}_p(S)$ for every $n \in \mathbb{N}^* \cup \{\infty\}$. Let $p \in [1, +\infty)$. Then,*

(a) $\mathcal{W}_p(\mu_n, \mu_\infty) \to 0$ *if and only if* $\begin{cases} (\alpha)\ \mu_n \overset{(S)}{\Longrightarrow} \mu_\infty \\ (\beta)\ \exists\, x_0 \in S, \int_S d(x_0,\xi)^p \mu_n(d\xi) \to \int_S d(x_0,\xi)^p \mu_\infty(d\xi) \end{cases}$.

(b) If

$$\exists\, x_0 \in S, \quad \lim_{R\to+\infty} \sup_{n\geq 1} \int_{d(x_0,\xi)^p \geq R} d(x_0,\xi)^p \mu_n(d\xi) = 0, \tag{2.1.11}$$

then $(\mu_n)_{n\geq 1}$ is relatively compact for the Wasserstein distance $\mathcal{W}_p$.

2.2 General quantization based characterizations on $\mathbb{R}^d$

This section is devoted to establishing a general criterion that positively answers to *Questions 1 and 2* in any dimension d, for any order p and any norm on $\mathbb{R}^d$. The idea is to design an approximate identity $(\varphi_\varepsilon)_{\varepsilon>0}$[(1)] based on the quantization error function $e_{N,p}(\mu,\cdot)$. Our construction of $(\varphi_\varepsilon)_{\varepsilon>0}$ relies on a purely geometrical idea: it is based on a specified Voronoï diagram containing a bounded open Voronoï cell that we introduce in Section 2.2.1. The static characterization is established in Theorem 2.2.1. Furthermore, Theorem 2.2.2 shows that a pointwise convergence of the quantization error functions is enough to imply the $\mathcal{W}_p$-convergence of a $\mathcal{P}_p(\mathbb{R}^d)$-valued sequence.

2.2.1 A review of Voronoï diagram, existence of a bounded cell

Let $\Gamma = \{x_1, \ldots, x_N\}$ be a quantizer of size N. The *Voronoï cell* generated by $x_i \in \Gamma$ is defined by

$$V_{x_i}(\Gamma) = \big\{\xi \in \mathbb{R}^d : |\xi - x_i| = \min_{1\leq j\leq N} |\xi - x_j|\,\big\}, \tag{2.2.1}$$

and $\big(V_{x_i}(\Gamma)\big)_{1\leq i\leq N}$ is called the *Voronoï diagram* of Γ, which is a finite covering of $\mathbb{R}^d$ (see Graf and Luschgy (2000)). A Borel measure partition $\big(C_{x_i}(\Gamma)\big)_{1\leq i\leq N}$ is called a *Voronoï partition* of $\mathbb{R}^d$ induced by Γ if for every $i \in \{1,\ldots,N\}$, $C_{x_i}(\Gamma) \subset V_{x_i}(\Gamma)$. We also define the *open Voronoï cell* generated by $x_i \in \Gamma$ by

$$V^o_{x_i}(\Gamma) = \big\{\xi \in \mathbb{R}^d : |\xi - x_i| < \min_{1\leq j\leq N, j\neq i} |\xi - x_j|\,\big\}. \tag{2.2.2}$$

(1) By approximate identity we mean $\varphi_\varepsilon \in L^1\big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda_d\big)$, $\varepsilon > 0$, such that $\int_{\mathbb{R}^d} \varphi_\varepsilon d\lambda_d = 1$, $\sup_{\varepsilon>0} \int_{\mathbb{R}^d} |\varphi_\varepsilon|\, d\lambda_d < +\infty$ and $\lim_{\varepsilon\to 0} \int_{\{|\xi|>\eta\}} \varphi_\varepsilon(\xi)\lambda_d(\xi) = 0$ for every $\eta > 0$.

If the norm $|\cdot|$ on $\mathbb{R}^d$ is strictly convex, we have $\mathring{V}_{x_i}(\Gamma) = V^o_{x_i}(\Gamma)$ and $\overline{V^o_{x_i}(\Gamma)} = V_{x_i}(\Gamma)$, where $\mathring{A}$ and $\overline{A}$ denote the interior and the closure of A. Examples of strictly convex norms are the isotropic ℓ_r-norms for $1 < r < +\infty$ defined by

$$\Big|(a^1, \ldots, a^d)\Big|_r = \big(\big|a^1\big|^r + \cdots + \big|a^d\big|^r \big)^{1/r}.$$

However, this is not true for any norm on $\mathbb{R}^d$, typically not for the ℓ^1-norm (see Graf and Luschgy (2000)[Figure 1.2]) or the $\ell^\infty-$norm.

We recall that $A \subset \mathbb{R}^d$ is star-shaped with respect to $a \in A$ if for every $b \in A$ and any $\lambda \in [0,1]$, $a + \lambda(b-a) \in A$.

Proposition 2.2.1. *(see Graf and Luschgy (2000)[Proposition 1.2]) Let $\Gamma = \{x_1, \ldots, x_N\}$ be a quantizer of size $N \geq 1$. For every $i \in \{1, \ldots, N\}$, $V_{x_i}(\Gamma)$ and $V^o_{x_i}(\Gamma)$ are star-shaped relative to x_i.*

Now we discuss a sufficient condition to obtain a Voronoï diagram containing a bounded open Voronoï cell. The first result in this direction is a rewriting Proposition 1.10 in Graf and Luschgy (2000) for Euclidean norms (stated here in view of our applications).

Proposition 2.2.2 ($|\cdot|$ Euclidean norm)**.** *Let $(b_1, \ldots, b_{d+1})$ be an affine basis of $\mathbb{R}^d$ and let $b_0 \in \mathrm{Conv}(\{b_1, \ldots, b_{d+1}\})^{\circ} \neq \varnothing$. Set $\Gamma = \{0, b_1 - b_0, \ldots, b_{d+1} - b_0\}$. Then, the open Voronoï cell $V^o_0(\Gamma)$ generated by 0 is bounded.*

Let us provide now a geometrical criterion for a general norm $|\cdot|$ on $\mathbb{R}^d$, let $\bar{B}_{|\cdot|}(x,r)$ denote the closed ball centered at x with radius r and let $S_{|\cdot|}(x,r)$ denote its sphere.

Proposition 2.2.3. *Let $a_1, \ldots, a_k \in S_{|\cdot|}(0,1)$ such that $S_{|\cdot|}(0,1) \subset \bigcup_{i=1}^k \bar{B}_{|\cdot|}(a_i, 1)$ (such a covering exists since $S_{|\cdot|}(0,1)$ is compact). If we choose $\Gamma = \{0, a_1, \ldots, a_k\}$, then the Voronoï open set $V^o_0(\Gamma) \subset \bar{B}_{|\cdot|}(0,1)$ and $\lambda_d\big(V^o_0(\Gamma)\big) > 0$.*

Proof. As $S_{|\cdot|}(0,1) \subset \bigcup_{i=1}^k \bar{B}_{|\cdot|}(a_i,1)$, for every $\xi \in S_{|\cdot|}(0,1)$, there exists $j \in \{1, \ldots, k\}$ such that $|\xi - a_j| \leq 1 = |\xi|$. If $\Gamma = \{0, a_1, \ldots, a_k\}$, then

$$\forall \xi \in S_{|\cdot|}(0,1), \quad \exists\, j \in \{1, \ldots, k\} \text{ such that } \xi \in V_{a_j}(\Gamma). \tag{2.2.3}$$

Assume that there exists $\xi \in V^o_0(\Gamma) \setminus \bar{B}_{|\cdot|}(0,1)$. Since $V^o_0(\Gamma)$ is star-shaped relatively to 0 and $\frac{1}{|\xi|} \in (0,1)$, we have $\frac{\xi}{|\xi|} \in S_{|\cdot|}(0,1) \cap V^o_0(\Gamma)$. This contradicts (2.2.3) since $V^o_0(\Gamma) \cap V_{a_j}(\Gamma) \neq \varnothing$, $j = 1, \ldots, k$. Consequently, $V^o_0(\Gamma) \subset \bar{B}_{|\cdot|}(0,1)$. Finally, $V^o_0(\Gamma)$ is an open set containing 0, therefore, $\lambda_d\big(V^o_0(\Gamma)\big) > 0$. □

The idea of the above proposition is to cover the unit sphere centered at the origin by a finite number of unit balls centered on the unit sphere. This leads us to introduce the following definition.

Definition 2.2.1. *We define the* minimal sphere covering number $c(d, |\cdot|)$ *as follows,*

$$c(d, |\cdot|) \coloneqq \min\left\{k : \exists\{a_1, \dots, a_k\} \subset S_{|\cdot|}(0,1) \text{ such that } S_{|\cdot|}(0,1) \subset \bigcup_{i=1}^{k} \bar{B}_{|\cdot|}(a_i, 1)\right\} < +\infty.$$

The index $c(d, |\cdot|)$ is finite since the unit sphere is a compact set in $\mathbb{R}^d$. Among all the possible norms, we will focus on the isotropic ℓ_r-norms on $\mathbb{R}^d$. We show some examples of the minimal covering number $c(d, |\cdot|_r)$ in the following proposition (whose proof is postponed to Appendix).

Proposition 2.2.4. *(i)* $c(1, |\cdot|) = 2$, *where* $|\cdot|$ *denotes the absolute value.*

(ii) $c(2, |\cdot|_1) = 2$ *and* $c(2, |\cdot|_r) = 3$ *for every* $1 < r < +\infty$.

(iii) $c(d, |\cdot|_\infty) = 2$ *for every dimension* d.

(iv) Let $r \geq 1$ *such that* $2^r \geq d$, *then* $c(d, |\cdot|_r) \leq 2d$.

2.2.2 A general condition for the probability measure characterization

Let $\Gamma = \{x_1, \dots, x_N\}$ be a quantizer in which there exists at least an $x_{i_0} \in \Gamma$ such that the open Voronoï cell $V^o_{x_{i_0}}(\Gamma)$ is bounded and non-empty. Based on such a quantizer, one can construct an approximate identity as follows. Let $\varphi : \mathbb{R}^d \to \mathbb{R}_+$ be the function defined by

$$\varphi(\xi) = \min_{a \in \Gamma \setminus \{x_{i_0}\}} |\xi - a|^p - \min_{a \in \Gamma} |\xi - a|^p .$$

The function φ is clearly nonnegative, continuous and $\{\varphi > 0\} = V^o_{x_{i_0}}(\Gamma)$ so that $\mathrm{supp}(\varphi) = \overline{V^o_{x_{i_0}}(\Gamma)}$ is compact. Hence, $\int \varphi \, d\lambda_d \in (0, +\infty)$ since $\varphi(x_{i_0}) = d\big(x_{i_0}, \Gamma \setminus \{x_{i_0}\}\big) > 0$ and we can normalize φ by setting $\varphi_1(\xi) \coloneqq \frac{\varphi(x_{i_0}+\xi)}{\int \varphi d\lambda_d}$. For every $\varepsilon > 0$, we define $\varphi_\varepsilon(\xi) \coloneqq \frac{1}{\varepsilon^d}\varphi_1\left(\frac{\xi}{\varepsilon}\right)$, then $(\varphi_\varepsilon)_{\varepsilon>0}$ is clearly an approximate identity (see Grafakos (2014)[Section 1.2.4]).

The following theorem gives conditions on the L^p-quantization error function to characterize a probability measure.

Theorem 2.2.1 (Static characterization)**.** *Let $p \in [1, +\infty)$, let $|\cdot|$ be a norm on $\mathbb{R}^d$ and let $N \geq c(d, |\cdot|) + 1$, or $N \geq d + 2$ if $|\cdot|$ is Euclidean. Then, the answer to Question 1 is positive i.e. if there exists a constant C such that $e^p_{N,p}(\mu, \cdot) = e^p_{N,p}(\nu, \cdot) + C$, $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, then $\mu = \nu$. The constant C is a posteriori 0.*

Proof. Following Proposition 2.2.2 and 2.2.3, we choose a quantizer $\Gamma = \{0, a_1, \ldots, a_{N-1}\}$ such that $V_0^o(\Gamma)$ is bounded and $\lambda_d\big(V_0^o(\Gamma)\big) > 0$. We define $\varphi : \mathbb{R}^d \to \mathbb{R}_+$, by

$$\varphi(\xi) = \min_{a \in \Gamma \setminus \{0\}} |\xi - a|^p - \min_{a \in \Gamma} |\xi - a|^p = \big(\min_{a \in \Gamma \setminus \{0\}} |\xi - a|^p - |\xi|^p \big)_+$$

and $(\varphi_\varepsilon)_{\varepsilon > 0}$ by $\varphi_\varepsilon(\xi) := \frac{1}{C_\varphi \varepsilon^d} \varphi\Big(\frac{\xi}{\varepsilon}\Big)$, where $C_\varphi = \int \varphi \, d\lambda_d$. For any $x \in \mathbb{R}^d$,

$$\begin{aligned}
\varphi_\varepsilon * \mu(x) &= \int_{\mathbb{R}^d} \varphi_\varepsilon(x - \xi) \mu(d\xi) = \int_{\mathbb{R}^d} \frac{1}{\varepsilon^d} \frac{\varphi(\frac{x-\xi}{\varepsilon})}{\int \varphi d\lambda_d} \mu(d\xi) \\
&= \frac{1}{C_\varphi \varepsilon^d} \int_{\mathbb{R}^d} \Big(\min_{a \in \Gamma \setminus \{0\}} \Big| \frac{x - \xi}{\varepsilon} - a \Big|^p - \min_{a \in \Gamma} \Big| \frac{x - \xi}{\varepsilon} - a \Big|^p \Big) \mu(d\xi) \\
&= \frac{1}{C_\varphi \varepsilon^{d+p}} \Big[\int_{\mathbb{R}^d} \min_{a \in \Gamma \setminus \{0\}} |x - \varepsilon a - \xi|^p \, \mu(d\xi) - \int_{\mathbb{R}^d} \min_{a \in \Gamma} |x - \varepsilon a - \xi|^p \, \mu(d\xi) \Big].
\end{aligned}$$

If we define two N-tuples $\tilde{x}$ and $\tilde{x}_0$ as $\tilde{x} = (x - \varepsilon a_1, x - \varepsilon a_1, x - \varepsilon a_2, \ldots, x - \varepsilon a_{N-1})$ and $\tilde{x}_0 = (x, x - \varepsilon a_1, x - \varepsilon a_2, \ldots, x - \varepsilon a_{N-1})$, then

$$\int_{\mathbb{R}^d} \min_{a \in \Gamma \setminus \{0\}} |x - \varepsilon a - \xi|^p \, \mu(d\xi) = e^p_{N,p}(\mu, \tilde{x}) \text{ and } \int_{\mathbb{R}^d} \min_{a \in \Gamma} |x - \varepsilon a - \xi|^p \, \mu(d\xi) = e^p_{N,p}(\mu, \tilde{x}_0).$$

Hence, $\varphi_\varepsilon * \mu(x) = \frac{1}{C_\varphi \varepsilon^{d+p}} \big(e^p_{N,p}(\mu, \tilde{x}) - e^p_{N,p}(\mu, \tilde{x}_0) \big)$.

The assumption $e^p_{N,p}(\mu, \cdot) = e^p_{N,p}(\nu, \cdot) + C$ implies that

$$e^p_{N,p}(\mu, \tilde{x}) - e^p_{N,p}(\mu, \tilde{x}_0) = e^p_{N,p}(\nu, \tilde{x}) - e^p_{N,p}(\nu, \tilde{x}_0),$$

so that, for every $x \in \mathbb{R}^d$ and every $\varepsilon > 0$, $\varphi_\varepsilon * \mu(x) = \varphi_\varepsilon * \nu(x)$.

One can finally conclude that $\mu = \nu$ by letting $\varepsilon \to 0$ since $(\varphi_\varepsilon)_{\varepsilon > 0}$ is an approximate identity (see Rudin (1991)[Theorem 6.32]). Hence $C = 0$. □

The following theorem shows that the pointwise convergence of the L^p-mean quantization error function is a necessary and sufficient condition for $\mathcal{W}_p$-convergence of probability distributions in $\mathcal{P}_p(\mathbb{R}^d)$.

Theorem 2.2.2 ($\mathcal{W}_p$-convergence characterization)**.** *Let $p \in [1, +\infty)$ and let $|\cdot|$ be any norm on $\mathbb{R}^d$. Let $\mu_n \in \mathcal{P}_p(\mathbb{R}^d)$ for $n \in \mathbb{N}^* \cup \{\infty\}$. The following properties are equivalent:*

(*i*) $\mathcal{W}_p(\mu_n, \mu_\infty) \xrightarrow{n\to+\infty} 0$,

(*ii*) $\forall N \geq 1, \quad e_{N,p}(\mu_n, \cdot) \xrightarrow{n\to+\infty} e_{N,p}(\mu_\infty, \cdot)$ *uniformly on* $\mathbb{R}^d$,

(*iii*) $\exists\, N \geq c(d, |\cdot|) + 1$ *or* $N \geq d+2$ *if* $|\cdot|$ *is Euclidean such that,* $e_{N,p}(\mu_n, \cdot) \xrightarrow{n\to+\infty} e_{N,p}(\mu_\infty, \cdot)$ *pointwise on* $\mathbb{R}^d$.

Proof of Theorem 2.2.2. (*i*) $\Rightarrow$ (*ii*) is obvious from (2.1.9).

(*ii*) $\Rightarrow$ (*iii*) is obvious.

(*iii*) $\Rightarrow$ (*i*) First of all, it follows from the convergence $e_{N,p}(\mu_n, \cdot) \xrightarrow{n\to+\infty} e_{N,p}(\mu_\infty, \cdot)$ that

$$e^p_{N,p}(\mu_n, \mathbf{0}) \xrightarrow{n\to+\infty} e^p_{N,p}(\mu_\infty, \mathbf{0}) \text{ i.e. } \int_{\mathbb{R}^d} |\xi|^p \, \mu_n(d\xi) \xrightarrow{n\to+\infty} \int_{\mathbb{R}^d} |\xi|^p \, \mu_\infty(d\xi) < +\infty, \tag{2.2.4}$$

where $\mathbf{0} = (0, \ldots, 0)$. In particular, the sequence $\left(\int_{\mathbb{R}^d} |\xi|^p \, \mu_n(d\xi) \right)_{n\geq 1}$ is bounded. Hence, the sequence of probability measures $(\mu_n)_{n\geq 1}$ is tight.

Let $\widetilde{\mu}_\infty$ be a weak limiting probability distribution of $(\mu_n)_{n\geq 1}$ i.e. there exists a subsequence $\alpha(n)$ of n such that $\mu_{\alpha(n)} \xRightarrow{(\mathbb{R}^d)} \widetilde{\mu}_\infty$ as $n \to +\infty$.

Let $x = (x_1, \ldots, x_N)$ be any N-tuple in $(\mathbb{R}^d)^N$. We define a continuous function $f_x : \mathbb{R}^d \to \mathbb{R}$ by

$$f_x(\xi) := \min_{1\leq i\leq N} |\xi - x_i|^p - |\xi|^p .$$

Hence, owing to the elementary inequality $v^p - u^p \leq pv^{p-1}(v-u)$ for any $0 \leq u \leq v < +\infty$, we derive

$$\big|f_x(\xi)\big| \leq \max_{i\in\{1,\ldots,N\}} p\big(\, |\xi| + |x_i| \,\big)^{p-1} |x_i| \leq C_{x,p}(1 + |\xi|^{p-1}), \tag{2.2.5}$$

where $C_{x,p}$ is a constant depending on x and p.

Owing to (2.2.4) and (2.2.5), the sequence $\left(\int f_x^{\frac{p}{p-1}} d\mu_n \right)_{n\geq 1}$ is bounded, hence f_x is uniformly integrable with respect to $(\mu_n)_{n\geq 1}$ since $\frac{p}{p-1} > 1$, so that f_x is uniformly integrable with respect to any subsequence $(\mu_{\alpha(n)})_{n\geq 1}$. It follows that

$$\int_{\mathbb{R}^d} f_x(\xi) \mu_{\alpha(n)}(d\xi) \to \int_{\mathbb{R}^d} f_x(\xi) \widetilde{\mu}_\infty(d\xi),$$

as $n \to +\infty$, where

$$\int_{\mathbb{R}^d} f_x(\xi)\mu_{\alpha(n)}(d\xi) = \int_{\mathbb{R}^d} \big(\min_{i\in\{1,\dots,N\}} |\xi - x_i|^p - |\xi|^p \big)\mu_{\alpha(n)}(d\xi)$$
$$= e^p_{N,p}(\mu_{\alpha(n)}, x) - e^p_{N,p}(\mu_{\alpha(n)}, \mathbf{0}),$$
$$\text{and } \int_{\mathbb{R}^d} f_x(\xi)\widetilde{\mu}_\infty(d\xi) = e^p_{N,p}(\widetilde{\mu}_\infty, x) - e^p_{N,p}(\widetilde{\mu}_\infty, \mathbf{0}).$$

On the other hand, $e^p_{N,p}(\mu_{\alpha(n)}, x) - e^p_{N,p}(\mu_{\alpha(n)}, \mathbf{0})$ converges to $e^p_{N,p}(\mu_\infty, x) - e^p_{N,p}(\mu_\infty, \mathbf{0})$ owing to the pointwise convergence in (iii) at $\mathbf{0} = (0, \dots, 0)$ and $x = (x_1, \dots, x_N)$.

Therefore,

$$e^p_{N,p}(\widetilde{\mu}_\infty, x) - e^p_{N,p}(\widetilde{\mu}_\infty, \mathbf{0}) = e^p_{N,p}(\mu_\infty, x) - e^p_{N,p}(\mu_\infty, \mathbf{0}),$$

which implies that for every $x \in (\mathbb{R}^d)^N$,

$$e^p_{N,p}(\widetilde{\mu}_\infty, x) - e^p_{N,p}(\mu_\infty, x) = C,$$

where $C = e^p_{N,p}(\widetilde{\mu}_\infty, \mathbf{0}) - e^p_{N,p}(\mu_\infty, \mathbf{0})$ is a real constant. It follows from Theorem 2.2.1 that $\widetilde{\mu}_\infty = \mu_\infty$, which implies that μ_∞ is the only limiting distribution of $(\mu_n)_{n\geq 1}$ for the weak convergence and consequently $\mu_n \xRightarrow{(\mathbb{R}^d)} \mu$. We have already proved that

$$\int_{\mathbb{R}^d} |\xi|^p \, \mu_n(d\xi) \xrightarrow{n\to+\infty} \int_{\mathbb{R}^d} |\xi|^p \, \mu_\infty(d\xi)$$

from (2.2.4), which finally shows that $\mathcal{W}_p(\mu_n, \mu_\infty) \xrightarrow{n\to+\infty} 0$ owing to Theorem 2.1.1. □

A careful reading of the proof shows that the following "à la Paul Lévy" characterization result holds for limiting functions of L^p-quantization error functions.

Corollary 2.2.1. *Let $p \in [1 + \infty)$. Let $(\mu_n)_{n\geq 1}$ be a $\mathcal{P}_p(\mathbb{R}^d)$-valued sequence. If*

$e_{N,p}(\mu_n, \cdot) \xrightarrow{n\to+\infty} f$ *pointwise for some N such that* static characterization *holds true (Question 1), then there exists $\mu_\infty \in \mathcal{P}_p(\mathbb{R}^d)$ such that $\mu_n \xRightarrow{(\mathbb{R}^d)} \mu_\infty$ as $n \to +\infty$ and*

$$f^p = e^p_{N,p}(\mu_\infty, \cdot) + \lim_n \int_{\mathbb{R}^d} |\xi|^p \, \mu_n(d\xi) - \int_{\mathbb{R}^d} |\xi|^p \, \mu_\infty(d\xi).$$

Now we will take advantage of what precedes to introduce a *quantization based distance* on $\mathcal{P}_p(\mathbb{R}^d)$. Let $\mathcal{C}_b\big((\mathbb{R}^d)^N, \mathbb{R}\big)$ denote the space of bounded $\mathbb{R}$-valued continuous functions

defined on $(\mathbb{R}^d)^N$ equipped with the sup norm $\|\cdot\|_{\sup}$. Let $p \in [1, +\infty)$. If $\mu \in \mathcal{P}_p(\mathbb{R}^d)$,

$$e_{N,p}(\mu, \cdot) - e_{N,p}(\delta_0, \cdot) \in \mathcal{C}_b\big((\mathbb{R}^d)^N, \mathbb{R}\big)$$

(note that $e_{N,p}\big(\delta_0, (x_1, \ldots, x_N)\big) = \min_{i=1,\ldots,N} |x_i|$) since inequality (2.1.8) implies that

$$\|e_{N,p}(\mu, \cdot) - e_{N,p}(\delta_0, \cdot)\|_{\sup} \leq \mathcal{W}_p(\mu, \delta_0) = \Big[\int_{\mathbb{R}^d} |\xi|^p \, \mu(d\xi)\Big]^{1/p} < +\infty.$$

Then, we define a function $\mathcal{Q}_{N,p}$ on $\mathcal{P}_p(\mathbb{R}^d)$ by

$$\begin{aligned}(\mu, \nu) \longmapsto \mathcal{Q}_{N,p}(\mu, \nu) &:= \big\| \big(e_{N,p}(\mu, \cdot) - e_{N,p}(\delta_0, \cdot)\big) - \big(e_{N,p}(\nu, \cdot) - e_{N,p}(\delta_0, \cdot)\big) \big\|_{\sup} \\ &= \|e_{N,p}(\mu, \cdot) - e_{N,p}(\nu, \cdot)\|_{\sup}. \end{aligned} \tag{2.2.6}$$

For any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, inequality (2.1.8) implies $\mathcal{Q}_{N,p}(\mu, \nu) \leq \mathcal{W}_p(\mu, \nu) < +\infty$ so that $\mathcal{Q}_{N,p}(\mu, \nu) \in [0, +\infty)$. Combining Theorems 2.2.1 and 2.2.2 implies the following result.

Corollary 2.2.2. *Let $p \in [1, +\infty)$.*

(a) $N_{d,p,|\cdot|} \leq c(d, |\cdot|) + 1$ for any norm and $N_{d,p,|\cdot|} \leq d + 2$ if $|\cdot|$ is Euclidean.

(b) If $N \geq c(d, |\cdot|) + 1$ or $N \geq d + 2$ if $|\cdot|$ is Euclidean, then $\mathcal{Q}_{N,p}$ defined by (2.2.6) is a distance on $\mathcal{P}_p(\mathbb{R}^d)$ and $\mathcal{Q}_{N,p}$ is topologically equivalent to the Wasserstein distance $\mathcal{W}_p$.

Comments on optimality. If we consider only the quadratic case $p = 2$ and a norm $|\cdot|$ induced by an inner product, the result in Corollary 2.2.2-(a) is in fact not optimal. In the next section, we will prove that in such a setting, $N_{d,2,|\cdot|} = 2$ and this result can also be extended to any separable (possibly infinite-dimensional) Hilbert space.

2.3 Quadratic quantization based characterization on a separable Hilbert space

Let H denote a separable Hilbert space with the inner product $(\cdot \,|\, \cdot)_H$. Let $|\cdot|_H$ denote the norm on H induced by $(\cdot \,|\, \cdot)_H$. When there is no ambiguity, we drop the index $_H$ and write $(\cdot \,|\, \cdot)$ and $|\cdot|$. The separable Hilbert space is a very common setup for applications, for example in functional data analysis: one can set $H = L^2\big([0, T], dt\big)$ and $X = (X_t)_{t \in [0,T]}$ a bi-measurable process such that $\int_0^T \mathbb{E} X_t^2 \, dt < +\infty$. For more information about functional data analysis with an L^2-setup, we refer to Hsing and Eubank (2015) among others.

We first prove in the quadratic case ($p = 2$), that both static (see further Propo-

sition 2.3.1) and $\mathcal{W}_2$-convergence (see further Theorem 2.3.1) characterizations can be obtained at level $N = 2$ by an analytical method. Then we will show that $N_{H,2} := N_{H,2,|\cdot|_H} = 2$ and for any $\mu, \nu \in \mathcal{P}_2(H)$, $\mathcal{Q}_{2,2}(\mu,\nu) := \|e_{2,2}(\mu,\cdot) - e_{2,2}(\nu,\cdot)\|_{\sup}$ is a well-defined distance on $\mathcal{P}_2(H)$ which is topologically equivalent to $\mathcal{W}_2$.

Proofs of quadratic quantization based characterizations rely on the following lemma.

Lemma 2.3.1. *(a) Let $\mu, \nu \in \mathcal{P}_2(H)$. If for every $u \in H, |u| = 1$,*

$$\mu \circ \big(\xi \mapsto (\xi \mid u)\big)^{-1} = \nu \circ \big(\xi \mapsto (\xi \mid u)\big)^{-1},$$

then $\mu = \nu$.
(b) Let $\mu_n \in \mathcal{P}_2(H)$ for every $n \in \mathbb{N}^* \cup \{\infty\}$. If $\int_H |\xi|^2 \mu_n(d\xi) \xrightarrow{n\to+\infty} \int_H |\xi|^2 \mu_\infty(d\xi)$ and for every $u \in H$, $|u| = 1$,

$$\mu_n \circ \big(\xi \mapsto (\xi \mid u)\big)^{-1} \overset{(\mathbb{R})}{\Longrightarrow} \mu_\infty \circ \big(\xi \mapsto (\xi \mid u)\big)^{-1},$$

then $\mathcal{W}_2(\mu_n, \mu_\infty) \to 0$.

Proof. As $(H, |\cdot|)$ is separable, let $(h_k)_{k\geq 1}$ be a countable orthonormal basis of $(H, |\cdot|)$.

(a) Let X, Y be random variables with respective distributions μ and ν and let $\lambda \in H$. We define for every $m \geq 1$, $X^{(m)} := \sum_{k=1}^m (X \mid h_k) h_k$, $Y^{(m)} := \sum_{k=1}^m (Y \mid h_k) h_k$ and $\lambda^{(m)} := \sum_{k=1}^m (\lambda \mid h_k) h_k$. For $m \geq 1$, let $u^{(m)} = \frac{\lambda^{(m)}}{|\lambda^{(m)}|}$ (convention $\frac{0}{|0|} = 0$), then we have

$$(\lambda \mid X^{(m)}) = \sum_{k=1}^{+\infty} (\lambda \mid h_k)(X^{(m)} \mid h_k) = \sum_{k=1}^{m} (\lambda \mid h_k)(X \mid h_k) = \big|\lambda^{(m)}\big| \big(X \,\big|\, u^{(m)}\big).$$

Similarly, $(\lambda \mid Y^{(m)}) = \big|\lambda^{(m)}\big| \big(Y \,\big|\, u^{(m)}\big)$. Let i be such that $i^2 = -1$. It follows that

$$\begin{aligned}\mathbb{E}\, e^{i(\lambda|X^{(m)})} = \mathbb{E}\, e^{i|\lambda^{(m)}|(X|u^{(m)})} &= \int_H e^{i\,|\lambda^{(m)}|\,\xi} \mu \circ \big(\xi \mapsto (u^{(m)} \mid \xi)\big)^{-1}(d\xi) \\ &= \int_H e^{i\,|\lambda^{(m)}|\,\xi} \nu \circ \big(\xi \mapsto (u^{(m)} \mid \xi)\big)^{-1}(d\xi) = \mathbb{E}\, e^{i(\lambda|Y^{(m)})}.\end{aligned}$$

Since we can arbitrarily choose λ, we have for every $m \geq 1$, $\mathrm{Law}(X^{(m)}) = \mathrm{Law}(Y^{(m)})$. Let $F : H \to \mathbb{R}$ be a bounded continuous function. Then, for every $m \geq 1$,

$$\mathbb{E}\, F(X^{(m)}) = \mathbb{E}\, F(Y^{(m)})$$

which implies $\mathbb{E}\, F(X) = \mathbb{E}\, F(Y)$ by letting $m \to +\infty$. Hence, $\mu = \nu$.

(b) For every $n \geq 1$, let X_n be random variables with distribution μ_n and let X_∞ be a

random variable with distribution μ_∞. We define for every $n \geq 1$ and for every $m \geq 1$,

$$X_n^{(m)} := \sum_{i=1}^{m} (X_n|h_i)h_i \text{ and } X_\infty^{(m)} := \sum_{i=1}^{m} (X_\infty|h_i)h_i.$$

Following the lines of item (a), we get for every $m \geq 1$, $X_n^{(m)} \overset{(H)}{\Longrightarrow} X_\infty^{(m)}$ as $n \to +\infty$, since the convergence of characteristic function implies weak convergence.

Now, let $F : H \to \mathbb{R}$ be a Lipschitz continuous function with Lipschitz coefficient $[F]_{\mathrm{Lip}} := \sup_{x,y\in H} \frac{|F(x)-F(y)|}{|x-y|}$. For every (temporarily) fixed $m \geq 1$,

$$\begin{aligned}
&\lim_n \left|\mathbb{E}\, F(X_n) - \mathbb{E}\, F(X_\infty)\right| \\
&\quad \leq \lim_n \left|\mathbb{E}\, F(X_n) - \mathbb{E}\, F(X_n^{(m)})\right| + \lim_n \left|\mathbb{E}\, F(X_n^{(m)}) - \mathbb{E}\, F(X_\infty^{(m)})\right| + \left|\mathbb{E}\, F(X_\infty^{(m)}) - \mathbb{E}\, F(X_\infty)\right| \\
&\quad \leq \lim_n \left|\mathbb{E}\, F(X_n) - \mathbb{E}\, F(X_n^{(m)})\right| + 0 + \left|\mathbb{E}\, F(X_\infty^{(m)}) - \mathbb{E}\, F(X_\infty)\right| \quad (\text{since } X_n^{(m)} \overset{(H)}{\Longrightarrow} X_\infty^{(m)}).
\end{aligned}$$

Then, for every $n \geq 1$,

$$\begin{aligned}
\left|\mathbb{E}\, F(X_n) - \mathbb{E}\, F(X_n^{(m)})\right| \leq \mathbb{E}\left|F(X_n) - F(X_n^{(m)})\right| &\leq [F]_{\mathrm{Lip}} \mathbb{E}\left|X_n - X_n^{(m)}\right| \\
&\leq [F]_{\mathrm{Lip}} \left\|X_n - X_n^{(m)}\right\|_2.
\end{aligned}$$

Similarly, we also have $\left|\mathbb{E}\, F(X_\infty^{(m)}) - \mathbb{E}\, F(X_\infty)\right| \leq [F]_{\mathrm{Lip}} \left\|X_\infty - X_\infty^{(m)}\right\|_2$.

It follows from Fatou's Lemma for the weak convergence and the convergence assumption made on $\mathbb{E}|X_n|^2$ that

$$\begin{aligned}
\limsup_n \left\|X_n - X_n^{(m)}\right\|_2^2 = \limsup_n \mathbb{E}\left|X_n - X_n^{(m)}\right|^2 &= \limsup_n \left[\mathbb{E}\left|X_n\right|^2 - \mathbb{E}\left|X_n^{(m)}\right|^2\right] \\
&= \mathbb{E}\left|X_\infty\right|^2 - \liminf_n \mathbb{E}\left|X_n^{(m)}\right|^2 \leq \mathbb{E}\left|X_\infty\right|^2 - \mathbb{E}\left|X_\infty^{(m)}\right|^2 \\
&= \left\|X_\infty - X_\infty^{(m)}\right\|_2^2.
\end{aligned}$$

Hence, for every $m \geq 1$,

$$\begin{aligned}
\lim_n \left|\mathbb{E}\, F(X_n) - \mathbb{E}\, F(X_\infty)\right| &\leq \limsup_n [F]_{\mathrm{Lip}} \left\|X_n - X_n^{(m)}\right\|_2 + [F]_{\mathrm{Lip}} \left\|X_\infty - X_\infty^{(m)}\right\|_2 \\
&\leq 2[F]_{\mathrm{Lip}} \left\|X_\infty - X_\infty^{(m)}\right\|_2.
\end{aligned}$$

Then,

$$\left\|X_\infty - X_\infty^{(m)}\right\|_2 \to 0 \text{ as } m \to +\infty$$

by the Lebesgue dominated convergence theorem since $\left|X_\infty - X_\infty^{(m)}\right| \leq \left|X_\infty\right| \in L^2(\mathbb{P})$ so that $\mathbb{E}\, F(X_n) \to \mathbb{E}\, F(X_\infty)$ as $n \to +\infty$. Thus, $X_n \overset{(H)}{\Longrightarrow} X_\infty$ and we can conclude that $\mathcal{W}_p(\mu_n, \mu_\infty) \to 0$ by applying Theorem 2.1.1. □

Proposition 2.3.1 (Static characterization)**.** *Let* $\mu, \nu \in \mathcal{P}_2(H)$*. If*

$$e_{2,2}^2(\mu, \cdot) = e_{2,2}^2(\nu, \cdot) + C$$

for some real constant C*, then* $\mu = \nu$ *and* $C = 0$*.*

Proof. Let $a, b \in H$, then $e_{2,2}^2\big(\mu, (a,b)\big) = \int_H |\xi - a|^2 \wedge |\xi - b|^2 \, \mu(d\xi)$.

As $e_{2,2}^2\big(\mu, (a,b)\big) = e_{2,2}^2\big(\nu, (a,b)\big) + C$ for every $(a,b) \in H^2$, in particular, if $a = b$, $\int_H |\xi - a|^2 \, \mu(d\xi) = \int_H |\xi - a|^2 \, \nu(d\xi) + C$. Hence, using that $(x-y)_+ = x - x \wedge y$, we have

$$\forall a, b \in H, \quad \int_H \big(|\xi - a|^2 - |\xi - b|^2 \big)_+ \mu(d\xi) = \int_H \big(|\xi - a|^2 - |\xi - b|^2 \big)_+ \nu(d\xi). \quad (2.3.1)$$

Note that $|\xi - a|^2 - |\xi - b|^2 = 2\Big(b - a \,\Big|\, \xi - \frac{a+b}{2}\Big)$. Hence, if we take $a = \lambda u$ and $b = \lambda' u$ with $\lambda, \lambda' \in \mathbb{R}$, $\lambda' > \lambda$ for some common $u \in H$ with $|u| = 1$, we obtain

$$\big(|\xi - a|^2 - |\xi - b|^2 \big)_+ = 2(\lambda' - \lambda) \left((\xi \mid u) - \frac{\lambda + \lambda'}{2} \right)_+ .$$

As a consequence of (2.3.1), we derive that

$$\forall \lambda, \lambda' \in \mathbb{R}, \lambda' > \lambda, \quad \int_H \left((\xi \mid u) - \frac{\lambda + \lambda'}{2} \right)_+ \mu(d\xi) = \int_H \left((\xi \mid u) - \frac{\lambda + \lambda'}{2} \right)_+ \nu(d\xi).$$

In turn, this implies, by letting $\lambda' \to \lambda$,

$$\forall u \in H, |u| = 1, \ \forall \lambda \in \mathbb{R}, \qquad \int_H \Big((\xi \mid u) - \lambda \Big)_+ \mu(d\xi) = \int_H \Big((\xi \mid u) - \lambda \Big)_+ \nu(d\xi). \quad (2.3.2)$$

The function $\lambda \mapsto \big((\xi \mid u) - \lambda\big)_+$ is right differentiable with $\mathbb{1}_{(\xi|u)>\lambda}$ as a right derivative and μ-integrable. Hence, by the Lebesgue differentiation theorem, we can right differentiate the equality (2.3.2) which yields for every $u \in H, |u| = 1$ and for every $\lambda \in \mathbb{R}$,

$$\mu\big((\xi \mid u) > \lambda\big) = \nu\big((\xi \mid u) > \lambda\big).$$

Hence, for every $u \in H, |u| = 1$, $\mu \circ \big(\xi \mapsto (\xi \mid u)\big)^{-1} = \nu \circ \big(\xi \mapsto (\xi \mid u)\big)^{-1}$ since they have the same survival function. We conclude by Lemma 2.3.1 (a) that $\mu = \nu$ and $C = 0$. □

The following theorem shows the equivalence of $\mathcal{W}_2$-convergence of $(\mu_n)_{n\geq 1}$ in $\mathcal{P}_2(H)$ and the pointwise convergence of quadratic quantization error function $\big(e_{2,2}(\mu_n, \cdot)\big)_{n\geq 1}$.

Theorem 2.3.1 ($\mathcal{W}_2$-convergence characterization)**.** *Let $\mu_n \in \mathcal{P}_2(H)$ for every $n \in \mathbb{N}^* \cup \{\infty\}$. The following properties are equivalent:*

(i) $\mathcal{W}_2(\mu_n, \mu_\infty) \xrightarrow{n\to+\infty} 0$,

(ii) $e_{2,2}(\mu_n, \cdot) \xrightarrow{n\to+\infty} e_{2,2}(\mu_\infty, \cdot)$ *uniformly,*

(iii) $e_{2,2}(\mu_n, \cdot) \xrightarrow{n\to+\infty} e_{2,2}(\mu_\infty, \cdot)$ *pointwise.*

Before proving Theorem 2.3.1, we recall the convergence of left and right derivatives of a converging sequence of convex functions. Let $\partial_- f$ (respectively $\partial_+ f$) denote the left derivative (resp. right derivative) of a convex function f.

Lemma 2.3.2. *(See e.g. Lacković (1982)[Theorems 2.5]) Let $f_n : \mathbb{R} \to \mathbb{R}, n \in \mathbb{N}^*$, be a sequence of convex functions converging pointwise to a function $f : \mathbb{R} \to \mathbb{R}$. Let $G := \{x \in \mathbb{R} \,|\, \partial_- f(x) \neq \partial_+ f(x)\}$. Then for every point $x \in \mathbb{R} \setminus G$,*

$$\lim_n \partial_+ f_n(x) = \lim_n \partial_- f_n(x) = f'(x).$$

Proof of Theorem 2.3.1. $(i) \Rightarrow (ii)$ is obvious from (2.1.9).

$(ii) \Rightarrow (iii)$ is obvious.

$(iii) \Rightarrow (i)$ For every $(a, b) \in H^2$,

$$e_{2,2}^2\big(\mu_n, (a,b)\big) = \int_H |\xi - a|^2 \wedge |\xi - b|^2 \, \mu_n(d\xi) \xrightarrow{n\to+\infty} e_{2,2}^2\big(\mu_\infty, (a,b)\big) = \int_H |\xi - a|^2 \wedge |\xi - b|^2 \mu_\infty(d\xi).$$

In particular, $\forall a \in H$, $\int_H |\xi - a|^2 \, \mu_n(d\xi) \xrightarrow{n\to+\infty} \int_H |\xi - a|^2 \, \mu_\infty(d\xi)$. Hence, using that $(x - y)_+ = x - x \wedge y$, we get

$$\forall a, b \in H, \quad \int_H \big(|\xi - a|^2 - |\xi - b|^2 \big)_+ \mu_n(d\xi) \xrightarrow{n\to+\infty} \int_H \big(|\xi - a|^2 - |\xi - b|^2 \big)_+ \mu_\infty(d\xi).$$

Following the lines of the proof of Proposition 2.3.1, we get

$$\forall \lambda \in \mathbb{R}, \forall u \in H, |u| = 1, \quad \int_H \Big((\xi \mid u) - \lambda \Big)_+ \mu_n(d\xi) \xrightarrow{n\to+\infty} \int_H \Big((\xi \mid u) - \lambda \Big)_+ \mu_\infty(d\xi). \tag{2.3.3}$$

For $\mu \in \mathcal{P}_2(H)$ and $u \in S_{|\cdot|}(0, 1)$, we define the real-valued convex function ϕ_μ by $\phi_\mu : \lambda \mapsto \int \big((\xi \mid u) - \lambda \big)_+ \mu(d\xi)$. It follows from (2.3.3) that $(\phi_{\mu_n})_{n\geq 0}$ converges pointwise to ϕ_{μ_∞}. Moreover, ϕ_{μ_n}, ϕ_{μ_∞} are right-differentiable and their right derivatives are given by

$\partial_+\phi_{\mu_n}(\lambda) = \mu_n\big((\xi \mid u) > \lambda\big)$ and $\partial_+\phi_{\mu_\infty}(\lambda) = \mu_\infty\big((\xi \mid u) > \lambda\big)$ respectively. Note that the functions $1-\partial_+\phi_{\mu_n}$ and $1-\partial_+\phi_{\mu_\infty}$ are the cumulative distribution functions of the probability distributions $\mu_n \circ \big(\xi \mapsto (\xi \mid u)\big)^{-1}$ and $\mu_\infty \circ \big(\xi \mapsto (\xi \mid u)\big)^{-1}$ and that the set of discontinuity points of $1-\partial_+\phi_{\mu_\infty}$ and $\partial_+\phi_{\mu_\infty}$, is $G = \{\lambda : \mu_\infty\big(\{\xi : (\xi \mid u) = \lambda\}\big) > 0\}$.

We know from Lemma 2.3.2 that for every $\lambda \in \mathbb{R} \setminus G$, $\partial_+\phi_{\mu_n}(\lambda) \xrightarrow{n\to+\infty} \partial_+\phi_{\mu_\infty}(\lambda)$ and that $\partial_-\phi_{\mu_\infty}$ is continuous on $\mathbb{R} \setminus G$. Hence

$$\forall u \in H, |u| = 1, \qquad \mu_n \circ \big(\xi \mapsto (\xi \mid u)\big)^{-1} \xRightarrow{(\mathbb{R})} \mu_\infty \circ \big(\xi \mapsto (\xi \mid u)\big)^{-1}. \tag{2.3.4}$$

Moreover, $e_{2,2}\big(\mu_n, (0,0)\big)$ converges to $e_{2,2}\big(\mu_\infty, (0,0)\big)$, which also reads $\int_H |\xi|^2 \mu_n(d\xi) \to \int_H |\xi|^2 \mu_\infty(d\xi)$. Consequently, it follows from Lemma 2.3.1-(b) that $\mathcal{W}_2(\mu_n, \mu_\infty) \to 0$ as $n \to +\infty$. □

Remark 2.3.1. Proposition 2.3.1 and Theorem 2.3.1 directly imply that $N_{H,2,|\cdot|_2} \leq 2$. In fact, for every $a \in H$,

$$e_{1,2}^2(\mu, a) = \int_H |\xi - a|_H^2 \, \mu(d\xi) = \int_H |\xi|_H^2 \, \mu(d\xi) - 2\Big(\int_H \xi\mu(d\xi) \,\Big|\, a\Big)_H + |a|_H^2 .$$

Thus, if $\mu, \nu \in \mathcal{P}_2(H)$ are such that

$$\int_H |\xi|_H^2 \, \mu(d\xi) = \int_H |\xi|_H^2 \, \nu(d\xi) \;\text{ and }\; \int_H \xi\mu(d\xi) = \int_H \xi\nu(d\xi), \tag{2.3.5}$$

then we have $e_{1,2}(\mu, \cdot) = e_{1,2}(\nu, \cdot)$. But condition (2.3.5) is clearly not sufficient to have $\mu = \nu$. Consequently, $N_{H,2,|\cdot|_2} = 2$.

Like what we did in Section 2.2.2, we define a function $\mathcal{Q}_{2,2}^H$ on $\big(\mathcal{P}_2(H)\big)^2$ by

$$(\mu, \nu) \mapsto \mathcal{Q}_{2,2}^H(\mu, \nu) = \|e_{2,2}(\mu, \cdot) - e_{2,2}(\nu, \cdot)\|_{\sup} .$$

Then inequality (2.1.8) implies that $\mathcal{Q}_{2,2}^H(\mu, \nu) \in [0, +\infty)$. Moreover, Proposition 2.3.1 and Theorem 2.3.1 lead to the following corollary.

Corollary 2.3.1. *The distances $\mathcal{Q}_{2,2}^H$ and $\mathcal{W}_2$ are topologically equivalent on $\mathcal{P}_2(H)$.*

We conclude this section by an "à la Paul Lévy" characterization of a limit of quantization error functions.

Theorem 2.3.2 (À la Paul Lévy characterization)**.** *Let $(H, |\cdot|_H)$ be a separable Hilbert space. Let $(\mu_n)_{n\geq 1}$ be a $\mathcal{P}_2(H)$-valued sequence and let $f : H^2 \to \mathbb{R}_+$ be such that*

$$e_{2,2}(\mu_n, \cdot) \xrightarrow{n\to+\infty} f \quad \textit{pointwise.}$$

Then there exists $\mu_\infty \in \mathcal{P}_2(H)$ *such that* $\mu_n \overset{(H_w)}{\Longrightarrow} \mu_\infty$ *(where* (H_w) *stands for the weak topology on* H*) and*

$$f^2 = e_{2,2}(\mu_\infty, \cdot)^2 + \lim_n \int_H |\xi|^2 \, \mu_n(d\xi) - \int_H |\xi|^2 \, \mu_\infty(d\xi).$$

Proof. The sequence $e_{2,2}\big(\mu_n, (0,0)\big)^2 = \int_H |\xi|^2 \mu_n(d\xi)$, $n \geq 1$, is bounded, hence the sequence $(\mu_n)_{n\geq 1}$ is tight for the weak topology (H_w) on H, which generates the same Borel σ-field as the strong one. Consequently there exists a subnet $\mu_{\varphi(n)} \overset{(H_w)}{\Longrightarrow} \mu_\infty \in \mathcal{P}_2(H)$ since the mapping $\xi \mapsto |\xi|^2$ is weakly lower semi-continuous and non-negative (see Topsoe (1974)[Lemma 2.3 and Theorem 3.1] and Kelley (1975)[Chapter 2] for the definition of subnet). Now note that, for a fixed $x = (x_1, x_2) \in H^2$, the mapping

$$\xi \mapsto \min\big(|\xi - x_1|^2, |\xi - x_2|^2\big) - |\xi|^2 = \min\big(|x_1|^2 - 2(x_1|\xi), |x_2|^2 - 2(x_2|\xi)\big)$$

is weakly continuous and $(\mu_n)_{n\geq 1}$-uniformly integrable since it is sublinear. Hence

$$\begin{aligned} e^2_{2,2}(\mu_{\varphi(n)}, x) \longrightarrow \int_H \min\big(|x_1|^2 - 2(x_1|\xi), |x_2|^2 - 2(x_2|\xi)\big)\mu_\infty(d\xi) + f^2\big((0,0)\big) \text{ as } n \to +\infty \\ = e^2_{2,2}(\mu_\infty, x) + f^2\big((0,0)\big) - \int_H |\xi|^2 \, \mu_\infty(d\xi). \end{aligned}$$

For two such limiting distributions μ_∞ and μ'_∞ it follows from what precedes that $e^2_{2,2}(\mu_\infty, \cdot) = e^2_{2,2}(\mu'_\infty, \cdot) + C_\infty$ for some real constant C_∞. Hence $\mu_\infty = \mu'_\infty$ by Proposition 2.3.1, which in turn implies that $\mu_n \overset{(H_w)}{\Longrightarrow} \mu_\infty$. □

2.4 Further quantization based characterizations on $\mathbb{R}$

Let $|\cdot|$ denote the absolute value on $\mathbb{R}$. Results from Section 2.2 (Theorem 2.2.1 and 2.2.2, Proposition 2.2.4-(i)) imply that $N_{1,p} := N_{1,p,|\cdot|} \leq 3$ for any $p \geq 1$. Moreover, Proposition 2.3.1 and Theorem 2.3.1 imply that $N_{1,2} = 2$. Other quantization based characterizations are developed in Section 2.4.1. Then we discuss the completeness of the distance $\mathcal{Q}_{1,1}$ (defined in (2.2.6)) on $\mathcal{P}_1(\mathbb{R})$ and of $\mathcal{Q}_{2,2}$ on $\mathcal{P}_2(\mathbb{R})$ with opposite answers in Section 2.4.2.

2.4.1 Quantization based characterization on $\mathbb{R}$

Proposition 2.4.1 $(p = 1)$**.** (a) *Let* $\mu, \nu \in \mathcal{P}_1(\mathbb{R})$*. If* $e_{1,1}(\mu, \cdot) = e_{1,1}(\nu, \cdot) + C$ *for some real constant* C*, then* $\mu = \nu$ *and* $C = 0$*.*

(*b*) *If* $\mu_n \in \mathcal{P}_1(\mathbb{R})$, $n \in \mathbb{N}^* \cup \{\infty\}$, *the following properties are equivalent:*

(*i*) $\mathcal{W}_1(\mu_n, \mu_\infty) \xrightarrow{n\to+\infty} 0$,

(*ii*) $e_{1,1}(\mu_n, \cdot) \xrightarrow{n\to+\infty} e_{1,1}(\mu_\infty, \cdot)$ *uniformly,*

(*iii*) $e_{1,1}(\mu_n, \cdot) \xrightarrow{n\to+\infty} e_{1,1}(\mu_\infty, \cdot)$ *pointwise.*

(*c*) *The distance* $\mathcal{Q}_{1,1}$ *and* $\mathcal{W}_1$ *are topologically equivalent on* $\mathcal{P}_1(\mathbb{R})$ *and* $N_{1,1} = 1$.

Proof. (*a*) The function $e_{1,1}(\mu, \cdot)$ reads $x \mapsto \int_{\mathbb{R}} |\xi - x|\, \mu(d\xi)$, hence it is convex and its right derivative is given by $x \mapsto -1 + 2\mu\big(]-\infty, x]\big)$. So if $e_{1,1}(\mu, \cdot) = e_{1,1}(\nu, \cdot) + C$, we have $\mu\big(]-\infty, x]\big) = \nu\big(]-\infty, x]\big)$ for all $x \in \mathbb{R}$, which implies $\mu = \nu$ (and $C = 0$).

(*b*) It is obvious that (*i*) $\Rightarrow$ (*ii*) and (*ii*) $\Rightarrow$ (*iii*). Now we prove (*iii*) $\Rightarrow$ (*i*).

For every $n \geq 1$, $e_{1,1}(\mu_n, \cdot)$ can also be written as $a \mapsto \int_{\mathbb{R}} |\xi - a|\, \mu_n(d\xi)$, which is convex with right derivative at a given by $-1 + 2\mu_n\big(]-\infty, a]\big)$. Consequently, if $e_{1,1}(\mu_n, \cdot)$ converges pointwise to $e_{1,1}(\mu_\infty, \cdot)$ on $\mathbb{R}$, then $\mu_n\big(]-\infty, a]\big)$ converges pointwise to $\mu_\infty\big(]-\infty, a]\big)$ for all $a \in \mathbb{R}$ such that $\mu_\infty(\{a\}) = 0$ by Lemma 2.3.2. This implies $\mu_n \xrightarrow{(\mathbb{R})} \mu_\infty$. The convergence of the first moment follows from $e_{1,1}(\mu_n, 0) \xrightarrow{n\to+\infty} e_{1,1}(\mu_\infty, 0)$. Hence, we conclude that $\mathcal{W}_1(\mu_n, \mu_\infty) \xrightarrow{n\to+\infty} 0$ by Theorem 2.1.1.

(*c*) The claim (*c*) is a direct result from (*a*) and (*b*). □

Proposition 2.4.2 (Even integer $p \geq 2$)**.** *Let* p *be an even integer,* $p \geq 2$.

(*a*) *Let* $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$ *such that* $e^p_{2,p}(\mu, \cdot) = e^p_{2,p}(\nu, \cdot) + C$ *for some real constant* C. *Then* $\mu = \nu$.

(*b*) *If* $\mu_n \in \mathcal{P}_p(\mathbb{R})$, $n \in \mathbb{N}^* \cup \{\infty\}$, *the following properties are equivalent:*

(*i*) $\mathcal{W}_p(\mu_n, \mu_\infty) \xrightarrow{n\to+\infty} 0$,

(*ii*) $e_{2,p}(\mu_n, \cdot) \xrightarrow{n\to+\infty} e_{2,p}(\mu_\infty, \cdot)$ *uniformly,*

(*iii*) $e_{2,p}(\mu_n, \cdot) \xrightarrow{n\to+\infty} e_{2,p}(\mu_\infty, \cdot)$ *pointwise.*

(*c*) *The distances* $\mathcal{Q}_{2,p}$ *and* $\mathcal{W}_p$ *are topologically equivalent on* $\mathcal{P}_p(\mathbb{R})$ *and* $N_{1,p} = 2$.

The proof of Proposition 2.4.2 is based on the following lemma.

Lemma 2.4.1. *Let* p *be an even number,* $p \geq 2$. *Let* $\mu \in \mathcal{P}_p(\mathbb{R})$ *be absolutely continuous*

with density f i.e. $\mu(d\xi) = f(\xi)d\xi$. *If f is continuous, then for any $a, b \in \mathbb{R}$ with $a < b$,*

$$e^{p-2}_{2,p-2}\big(\mu,(a,b)\big) = \frac{1}{p(p-1)}\left(\frac{\partial^2 e^p_{2,p}}{\partial a^2}\big(\mu,(a,b)\big) + \frac{\partial^2 e^p_{2,p}}{\partial b^2}\big(\mu,(a,b)\big) - 2\frac{\partial^2 e^p_{2,p}}{\partial a \partial b}\big(\mu,(a,b)\big)\right). \tag{2.4.1}$$

Proof of Lemma 2.4.1. Assume that $a < b$, then $e^p_{2,p}\big(\mu,(a,b)\big) = \int_{-\infty}^{\frac{a+b}{2}} |\xi - a|^p f(\xi)d\xi + \int_{\frac{a+b}{2}}^{+\infty} |\xi - b|^p f(\xi)d\xi$. Hence, the function $e^p_{2,p}\big(\mu,(a,b)\big)$ is continuously differentiable in a, since, for any even number $p \geq 2$, we have $\frac{\partial |\xi - a|^p f(\xi)}{\partial a} = p(a-\xi)^{p-1} f(\xi)$ and

$$\sup_{a' \in (a-1, a+1)} \big|p(a'-\xi)^{p-1} f(\xi)\big| \leq p 2^{p-1} f(\xi)\big[\,|a+1|^{p-1} \vee |a-1|^{p-1} + |\xi|^{p-1}\,\big] \in L^1(\lambda)$$

since $\int_{\mathbb{R}} |\xi|^p f(\xi)d\xi < +\infty$. Likewise, $e^p_{2,p}\big(\mu,(a,b)\big)$ is continuously differentiable in b with partial derivatives

$$\frac{\partial e^p_{2,p}\big(\mu,(a,b)\big)}{\partial a} = p\int_{-\infty}^{\frac{a+b}{2}} (a-\xi)^{p-1} f(\xi)d\xi \text{ and } \frac{\partial e^p_{2,p}\big(\mu,(a,b)\big)}{\partial b} = p\int_{\frac{a+b}{2}}^{+\infty} (b-\xi)^{p-1} f(\xi)d\xi.$$

Moreover, we have $\frac{\partial (a-\xi)^{p-1} f(\xi)}{\partial a} = (p-1)(a-\xi)^{p-2} f(\xi)$ and

$$\sup_{a' \in (a-1, a+1)} \big|(p-1)(a'-\xi)^{p-2} f(\xi)\big| \leq (p-1) 2^{p-2} f(\xi)\big[\,|a+1|^{p-2} \vee |a-1|^{p-2} + |\xi|^{p-2}\,\big] \in L^1(d\xi)$$

since $\int_{\mathbb{R}} |\xi|^p f(\xi)d\xi < +\infty$. By a similar reasoning, one derives that $e^p_{2,p}\big(\mu,(a,b)\big)$ is continuously twice differentiable with second order partial derivatives

$$\begin{aligned}
\frac{\partial^2 e^p_{2,p}}{\partial a^2}\big(\mu,(a,b)\big) &= p\Big[\int_{-\infty}^{\frac{a+b}{2}} (p-1)(a-\xi)^{p-2} f(\xi)d\xi - \frac{1}{2^p}(b-a)^{p-1} f\big(\frac{a+b}{2}\big)\Big],\\
\frac{\partial^2 e^p_{2,p}}{\partial b^2}\big(\mu,(a,b)\big) &= p\Big[\int_{\frac{a+b}{2}}^{+\infty} (p-1)(b-\xi)^{p-2} f(\xi)d\xi - \frac{1}{2^p}(b-a)^{p-1} f\big(\frac{a+b}{2}\big)\Big],\\
\frac{\partial^2 e^p_{2,p}}{\partial a \partial b}\big(\mu,(a,b)\big) &= \frac{\partial^2 e^p_{2,p}}{\partial b \partial a}\big(\mu,(a,b)\big) = -p\frac{1}{2^p}(b-a)^{p-1} f\Big(\frac{a+b}{2}\Big).
\end{aligned}$$

Hence, for every $(a,b) \in \mathbb{R}^2$ such that $a < b$,

$$\frac{\partial^2 e^p_{2,p}}{\partial a^2}\big(\mu,(a,b)\big) + \frac{\partial^2 e^p_{2,p}}{\partial b^2}\big(\mu,(a,b)\big) - 2\frac{\partial^2 e^p_{2,p}}{\partial a \partial b}\big(\mu,(a,b)\big) = p(p-1)e^{p-2}_{2,p-2}\big(\mu,(a,b)\big).$$

Proof of Proposition 2.4.2. (a) *Step 1: μ and ν are absolutely continuous with continuous density functions.* Note that $e^p_{2,p}(\mu,\cdot) = e^p_{2,p}(\nu,\cdot) + C$ implies either $\mu = \nu$ by

Proposition 2.3.1 if $p=2$, or, if $p>2$ $e^{p-2}_{2,p-2}(\mu,\cdot)=e^{p-2}_{2,p-2}(\nu,\cdot)$ (after differentiation) by Lemma 2.4.1. We can conclude by induction.

Step 2 (General case). Let X,Y be two random variables with the respective distributions μ and ν, such that

$$\forall (a,b)\in\mathbb{R}^2,\quad e^p_{2,p}\big(X,(a,b)\big)=e^p_{2,p}\big(Y,(a,b)\big)+C. \tag{2.4.2}$$

Let Z be a random variable with probability distribution $\mathbb{P}_Z=\mathcal{N}(0,1)$, independent of X and Y. For every $\varepsilon>0$,

$$e^p_{2,p}\big(X+\varepsilon Z,(a,b)\big)=\iint \min_{x\in\{a,b\}}|\xi+\varepsilon z-x|^p\,\mu(d\xi)\mathbb{P}_Z(dz)=\int e^p_{2,p}\big(X,(a,b)-\varepsilon z\big)\mathbb{P}_Z(dz). \tag{2.4.3}$$

We derive from (2.4.2) and (2.4.3) that

$$\forall (a,b)\in\mathbb{R}^2,\quad e^p_{2,p}\big(X+\varepsilon Z,(a,b)\big)=e^p_{2,p}\big(Y+\varepsilon Z,(a,b)\big)+C. \tag{2.4.4}$$

Moreover, the random variables $X+\varepsilon Z$ and $Y+\varepsilon Z$ have distributions $\mathcal{N}(0,\varepsilon^2)*\mu$ and $\mathcal{N}(0,\varepsilon^2)*\nu$ respectively, both with continuous densities. It follows from *Step 1* that $\mathrm{Law}(X+\varepsilon Z)=\mathrm{Law}(Y+\varepsilon Z)$ for every $\varepsilon>0$ so that $\mathrm{Law}(X)$=$\mathrm{Law}(Y)$ by letting $\varepsilon\to 0$.

(b) It is obvious that $(i)\Rightarrow(ii)$ and $(ii)\Rightarrow(iii)$. Now we prove $(iii)\Rightarrow(i)$. It follows from Lemma 2.4.1 that $e_{2,p}(\mu_n,\cdot)\xrightarrow{n\to+\infty}e_{2,p}(\mu_\infty,\cdot)$ implies $e_{2,p-2}(\mu_n,\cdot)\xrightarrow{n\to+\infty}e_{2,p-2}(\mu_\infty,\cdot)$ and, by induction, yields $e_{2,2}(\mu_n,\cdot)\xrightarrow{n\to+\infty}e_{2,2}(\mu_\infty,\cdot)$, so that Theorem 2.3.1 and Theorem 2.1.1 imply that μ_n converges weakly to μ_∞. The convergence of the p-th moment follows from $e_{2,p}(\mu_n,\mathbf{0})\xrightarrow{n\to+\infty}e_{2,p}(\mu_\infty,\mathbf{0})$. Hence $\mathcal{W}_p(\mu_n,\mu_\infty)\xrightarrow{n\to+\infty}0$ by Theorem 2.1.1.

(c) The claim (a) and (b) directly imply that if p is an even integer, $p\geq 2$, the distances $\mathcal{Q}_{2,p}$ and $\mathcal{W}_p$ are topologically equivalent on $\mathcal{P}_p(\mathbb{R})$ and $N_{1,p}\leq 2$. Now we prove that $N_{1,p}=2$. Note that for every $x\in\mathbb{R}$,

$$e^p_{1,p}(\mu,x)=\int_{\mathbb{R}}|\xi-x|^p\,\mu(d\xi)=\int_{\mathbb{R}}(\xi^2-2\xi x+x^2)^{\frac{p}{2}}\mu(d\xi),$$

which is polynomial in x and whose coefficients are the k-th moments of μ, $k\in\{1,...,p\}$. Thus, as soon as two different distributions μ and ν have the same first p moments, $e^p_{1,p}(\mu,\cdot)=e^p_{1,p}(\nu,\cdot)$. This implies $N_{1,p}>1$. □

2.4.2 About completeness of $\big(\mathcal{P}_1(\mathbb{R}), \mathcal{Q}_{1,1}\big)$ and $\big(\mathcal{P}_2(\mathbb{R}), \mathcal{Q}_{N,2}\big)$

We know from Bolley (2008) that for $p \geq 1$, $(\mathcal{P}_p(\mathbb{R}), \mathcal{W}_p)$ is a complete space and we have proved that $\mathcal{Q}_{1,1}$ (respectively $Q_{2,2}$) is topologically equivalent to $\mathcal{W}_1$ (resp. $\mathcal{W}_2$) on $\mathcal{P}_1(\mathbb{R})$ (resp. $\mathcal{P}_2(\mathbb{R})$). Now we discuss whether $\mathcal{Q}_{1,1}$ and $\mathcal{Q}_{2,2}$ are complete distances.

Proposition 2.4.3. *The metric space* $\big(\mathcal{P}_1(\mathbb{R}), \mathcal{Q}_{1,1}\big)$ *is complete.*

Proof. The inequality (2.1.8) directly implies that a Cauchy sequence in $\big(\mathcal{P}_1(\mathbb{R}), \mathcal{W}_1\big)$ is also a Cauchy sequence in $\big(\mathcal{P}_1(\mathbb{R}), \mathcal{Q}_{1,1}\big)$. Now let $(\mu_n)_{n\geq 1}$ be a Cauchy sequence in $\big(\mathcal{P}_1(\mathbb{R}), \mathcal{Q}_{1,1}\big)$. It follows from the definition of $\mathcal{Q}_{1,1}$ that $\big(e_{1,1}(\mu_n, \cdot) - e_{1,1}(\delta_0, \cdot)\big)_{n\geq 1}$ is a Cauchy sequence in $\big(\mathcal{C}_b(\mathbb{R}, \mathbb{R}), \|\cdot\|_{\sup}\big)$.

As $\big(\mathcal{C}_b(\mathbb{R}, \mathbb{R}), \|\cdot\|_{\sup}\big)$ is complete, there exists a function $g \in \mathcal{C}_b(\mathbb{R}, \mathbb{R})$ such that

$$\big\|\big(e_{1,1}(\mu_n, \cdot) - e_{1,1}(\delta_0, \cdot)\big) - g\big\|_{\sup} \xrightarrow{n\to+\infty} 0. \tag{2.4.5}$$

Note that for any $a \in \mathbb{R}$, $e_{1,1}(\delta_0, a) = |a|$. The sequence $e_{1,1}(\mu_n, 0) - e_{1,1}(\delta_0, 0) = e_{1,1}(\mu_n, 0)$ is also a Cauchy sequence in $\mathbb{R}$. Therefore, $\big(e_{1,1}(\mu_n, 0)\big)_{n\geq 1} = \big(\int_{\mathbb{R}} |\xi|\, \mu_n(d\xi)\big)_{n\geq 1}$ is bounded, which implies that $(\mu_n)_{n\geq 1}$ is tight. It follows from Prohorov's theorem that there exists a subsequence $(\mu_{\varphi(n)})_{n\geq 1}$ weakly converging to $\widetilde{\mu}_\infty$. Moreover, by Fatou's lemma in distribution, $\widetilde{\mu}_\infty \in \mathcal{P}_1(\mathbb{R})$ since $\int_{\mathbb{R}} |\xi|\, \widetilde{\mu}_\infty(d\xi) \leq \liminf_n \int_{\mathbb{R}} |\xi|\, \mu_{\varphi(n)}(d\xi) < +\infty$.

Now, we prove that $g = e_{1,1}(\widetilde{\mu}, \cdot) - e_{1,1}(\delta_0, \cdot)$. First, let us define a function $f_a(\xi) := |\xi - a| - |\xi|$. For every $a \in \mathbb{R}$, f_a is bounded and continuous. Hence, the weak convergence of $(\mu_{\varphi(n)})_{n\geq 1}$ implies that $\displaystyle\int_{\mathbb{R}} f_a(\xi)\mu_{\varphi(n)}(d\xi) \xrightarrow{n\to+\infty} \int_{\mathbb{R}} f_a(\xi)\widetilde{\mu}_\infty(d\xi)$.

Besides,

$$\int_{\mathbb{R}} f_a(\xi)\mu_{\varphi(n)}(d\xi) = \int_{\mathbb{R}} \big[\, |\xi - a| - |\xi| \,\big]\mu_{\varphi(n)}(d\xi) = e_{1,1}(\mu_{\varphi(n)}, a) - e_{1,1}(\mu_{\varphi(n)}, 0),$$

which converges to $\big(g(a) + e_{1,1}(\delta_0, a)\big) - \big(g(0) + e_{1,1}(\delta_0, 0)\big)$ as $n \to +\infty$ by (2.4.5). Hence, for every $a \in \mathbb{R}$,

$$\big(g(a) + e_{1,1}(\delta_0, a)\big) - \big(g(0) + \underbrace{e_{1,1}(\delta_0, 0)}_{=0}\big) = \int_{\mathbb{R}} f_a(\xi)\widetilde{\mu}_\infty(d\xi) = e_{1,1}(\widetilde{\mu}_\infty, a) - e_{1,1}(\widetilde{\mu}_\infty, 0),$$

i.e. $e_{1,1}(\widetilde{\mu}_\infty, a) - e_{1,1}(\delta_0, a) - g(a) = e_{1,1}(\widetilde{\mu}_\infty, 0) - g(0)$. Setting $C = e_{1,1}(\widetilde{\mu}_\infty, 0) - g(0)$, we derive that for every $a \in \mathbb{R}$,

$$e_{1,1}(\widetilde{\mu}_\infty, a) - e_{1,1}(\delta_0, a) - g(a) = C. \tag{2.4.6}$$

Now we prove that $C = 0$. Generally, for any $\nu \in \mathcal{P}_1(\mathbb{R})$, one has

$$\lim_{a\to+\infty}\big(e_{1,1}(\nu,a) - e_{1,1}(\delta_0,a)\big) = \lim_{a\to+\infty}\big(e_{1,1}(\nu,a) - |a|\,\big) = \lim_{a\to+\infty}\big(e_{1,1}(\nu,a) - a\big)$$
$$= \lim_{a\to+\infty}\Big(\int_{\mathbb{R}}|\xi-a|\,\nu(d\xi) - a\Big) = \lim_{a\to+\infty}\Big(\int_{\{\xi\ge a\}}(\xi-a)\nu(d\xi) + \int_{\{\xi<a\}}(a-\xi)\nu(d\xi) - a\Big)$$
$$= \lim_{a\to+\infty}\Big(\int_{\{\xi\ge a\}}\xi\nu(d\xi) - 2\int_{\{\xi\ge a\}}a\nu(d\xi) + \int_{\{\xi<a\}}(-\xi)\nu(d\xi)\Big).$$

As $\nu \in \mathcal{P}_1(\mathbb{R})$ i.e. $\int_{\mathbb{R}}|\xi|\,\nu(d\xi) < +\infty$, we derive that $\lim_{a\to+\infty}\int_{\xi<a}(-\xi)\nu(d\xi) = \int_{\mathbb{R}}(-\xi)\nu(d\xi)$ and $\lim_{a\to+\infty}\int_{\{\xi\ge a\}}\xi\nu(d\xi) = 0$. This implies

$$0 \le \lim_{a\to+\infty}\int_{\{\xi\ge a\}} a\,\nu(d\xi) \le \lim_{a\to+\infty}\int_{\{\xi\ge a\}}\xi\,\nu(d\xi) = 0.$$

After a similar calculation with $\lim_{a\to-\infty}\big(e_{1,1}(\nu,a) - e_{1,1}(\delta_0,a)\big)$, we get

$$\lim_{a\to+\infty}\big[e_{1,1}(\nu,a) - e_{1,1}(\delta_0,a)\big] = \int_{\mathbb{R}}(-\xi)\nu(d\xi)$$
$$\text{and } \lim_{a\to-\infty}\big[e_{1,1}(\nu,a) - e_{1,1}(\delta_0,a)\big] = \int_{\mathbb{R}}\xi\nu(d\xi). \tag{2.4.7}$$

Combining (2.4.6) and (2.4.7) with $\nu = \widetilde{\mu}_\infty$ shows that

$$\lim_{a\to+\infty} g(a) = -C - \int_{\mathbb{R}}\xi\widetilde{\mu}_\infty(d\xi) \text{ and } \lim_{a\to-\infty} g(a) = -C + \int_{\mathbb{R}}\xi\widetilde{\mu}_\infty(d\xi).$$

On the other hand, for every $n \ge 1$, (2.4.7) applied to $\nu = \mu_{\varphi(n)}$ implies

$$\lim_{a\to\pm\infty} e_{1,1}(\mu_{\varphi(n)},a) - e_{1,1}(\delta_0,a) = \mp\int_{\mathbb{R}}\xi\mu_{\varphi(n)}(d\xi).$$

Up to a new extraction of $\mu_{\varphi(n)}$, still denoted by $\mu_{\varphi(n)}$, we may assume that

$$\int_{\mathbb{R}}\xi\mu_{\varphi(n)}(d\xi) \to \widetilde{C} \in \mathbb{R}$$

as $n \to +\infty$ since $\big(e_{1,1}(\mu_n,0)\big)_{n\ge1} = \big(\int_{\mathbb{R}}|\xi|\,\mu_n(d\xi)\big)_{n\ge1}$ is bounded.

Now the uniform convergence (2.4.5) implies that

$$\lim_n \lim_{a\to\pm\infty}\Big[e_{1,1}(\mu_{\varphi(n)},a) - e_{1,1}(\delta_0,a) - g(a)\Big] = 0$$

so that $\widetilde{C} = C + \int_{\mathbb{R}}\xi\widetilde{\mu}_\infty(d\xi) = -C + \int_{\mathbb{R}}\xi\widetilde{\mu}_\infty(d\xi)$, which in turn implies $C = 0$, i.e.

$g = e_{1,1}(\widetilde{\mu}_\infty, \cdot) - e_{1,1}(\delta_0, \cdot)$. Then it follows from (2.4.5) that

$$\begin{aligned}&\big\|\big(e_{1,1}(\mu_n, \cdot) - e_{1,1}(\delta_0, \cdot)\big) - \big(e_{1,1}(\widetilde{\mu}_\infty, \cdot) - e_{1,1}(\delta_0, \cdot)\big)\big\|_{\sup}\\ &\quad = \|e_{1,1}(\mu_n, \cdot) - e_{1,1}(\widetilde{\mu}_\infty, \cdot)\|_{\sup} \xrightarrow{n\to+\infty} 0\end{aligned}$$

Hence, $\mathcal{W}_1(\mu_n, \widetilde{\mu}_\infty) \to 0$ by applying Proposition 2.4.1. The completeness of $(\mathcal{P}_1(\mathbb{R}), \mathcal{W}_1)$ implies that $\widetilde{\mu}_\infty$ is the unique limit of $(\mu_n)_{n\geq 1}$, which in turn implies that $(\mathcal{P}_1(\mathbb{R}), \mathcal{Q}_{1,1})$ is complete. □

Theorem 2.4.1. *For any $N \geq 2$, the metric space $\big(\mathcal{P}_2(\mathbb{R}), \mathcal{Q}_{N,2}\big)$ is not complete.*

We will build a sequence on $\mathcal{P}_2(\mathbb{R})$ which is Cauchy for $\mathcal{Q}_{N,2}$ but not for $\mathcal{W}_2$. First, we have the following result.

Lemma 2.4.2. *Let $(\mu_n)_{n\geq 1}$ be a $\mathcal{P}_2(\mathbb{R}^d)$-valued sequence which converges weakly to μ_∞ and, for $n \in \mathbb{N}^* \cup \{\infty\}$, let X_n denote a μ_n-distributed random variable . Assume that $\lim_n \mathbb{E}\,|X_n|^2$ exists and is finite. Then*

$$\sup_{a\in\mathbb{R}^d} \Big|e_{2,2}\big(\mu_n, (a,a)\big) - \sqrt{e^2_{2,2}\big(\mu_\infty, (a,a)\big) + C_0}\Big| \xrightarrow{n\to+\infty} 0, \tag{2.4.8}$$

where $C_0 = \lim\limits_n \mathbb{E}\,|X_n|^2 - \mathbb{E}\,|X_\infty|^2 \in [0, +\infty)$.

Proof of Lemma 2.4.2. An elementary computation shows that

$$e^2_{2,2}\big(\mu_n, (a,a)\big) = \int_{\mathbb{R}^d} |\xi - a|^2\, \mu_n(d\xi) = \int_{\mathbb{R}^d} |\xi|^2\, \mu_n(d\xi) - 2\Big(\int_{\mathbb{R}^d} \xi\mu_n(d\xi)\,\Big|\, a\Big) + |a|^2\,.$$

As $\Big(\int_{\mathbb{R}^d} |\xi|^2\, \mu_n(d\xi)\Big)_{n\geq 1}$ is bounded and $\mu_n \xRightarrow{(\mathbb{R}^d)} \mu_\infty$, we have $\int_{\mathbb{R}^d} \xi\mu_n(d\xi) \to \int_{\mathbb{R}^d} \xi\mu_\infty(d\xi)$. It follows that

$$\begin{aligned}e^2_{2,2}\big(\mu_n, (a,a)\big) &= \int_{\mathbb{R}^d} |\xi|^2\, \mu_n(d\xi) - 2\Big(\int_{\mathbb{R}^d} \xi\mu_n(d\xi)\,\Big|\, a\Big) + |a|^2\\ &\xrightarrow{n\to+\infty} \int_{\mathbb{R}^d} |\xi|^2\, \mu_\infty(d\xi) + C_0 - 2\Big(\int_{\mathbb{R}^d} \xi\mu_\infty(d\xi)\,\Big|\, a\Big) + |a|^2 = e^2_{2,2}\big(\mu_\infty, (a,a)\big) + C_0.\end{aligned}$$

Therefore, for every compact set K in $\mathbb{R}^d$, we have

$$\sup_{a\in K} \Big|e_{2,2}\big(\mu_n, (a,a)\big) - \sqrt{e^2_{2,2}\big(\mu_\infty, (a,a)\big) + C_0}\Big| \xrightarrow{n\to+\infty} 0, \tag{2.4.9}$$

owing to Arzelá-Ascoli theorem, since all functions $e_{N,p}$ are 1-Lipschitz continuous (see (2.1.5)). On the other hand, we have

$$\Big|e_{2,2}\big(\mu_n, (a,a)\big) - \sqrt{e^2_{2,2}(\mu_\infty, (a,a)) + C_0}\Big|$$

$$
\begin{aligned}
&= \frac{\left| e_{2,2}^2\big(\mu_n,(a,a)\big) - \Big(e_{2,2}^2\big(\mu_\infty,(a,a)\big) + C_0\Big)\right|}{e_{2,2}\big(\mu_n,(a,a)\big) + \sqrt{e_{2,2}^2\big(\mu_\infty,(a,a)\big) + C_0}} \\
&\leq \frac{\left|\mathbb{E}\big(\,|X_n|^2 - 2(a\,|\,X_n) + |a|^2\,\big) - \mathbb{E}\big(\,|X_\infty|^2 - 2(a\,|\,X_\infty) + |a|^2\,\big) - C_0\right|}{\|X_n - a\|_2 + \|X_\infty - a\|_2} \\
&\leq \frac{2\,|(a\mid \mathbb{E}X_\infty - \mathbb{E}X_n)| + \left|\mathbb{E}\,|X_n|^2 - \mathbb{E}\,|X_\infty|^2 - C_0\right|}{\|X_n - a\|_2 + \|X_\infty - a\|_2} \\
&\leq \frac{2\,|a|\,|\mathbb{E}X_\infty - \mathbb{E}X_n| + \left|\mathbb{E}\,|X_n|^2 - \mathbb{E}\,|X_\infty|^2 - C_0\right|}{\big|\,\|X_n\|_2 - |a|\,\big| + \big|\,\|X_\infty\|_2 - |a|\,\big|}.
\end{aligned}
\tag{2.4.10}
$$

Let $A := 2\sup_{n\in\mathbb{N}\cup\{\infty\}} \mathbb{E}\,|X_n|^2$, then

$$
\begin{aligned}
&\sup_{|a|>A} \left| e_{2,2}\big(\mu_n,(a,a)\big) - \sqrt{e_{2,2}^2(\mu_\infty,(a,a)) + C_0}\right| \\
&\quad\leq \sup_{|a|>A} \frac{2\,|a|\,|\mathbb{E}X_\infty - \mathbb{E}X_n| + \left|\mathbb{E}\,|X_n|^2 - \mathbb{E}\,|X_\infty|^2 - C_0\right|}{|a| - \|X_n\|_2 + |a| - \|X_\infty\|_2} \\
&\quad\leq \sup_{|a|>A} \frac{2\,|a|\,|\mathbb{E}X_\infty - \mathbb{E}X_n| + \left|\mathbb{E}\,|X_n|^2 - \mathbb{E}\,|X_\infty|^2 - C_0\right|}{2\,|a| - A} \\
&\quad\leq \sup_{|a|>A} 2\,|\mathbb{E}X_\infty - \mathbb{E}X_n| + \frac{\left|\mathbb{E}\,|X_n|^2 - \mathbb{E}\,|X_\infty|^2 - C_0\right|}{A} \xrightarrow{n\to+\infty} 0
\end{aligned}
\tag{2.4.11}
$$

Hence, (2.4.9) and (2.4.11) imply that

$$
\sup_{a\in\mathbb{R}^d} \left| e_{2,2}\big(\mu_n,(a,a)\big) - \sqrt{e_{2,2}^2\big(\mu_\infty,(a,a)\big) + C_0}\right| \xrightarrow{n\to+\infty} 0. \qquad \square
$$

Let $Z:\Omega\to\mathbb{R}$ be $\mathcal{N}(0,1)$-distributed. We define for every $n\in\mathbb{N}$,

$$
X_n := e^{\frac{n}{2}Z - \frac{n^2}{4}}. \tag{2.4.12}
$$

For $n\geq 1$, let μ_n denote the probability distribution of X_n. It is obvious that X_n converges a.s. to $X_\infty = 0$, so that $\mu_\infty = \delta_0$. Moreover, for every $p>0$, $\mathbb{E}\,X_n^p = e^{\frac{pn^2}{8}(p-2)}$. Hence, $\mathbb{E}\,X_n = e^{-\frac{n^2}{8}} \longrightarrow 0 = \mathbb{E}\,X_\infty$ as $n\to+\infty$ so that $\mathcal{W}_1(\mu_n,\mu_\infty)\to 0$ whereas $\mathbb{E}\,X_n^2 = 1$ for every $n\in\mathbb{N}$.

Hence $\mathbb{E}X_n^2$ does not converge to $\mathbb{E}\,X_\infty^2 = 0$, which entails that μ_n does not converge to μ_∞ for the Wasserstein distance $\mathcal{W}_2$ and thus μ_n is not a $\mathcal{W}_2$-Cauchy sequence. We first prove $(\mu_n)_{n\geq 1}$ is a Cauchy sequence in $\big(\mathcal{P}_2(\mathbb{R}),\mathcal{Q}_{2,2}\big)$. The proof relies on the following three lemmas.

Lemma 2.4.3. *Let $Z:\Omega\to\mathbb{R}$ be $\mathcal{N}(0,1)$-distributed. Then, $\forall z>0$, $\mathbb{P}(Z\geq z)\leq \frac{e^{-\frac{z^2}{2}}}{z\sqrt{2\pi}}$.*

Proof. $\mathbb{P}(Z\geq z)=\int_z^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx\leq\int_z^{+\infty}\frac{x}{z}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=\frac{e^{-\frac{z^2}{2}}}{z\sqrt{2\pi}}$. □

Lemma 2.4.4. *Define (X_n) as in (2.4.12), then $\sup_{K\geq 0}K\,\mathbb{E}(X_n-K)_+\to 0$ as $n\to+\infty$.*

Proof. We have

$$\begin{aligned}
K\,\mathbb{E}(X_n-K)_+ &= K\int_0^{\infty}\mathbb{P}\Big((X_n-K)_+\geq u\Big)du = K\int_0^{+\infty}\mathbb{P}(X_n>u+K)du\\
&= K\int_K^{+\infty}\mathbb{P}(X_n\geq v)dv = K\int_K^{+\infty}\mathbb{P}\Big(e^{\frac{n}{2}Z-\frac{n^2}{4}}\geq v\Big)dv\\
&= K\int_K^{+\infty}\mathbb{P}\Big(Z\geq\frac{n}{2}+\frac{2}{n}\ln v\Big)dv = K\int_{\ln K}^{\infty}\mathbb{P}\Big(Z\geq\frac{n}{2}+\frac{2}{n}u\Big)e^u du\ \ (\text{setting } u=\ln v).
\end{aligned}$$

By Lemma 2.4.3,

$$\mathbb{P}\Big(Z\geq\frac{n}{2}+\frac{2}{n}u\Big)\leq\frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{1}{2}(\frac{n}{2}+\frac{2}{n}u)^2}}{\frac{n}{2}+\frac{2}{n}u}=\frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{n^2}{8}-\frac{2}{n^2}u^2-u}}{\frac{n}{2}+\frac{2}{n}u}.$$

It follows that,

$$\begin{aligned}
K\,\mathbb{E}(X_n-K)_+ &\leq K\int_{\ln K}^{\infty}\frac{e^{-\frac{n^2}{8}-\frac{2}{n^2}u^2-u}}{\frac{n}{2}+\frac{2}{n}u}e^u\frac{du}{\sqrt{2\pi}}\leq\frac{Ke^{-\frac{n^2}{8}}}{\frac{n}{2}+\frac{2}{n}\ln K}\int_{\ln K}^{\infty}e^{-\frac{2}{n^2}u^2}\frac{du}{\sqrt{2\pi}}\\
&= \frac{Ke^{-\frac{n^2}{8}}}{\frac{n}{2}+\frac{2}{n}\ln K}\int_{\frac{2}{n}\ln K}^{\infty}e^{-\frac{w^2}{2}}\frac{n}{2}\frac{dw}{\sqrt{2\pi}}\ \ (\text{by setting } w=\frac{2}{n}u)\\
&= \frac{Ke^{-\frac{n^2}{8}}}{\frac{n}{2}+\frac{2}{n}\ln K}\frac{n}{2}\,\mathbb{P}\Big(Z\geq\frac{2}{n}\ln K\Big)\leq\frac{nKe^{-\frac{n^2}{8}}}{2(\frac{n}{2}+\frac{2}{n}\ln K)}\frac{e^{-\frac{1}{2}\frac{4}{n^2}(\ln K)^2}}{\sqrt{2\pi}\frac{2}{n}\ln K}\ \ (\text{by Lemma 2.4.3})\\
&= \frac{n}{2\sqrt{2\pi}}e^{-\frac{n^2}{8}}\frac{Ke^{-\frac{2}{n^2}(\ln K)^2}}{(1+\frac{4}{n^2}\ln K)\ln K}=\frac{n}{2\sqrt{2\pi}}e^{-\frac{n^2}{8}}\frac{e^{\ln K(1-\frac{2}{n^2}\ln K)}}{(1+\frac{4}{n^2}\ln K)\ln K}.
\end{aligned}\tag{2.4.13}$$

Since the function $u\mapsto u(1-\frac{2}{n^2}u)$ attains its maximum at $u=\frac{n^2}{4}$ with maximum value $\frac{n^2}{8}$, we will discuss the value of $K\,\mathbb{E}(X_n-K)_+$ in the following two cases:

$$(i)\ K\geq e^{\rho\frac{n^2}{4}},\qquad (ii)\ 0\leq K\leq e^{\rho\frac{n^2}{4}}\ ,$$

with the same fixed $\rho\in(0,\frac{1}{2})$ in both (i) and (ii).

Case (i): If $K \geq e^{\rho \frac{n^2}{4}}$, then $\ln K \geq \rho \frac{n^2}{4}$. It follows that

$$\begin{aligned} K\,\mathbb{E}(X_n - K)_+ &\leq \frac{n e^{-\frac{n^2}{8}}}{2\sqrt{2\pi}} \frac{e^{\ln K(1-\frac{2}{n^2}\ln K)}}{(1+\frac{4}{n^2}\ln K)\ln K} \leq \frac{n e^{-\frac{n^2}{8}}}{2\sqrt{2\pi}} \frac{e^{\frac{n^2}{8}}}{(1+\frac{4}{n^2}\times \rho\frac{n^2}{4})\rho\frac{n^2}{4}} \\ &= \frac{2}{n(1+\rho)\rho\sqrt{2\pi}} \to 0. \end{aligned}$$

Case (ii): If $0 \leq K \leq e^{\rho \frac{n^2}{4}}$, then

$$K\,\mathbb{E}(X_n - K)_+ \leq e^{\frac{\rho}{4}n^2}\mathbb{E}X_n = e^{\frac{\rho}{4}n^2} \cdot e^{-\frac{n^2}{8}} = e^{\frac{1}{4}(\rho-\frac{1}{2})n^2} \xrightarrow{n\to+\infty} 0.$$

Therefore, $\sup_{K>0}\ K\,\mathbb{E}(X_n - K)_+ \xrightarrow{n\to+\infty} 0.$ □

By Lemma 2.4.2,

$$\sup_{a\in\mathbb{R}^d} \left| e_{2,2}\big(\mu_n,(a,a)\big) - \sqrt{e^2_{2,2}\big(\mu_\infty,(a,a)\big)+C_0} \right| \xrightarrow{n\to+\infty} 0.$$

Consequently, it is reasonable to guess that

$$e_{N,2}(\mu_n,\cdot) \xrightarrow[n\to+\infty]{\|\cdot\|_{\sup}} \sqrt{e^2_{N,2}(\mu_\infty,\cdot)+1}$$

so that $(\mu_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $(\mathcal{P}_2(\mathbb{R}^d),\mathcal{Q}_{N,2})$. Let $g_N : \mathbb{R}^N \to \mathbb{R}_+$ be defined by

$$(a_1,\ldots,a_N) \mapsto g_N\big((a_1,\ldots,a_N)\big) \coloneqq \sqrt{e^2_{N,2}\big(\mu_\infty,(a_1,\ldots,a_N)\big)+1} = \sqrt{\min_{1\leq i\leq N}|a_i|^2+1}.$$

Proposition 2.4.4. *For every $N \geq 2$,*

$$\sup_{(a_1,\ldots,a_N)\in\mathbb{R}^N} \left| e_{N,2}\big(\mu_n,(a_1,\ldots,a_N)\big) - g_N\big((a_1,\ldots,a_N)\big)\right| \xrightarrow{n\to+\infty} 0.$$

Therefore, $(\mu_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $(\mathcal{P}_2(\mathbb{R}),\mathcal{Q}_{N,2})$ by the definition of $\mathcal{Q}_{N,2}$.

Proof. We proceed by induction.

$\triangleright$ $N=2$. Since the functions g_2 and $e_{2,2}(\mu_n,\cdot)$ are symmetric, it is only necessary to show that

$$\sup_{(a,b)\in\mathbb{R}^2,\,|a|\leq|b|} \left| e_{2,2}\big(\mu_n,(a,b)\big) - g_2(a,b)\right| \xrightarrow{n\to+\infty} 0.$$

Note that when $|a| \le |b|$, $g_2(a,b) = \sqrt{|a|^2+1} = g_2(a,a)$. We discuss now the value of $\left|e_{2,2}\big(\mu_n,(a,b)\big) - g_2(a,b)\right|$ in the following four cases,

(i) $0 \le a \le b$,

(ii) $a \le 0 \le b$, $\begin{cases} (ii,\alpha) & a \le 0 \le b \text{ with } |a| \le \frac{1}{2}|b| \\ (ii,\beta) & a \le 0 \le b \text{ with } \frac{1}{2}|b| \le |a| \le |b|, \end{cases}$

(iii) $b \le 0 \le a$, with $|a| \le |b|$,

(iv) $b \le a \le 0$.

Cases (iii) and (iv): $\mathbf{b<0}$ and $\frac{\mathbf{a+b}}{\mathbf{2}} < \mathbf{0}$. The random variables X_n are positive so that $|x-a| \le |x-b|$. Hence $e_{2,2}\big(\mu_n,(a,b)\big) = e_{2,2}\big(\mu_n,(a,a)\big)$. With a slight abuse of notation, we will write in what follows $(a,b) \in (iii)$ for

$$(a,b) \in \{(a,b) \in \mathbb{R}^2 \mid b \le 0 \le a, \text{and } |a| \le |b|\}.$$

We will adopt the same notation for other cases too. Then for the case (iii) and (iv), it is obvious by applying Lemma 2.4.2 that

$$\sup_{(a,b)\in(iii)\cup(iv)} \left|e_{2,2}\big(\mu_n,(a,b)\big) - g_2(a,b)\right| = \sup_{(a,b)\in(iii)\cup(iv)} \left|e_{2,2}\big(\mu_n,(a,a)\big) - g_2(a,a)\right| \xrightarrow{n\to+\infty} 0.$$

Case (i): $\mathbf{0 \le a \le b}$. We have

$$\begin{aligned}
&\sup_{(a,b)\in(i)} \left|e_{2,2}\big(\mu_n,(a,b)\big) - g_2(a,b)\right| \\
&\le \sup_{(a,b)\in(i)} \left|e_{2,2}\big(\mu_n,(a,b)\big) - e_{2,2}\big(\mu_n,(a,a)\big)\right| + \left|e_{2,2}\big(\mu_n,(a,a)\big) - g_2(a,a)\right| \\
&\le \sup_{(a,b)\in(i)} \left|\sqrt{\int_{\mathbb{R}} |\xi-a|^2 \wedge |\xi-b|^2 \,\mu_n(d\xi)} - \sqrt{\int_{\mathbb{R}} |\xi-a|^2 \,\mu_n(d\xi)}\right| \\
&\quad + \left|e_{2,2}\big(\mu_n,(a,a)\big) - g_2(a,a)\right| \\
&\le \sup_{(a,b)\in(i)} \sqrt{\int_{\mathbb{R}} \Big[\,|\xi-a|^2 - \big(\,|\xi-a|^2 \wedge |\xi-b|^2\,\big)\Big]\mu_n(d\xi)} + \left|e_{2,2}\big(\mu_n,(a,a)\big) - g_2(a,a)\right| \\
&\quad \text{(since } \left|\sqrt{\alpha} - \sqrt{\beta}\right| \le \sqrt{\beta-\alpha} \text{ for } \beta > \alpha > 0) \\
&\le \sup_{(a,b)\in(i)} \sqrt{\int_{\mathbb{R}} \Big(\,|\xi-a|^2 - |\xi-b|^2\Big)_+ \mu_n(d\xi)} + \left|e_{2,2}\big(\mu_n,(a,a)\big) - g_2(a,a)\right| \\
&\le \sup_{(a,b)\in(i)} \sqrt{\int_{\mathbb{R}} 2(b-a)\Big(\xi - \frac{b+a}{2}\Big)_+ \mu_n(d\xi)} + \left|e_{2,2}\big(\mu_n,(a,a)\big) - g_2(a,a)\right| \\
&\le \sup_{(a,b)\in(i)} 2\sqrt{\int_{\mathbb{R}} \frac{b}{2}\Big(\xi - \frac{b}{2}\Big)_+ \mu_n(d\xi)} + \left|e_{2,2}\big(\mu_n,(a,a)\big) - g_2(a,a)\right|
\end{aligned}$$

$$\leq\ 2\sqrt{\sup_{K\geq 0} K\,\mathbb{E}(X_n-K)_+}+\sup_{a\in\mathbb{R}}\big|e_{2,2}\big(\mu_n,(a,a)\big)-g_2(a,a)\big|\ \xrightarrow{n\to+\infty}\ 0.$$

Case (ii,α): $\mathbf{a\leq 0\leq b}$, with $\mathbf{|a|\leq\frac{1}{2}\,|b|}$. We have

$$\begin{aligned}
&\sup_{(a,b)\in(ii,\alpha)}\big|e_{2,2}\big(\mu_n,(a,b)\big)-g_2(a,b)\big|\\
&\qquad\leq\ \sup_{(a,b)\in(ii,\alpha)}\big|e_{2,2}\big(\mu_n,(a,b)\big)-e_{2,2}\big(\mu_n,(a,a)\big)\big|+\big|e_{2,2}\big(\mu_n,(a,a)\big)-g_2(a,a)\big|\\
&\qquad\leq\ \sup_{(a,b)\in(ii,\alpha)}\sqrt{\int_{\mathbb{R}}2(b-a)\Big(\xi-\frac{b+a}{2}\Big)_+\mu_n(d\xi)}+\big|e_{2,2}\big(\mu_n,(a,a)\big)-g_2(a,a)\big|\\
&\qquad\leq\ \sup_{(a,b)\in(ii,\alpha)}\sqrt{\int_{\mathbb{R}}3\cdot b\Big(\xi-\frac{b}{4}\Big)_+\mu_n(d\xi)}+\big|e_{2,2}\big(\mu_n,(a,a)\big)-g_2(a,a)\big|\\
&\qquad\leq\ 2\sqrt{3}\cdot\sqrt{\sup_{K\geq 0} K\,\mathbb{E}(X_n-K)_+}+\sup_{a\in\mathbb{R}}\big|e_{2,2}\big(\mu_n,(a,a)\big)-g_2(a,a)\big|\ \xrightarrow{n\to+\infty}\ 0.
\end{aligned}$$

Case (ii,β): $\mathbf{a\leq 0\leq b}$, with $\mathbf{\frac{1}{2}\,|b|\leq|a|\leq|b|}$. One has

$$\begin{aligned}
&\sup_{(a,b)\in(ii,\beta)}\big|e_{2,2}\big(\mu_n,(a,b)\big)-g_2(a,b)\big|\\
&\qquad\leq\ \sup_{(a,b)\in(ii,\beta)}\big|e_{2,2}\big(\mu_n,(a,b)\big)-e_{2,2}\big(\mu_n,(a,a)\big)\big|+\big|e_{2,2}\big(\mu_n,(a,a)\big)-g_2(a,a)\big|\\
&\qquad\leq\ \sup_{(a,b)\in(ii,\beta)}\frac{\big|e_{2,2}^2\big(\mu_n,(a,b)\big)-e_{2,2}^2\big(\mu_n,(a,a)\big)\big|}{e_{2,2}\big(\mu_n,(a,b)\big)+e_{2,2}\big(\mu_n,(a,a)\big)}+\big|e_{2,2}\big(\mu_n,(a,a)\big)-g_2(a,a)\big|\\
&\qquad\leq\ \sup_{(a,b)\in(ii,\beta)}\frac{\int_{\mathbb{R}}2(b-a)\big(\xi-\frac{b+a}{2}\big)_+\mu_n(d\xi)}{\|X_n-a\|_2}+\sup_{a\in\mathbb{R}}\big|e_{2,2}\big(\mu_n,(a,a)\big)-g_2(a,a)\big|\\
&\qquad\leq\ \sup_{(a,b)\in(ii,\beta)}\frac{2(b-a)\,\mathbb{E}\left(X_n-\frac{b+a}{2}\right)_+}{\|X_n-a\|_2}+\sup_{a\in\mathbb{R}}\big|e_{2,2}\big(\mu_n,(a,a)\big)-g_2(a,a)\big|.
\end{aligned}$$

As $\|X_n-a\|_2=\big(\underbrace{\mathbb{E}X_n^2}_{=1}\underbrace{-2a\,\mathbb{E}X_n}_{\geq 0}+|a|^2\big)^{1/2}\geq\sqrt{1+|a|^2}$, we have

$$\begin{aligned}
&\sup_{(a,b)\in(ii,\beta)}\big|e_{2,2}\big(\mu_n,(a,b)\big)-g_2(a,b)\big|\\
&\qquad\leq\ \sup_{(a,b)\in(ii,\beta)}\frac{2(b+|a|)\mathbb{E}\big[X_n-\frac{b+a}{2}\big]_+}{\sqrt{1+|a|^2}}+\sup_{a\in\mathbb{R}}\big|e_{2,2}\big(\mu_n,(a,a)\big)-g_2(a,a)\big|\\
&\qquad\leq\ \sup_{(a,b)\in(ii,\beta)}\frac{4b\,\mathbb{E}X_n}{\sqrt{1+\frac{b^2}{4}}}+\sup_{a\in\mathbb{R}}\big|e_{2,2}\big(\mu_n,(a,a)\big)-g_2(a,a)\big|.
\end{aligned}$$

$$\leq 8\,\mathbb{E}X_n + \sup_{a\in\mathbb{R}} \left|e_{2,2}\big(\mu_n,(a,a)\big) - g_2(a,a)\right| \xrightarrow{n\to+\infty} 0.$$

▷ **From N to $N+1$.** Assume now that

$$\sup_{(a_1,\ldots,a_N)\in\mathbb{R}^N} \left|e_{N,2}\big(\mu_n,(a_1,\ldots,a_N)\big) - g_N(a_1,\ldots,a_N)\right| \to 0 \text{ as } n\to+\infty.$$

Then, for the level $N+1$, we assume without loss of generality that $|a_1|\leq|a_2|\leq\cdots\leq|a_{N+1}|$ since g_{N+1} and $e_{N,2}(\mu_n,\cdot)$ are symmetric. Under this assumption,

$$g_{N+1}(a_1,\ldots,a_{N+1}) = g_2(a_1,a_1) = \sqrt{|a_1|^2+1}. \tag{2.4.14}$$

We discuss now the value of

$$\sup_{(a_1,\ldots,a_{N+1})\in\mathbb{R}^{N+1}} \left|e_{N+1,2}\big(\mu_n,(a_1,\ldots,a_{N+1})\big) - g_{N+1}(a_1,\ldots,a_{N+1})\right|$$

in the following cases:

(i) $\exists\, i\in\{2,\ldots,N+1\}$ such that $a_i<0$,

(ii) $0\leq a_1\leq a_2\leq\cdots\leq a_{N+1}$,

(iii) $a_1\leq 0\leq a_2\leq\cdots\leq a_{N+1}$,

$$\begin{cases} (iii,\alpha)\ a_1\leq 0\leq a_2\leq\cdots\leq a_{N+1}, \text{ with } |a_1|\leq\frac{1}{2}|a_{N+1}| \\ (iii,\beta)\ a_1\leq 0\leq a_2\leq\cdots\leq a_{N+1}, \text{ with } |a_1|\geq\frac{1}{2}|a_{N+1}| \end{cases}.$$

Case (i): $\exists\,\mathbf{i}\in\{2,\ldots,N+1\}$ such that $a_i<0$. For every $n\geq 1$, X_n is a.s. positive. Hence, $|X_n-a_1|\leq|X_n-a_i|$ a.s. since we assume that $|a_1|\leq|a_2|\leq\cdots\leq|a_{N+1}|$. Therefore,

$$e_{N+1,2}\big(\mu_n,(a_1,\ldots,a_{N+1})\big) = e_{N,2}\big(\mu_n,(a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_{N+1})\big).$$

It follows from (2.4.14) that

$$\begin{aligned}&\sup_{(a_1,\ldots,a_{N+1})\in\mathbb{R}^{N+1}} \left|e_{N+1,2}\big(\mu_n,(a_1,\ldots,a_{N+1})\big) - g_{N+1}(a_1,\ldots,a_{N+1})\right| \\ &= \sup_{(a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_{N+1})\in\mathbb{R}^N} \Big|e_{N,2}\big(\mu_n,(a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_{N+1})\big) \\ &\qquad\qquad - g_N(a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_{N+1})\Big|,\end{aligned}$$

which converges to 0 as $n\to+\infty$ owing to the assumption on the level N.

Case (ii)**:** $\mathbf{0 \le a_1 \le a_2 \le \cdots \le a_{N+1}}$.

$$\begin{aligned}
&\sup_{0\le a_1\le a_2\le\cdots\le a_{N+1}} \left|e_{N+1,2}\big(\mu_n,(a_1,\dots,a_{N+1})\big) - g_{N+1}(a_1,\dots,a_{N+1})\right| \\
&\qquad\le \sup_{0\le a_1\le a_2\le\cdots\le a_{N+1}} \left|e_{N+1,2}\big(\mu_n,(a_1,\dots,a_{N+1})\big) - e_{N,2}\big(\mu_n,(a_1,\dots,a_N)\big)\right| \\
&\qquad\quad + \sup_{0\le a_1\le a_2\le\cdots\le a_{N+1}} \left|e_{N,2}\big(\mu_n,(a_1,\dots,a_N)\big) - g_N(a_1,\dots,a_N)\right|. \quad (2.4.15)
\end{aligned}$$

The second term on the right hand side of (2.4.15) converges to 0 as $n \to +\infty$ owing to the assumption on the level N.

For the first term on the right hand side of (2.4.15), we have

$$\begin{aligned}
&\sup_{0\le a_1\le a_2\le\cdots\le a_{N+1}} \left|e_{N+1,2}\big(\mu_n,(a_1,\dots,a_{N+1})\big) - e_{N,2}\big(\mu_n,(a_1,\dots,a_N)\big)\right| \\
&= \sup_{0\le a_1\le a_2\le\cdots\le a_{N+1}} \sqrt{\int_{\mathbb{R}} \min_{1\le i\le N} |\xi - a_i|^2 \,\mu_n(d\xi)} \\
&\qquad - \sqrt{\int_{\mathbb{R}} \Big[\min_{1\le i\le N} |\xi - a|^2\Big] \wedge |\xi - a_{N+1}|^2 \,\mu_n(d\xi)} \\
&\le \sup_{0\le a_1\le a_2\le\cdots\le a_{N+1}} \sqrt{\int_{\mathbb{R}} \min_{1\le i\le N} |\xi - a_i|^2 - \Big[\min_{1\le i\le N} |\xi - a_i|^2\Big] \wedge |\xi - a_{N+1}|^2 \,\mu_n(d\xi)} \\
&= \sup_{0\le a_1\le a_2\le\cdots\le a_{N+1}} \sqrt{\int_{\mathbb{R}} \big(\min_{1\le i\le N} |\xi - a_i|^2 - |\xi - a_{N+1}|^2\big)_+ \mu_n(d\xi)} \\
&\le \sup_{0\le a_1\le a_2\le\cdots\le a_{N+1}} \sqrt{\int_{\mathbb{R}} \big(|\xi - a_1|^2 - |\xi - a_{N+1}|^2\big)_+ \mu_n(d\xi)} \\
&= \sup_{0\le a_1\le a_2\le\cdots\le a_{N+1}} \sqrt{\int_{\mathbb{R}} 2(a_{N+1} - a_1)\big(\xi - \frac{a_1 + a_{N+1}}{2}\big)_+ \mu_n(d\xi)} \\
&\le \sup_{0\le a_1\le a_2\le\cdots\le a_{N+1}} \sqrt{\int_{\mathbb{R}} 2\cdot a_{N+1}\big(\xi - \frac{a_{N+1}}{2}\big)_+ \mu_n(d\xi)} \\
&\le 2\cdot\sqrt{\sup_{K\ge 0} K\,\mathbb{E}(X_n - K)_+} \xrightarrow{n\to+\infty} 0.
\end{aligned}$$

Case (iii,α)**:** $\mathbf{a_1 \le 0 \le a_2 \le \cdots \le a_{N+1}}$ **with** $\mathbf{|a_1| \le \frac{1}{2}|a_{N+1}|}$.

$$\begin{aligned}
&\sup_{(a_1,\dots,a_{N+1})\in(iii,\alpha)} \left|e_{N+1,2}\big(\mu_n,(a_1,\dots,a_{N+1})\big) - g_{N+1}(a_1,\dots,a_{N+1})\right| \\
&\le \sup_{(a_1,\dots,a_{N+1})\in(iii,\alpha)} \left|e_{N+1,2}\big(\mu_n,(a_1,\dots,a_{N+1})\big) - e_{N,2}\big(\mu_n,(a_1,\dots,a_N)\big)\right| \\
&\qquad + \sup_{(a_1,\dots,a_{N+1})\in(iii,\alpha)} \left|e_{N,2}\big(\mu_n,(a_1,\dots,a_N)\big) - g_N(a_1,\dots,a_N)\right|. \quad (2.4.16)
\end{aligned}$$

Like in Case (ii), the second term on the right hand side of (2.4.16) converges to 0 as $n \to +\infty$. For the first term of the right hand side of (2.4.16), we have

$$\begin{aligned}
&\sup_{(a_1,\ldots,a_{N+1})\in(iii,\alpha)} \left|e_{N+1,2}\big(\mu_n,(a_1,\ldots,a_{N+1})\big) - e_{N,2}\big(\mu_n,(a_1,\ldots,a_N)\big)\right| \\
&\quad\le \sup_{(a_1,\ldots,a_{N+1})\in(iii,\alpha)} \sqrt{\int_{\mathbb{R}} 2(a_{N+1}-a_1)\big(\xi - \frac{a_1+a_{N+1}}{2}\big)_+ \mu_n(d\xi)} \\
&\quad\le \sup_{(a_1,\ldots,a_{N+1})\in(iii,\alpha)} \sqrt{\int_{\mathbb{R}} 3\cdot a_{N+1}\big(\xi - \frac{a_{N+1}}{4}\big)_+ \mu_n(d\xi)} \\
&\quad\le 2\sqrt{3}\cdot\sqrt{\sup_{K\ge 0} K\,\mathbb{E}(X_n-K)_+} \longrightarrow 0.
\end{aligned}$$

Case (iii,β)**:** $\mathbf{a_1 \le 0 \le a_2 \le \cdots \le a_{N+1}}$ **with** $\mathbf{|a_1| \ge \frac{1}{2}\,|a_{N+1}|}$.

Since we assume $|a_1| \le |a_2| \le \cdots \le |a_{N+1}|$, then for any $i \in \{2,\ldots,N+1\}$, we have $\frac{1}{2}\,|a_i| \le |a_1| \le |a_i|$. It follows that

$$\begin{aligned}
&\sup_{(a_1,\ldots,a_{N+1})\in(iii,\beta)} \left|e_{N+1,2}\big(\mu_n,(a_1,\ldots,a_{N+1})\big) - g_{N+1}(a_1,\ldots,a_{N+1})\right| \\
&\qquad\le \sup_{(a_1,\ldots,a_{N+1})\in(iii,\beta)} \left|e_{N+1,2}\big(\mu_n,(a_1,\ldots,a_{N+1})\big) - e_{2,2}\big(\mu_n,(a_1,a_1)\big)\right| \\
&\qquad\quad + \sup_{a_1\in\mathbb{R}} \left|e_{2,2}\big(\mu_n,(a_1,a_1)\big) - g_N(a_1,a_1)\right|. \qquad (2.4.17)
\end{aligned}$$

The second part of (2.4.17), $\sup_{a_1\in\mathbb{R}} \left|e_{2,2}\big(\mu_n,(a_1,a_1)\big) - g_N(a_1,a_1)\right|$ converges to 0 as $n \to +\infty$ owing to Lemma 2.4.2. Then for the first part of (2.4.17), we have

$$\begin{aligned}
&\sup_{(a_1,\ldots,a_{N+1})\in(iii,\beta)} \left|e_{N+1,2}\big(\mu_n,(a_1,\ldots,a_{N+1})\big) - e_{2,2}\big(\mu_n,(a_1,a_1)\big)\right| \\
&\quad= \sup_{(a_1,\ldots,a_{N+1})\in(iii,\beta)} \frac{e^2_{2,2}\big(\mu_n,(a_1,a_1)\big) - e^2_{N+1,2}\big(\mu_n,(a_1,\ldots,a_{N+1})\big)}{e_{N+1,2}\big(\mu_n,(a_1,\ldots,a_{N+1})\big) + e_{2,2}\big(\mu_n,(a_1,a_1)\big)} \\
&\quad\le \sup_{(a_1,\ldots,a_{N+1})\in(iii,\beta)} \frac{\int_{\mathbb{R}} |\xi-a_1|^2 - \min_{1\le i\le N+1} |\xi-a_i|^2\,\mu_n(d\xi)}{\|X_n-a_1\|_2} \\
&\quad\le \sup_{(a_1,\ldots,a_{N+1})\in(iii,\beta)} \frac{\int_{\mathbb{R}} \big(|\xi-a_1|^2 - \min_{2\le i\le N+1} |\xi-a_i|^2\big)_+ \mu_n(d\xi)}{\|X_n-a_1\|_2} \\
&\quad\le \sup_{(a_1,\ldots,a_{N+1})\in(iii,\beta)} \frac{1}{\|X_n-a_1\|_2}\Big[\sum_{i=2}^{N+1}\int_{\mathbb{R}} \big(|\xi-a_1|^2 - |\xi-a_i|^2\big)_+ \mu_n(d\xi)\Big]
\end{aligned}$$

Since $a_1 < 0$, $\|X_n - a_1\|_2 = \left(\mathbb{E}X_n^2 - 2a_1\mathbb{E}X_n + |a_1|^2\right)^{1/2} \geq \sqrt{1+|a_1|^2}$. Therefore,

$$\frac{\int_{\mathbb{R}} \left(|\xi - a_1|^2 - |\xi - a_i|^2\right)_+ \mu_n(d\xi)}{\|X_n - a_1\|_2} = \frac{\int_{\mathbb{R}} 2(a_i - a_1)\left(\xi - \frac{a_i+a_1}{2}\right)_+ \mu_n(d\xi)}{\|X_n - a_1\|_2}$$
$$\leq \frac{4a_i\mathbb{E}X_n}{\sqrt{1+|a_1|^2}} \leq \frac{4a_i\mathbb{E}X_n}{\frac{1}{2}a_i} = 8\,\mathbb{E}X_n.$$

for $i \in \{2, \ldots, N+1\}$. Consequently,

$$\sup_{(a_1,\ldots,a_{N+1})\in(iii,\beta)} \left|e_{N+1,2}\big(\mu_n, (a_1, \ldots, a_{N+1})\big) - e_{2,2}\big(\mu_n, (a_1, a_1)\big)\right| \leq 8N\,\mathbb{E}X_n = 8\,Ne^{-n^2/8} \longrightarrow 0.$$

This completes the proof. □

Proof of Theorem 2.4.1. Let μ_n be the probability distribution of X_n defined in (2.4.12). If for some $N \geq 2$, $\big(\mathcal{P}_2(\mathbb{R}), \mathcal{Q}_{N,2}\big)$ were complete, then there exists a probability measure $\widetilde{\mu}$ in $\mathcal{P}_2(\mathbb{R})$ such that $\mathcal{Q}_{N,2}(\mu_n, \widetilde{\mu}) \longrightarrow 0$. Then, $\mathcal{W}_2(\mu_n, \widetilde{\mu}) \longrightarrow 0$ by applying Proposition 2.4.2, which creates a contradiction. □

Remark 2.4.1. The extension of this result to a Hilbert or simply multidimensional setting, although likely, is not straightforward.

2.5 Appendix: some examples of $c(d, |\cdot|_r)$

Proof of Proposition 2.2.4. (i) is obvious.
(ii) $c(2, |\cdot|_1) = 2$ is obvious (see Figure 2.1). Now we prove that $c(2, |\cdot|_r) = 3$ for every $r \in (1, +\infty)$.

We choose $a_1 = (0, 1)$, $a_2 = \big((1-2^{-r})^{\frac{1}{r}}, -\frac{1}{2}\big)$ and $a_3 = \big(-(1-2^{-r})^{\frac{1}{r}}, -\frac{1}{2}\big)$. We will first show that $S_{|\cdot|_r}(0,1) \subset \bigcup_{1\leq i\leq 3} \bar{B}_{|\cdot|_r}(a_i, 1)$.

Let (x, y) be any point on $S_{|\cdot|_r}(0,1)$, then $|x|^r + |y|^r = 1$.

- If $\frac{1}{2} \leq y \leq 1$, then $(1-y)^r \leq y^r$ so that

$$\big|(x,y) - a_1\big|_r^r = |x|^r + (1-y)^r = 1 - y^r + (1-y)^r \leq 1,$$

that is, $(x, y) \in \bar{B}_{|\cdot|_r}(a_1, 1)$.

- If $-1 \leq y \leq \frac{1}{2}$ and $x \geq 0$, then

$$\big|(x,y) - a_2\big|_r^r = \big|x - (1-2^{-r})^{\frac{1}{r}}\big|^r + \big|y + \frac{1}{2}\big|^r = \big|(1-|y|^r)^{\frac{1}{r}} - (1-2^{-r})^{\frac{1}{r}}\big|^r + \big|y + \frac{1}{2}\big|^r$$

$$\leq \big||y|^r - 2^{-r}\big| + \big|y + \frac{1}{2}\big|^r,$$

the last inequality is due to the fact that the function $u \mapsto u^{-\frac{1}{r}}$ is $\frac{1}{r}$-Hölder. As $r \geq 1$, the function $y \mapsto \big||y|^r - 2^{-r}\big| + \big|y + \frac{1}{2}\big|^r$ is convex over $[-1, \frac{1}{2}]$. Consequently, it attains its maximum either at -1 or at $\frac{1}{2}$. Hence, $\big|(x,y) - a_2\big|_r^r$ is upper bounded by 1 since

$$\begin{aligned} &\text{if } \ y = -1, \ \big||y|^r - 2^{-r}\big| + \big|y + \tfrac{1}{2}\big|^r = 1 - 2^{-r} + 2^{-r} = 1, \\ &\text{if } \ y = \tfrac{1}{2}, \quad \Big||y|^r - 2^{-r}\Big| + \big|y + \frac{1}{2}\big|^r = \big|2^{-r} - 2^{-r}\big| + 1^r = 1. \end{aligned}$$

This implies that $(x,y) \in \bar{B}_{|\cdot|_r}(a_2, 1)$.

- If $-1 \leq y \leq \frac{1}{2}$ and $x \leq 0$, then $(x,y) \in \bar{B}_{|\cdot|_r}(a_3, 1)$ by the symmetry of the unit sphere.

Next, we will show $c(2, |\cdot|_r) > 2$ for every $1 < r < +\infty$. Let a_1 and a_2 denote the two centers of balls on the sphere $S_{|\cdot|}(0,1)$. Since the ℓ^r-ball is centrally symmetric with respect to $(0,0)$, we fix $a_1 = (x,y)$ such that $x \in [(\frac{1}{2})^{\frac{1}{r}}, 1]$, $y \in [0, (\frac{1}{2})^{\frac{1}{r}}]$ and $x^r + y^r = 1$.

We first prove that if $r > 1$, $x \in [(\frac{1}{2})^{\frac{1}{r}}, 1]$, $y \in (0, (\frac{1}{2})^{\frac{1}{r}}]$ s.t. $x^r + y^r = 1$, then $(x+y)^r > 1$. Let $q = r - 1$, then $q > 0$ and

$$\begin{aligned} (x+y)^r = (x+y)^{1+q} &= (x+y)(x+y)^q = x(x+y)^q + y(x+y)^q \\ &> xx^q + yy^q = x^r + y^r = 1. \end{aligned}$$

- *Case 1.* We choose a_2 such that a_2 is centrally symmetric to a_1 with respect to the center $(0,0)$, i.e. $a_2 = (-x, -y)$.

  We prove $z_1 = (y, -x) \notin \cup_{i=1,2} \bar{B}_{|\cdot|_r}(a_i, 1)$ and $z_2 = (-y, x) \notin \cup_{i=1,2} \bar{B}_{|\cdot|_r}(a_i, 1)$. In fact, if $y = 0$, then $\big|a_1 - z_1\big|_r = \big|a_2 - z_1\big|_r = 2 > 1$. If $y > 0$, then

$$\begin{aligned} \big|a_1 - z_1\big|_r^r = \big|a_2 - z_1\big|_r^r = \big|a_1 - z_2\big|_r^r = \big|a_2 - z_2\big|_r^r &= (x+y)^r + (x-y)^r \\ &\geq (x+y)^r > 1. \end{aligned}$$

- *Case 2.* The point a_2 is not centrally symmetric to a_1.

  Let $H_{a_1} := \{\eta = (\eta_1, \eta_2) \in \mathbb{R}^2 \text{ s.t. } x \cdot \eta_2 = y \cdot \eta_1\}$, which is the straight line (with respect to the Euclidean distance) across the origin and a_1. Then between z_1 and z_2, there exists at least one point which is not in the same side of H_{a_1} as a_2, and this point can not be covered by $\cup_{i=1,2} \bar{B}_{|\cdot|_r}(a_i, 1)$.

Figure 2.2 illustrates that $c(2, |\cdot|_r) = 3$ when $r = 3$.

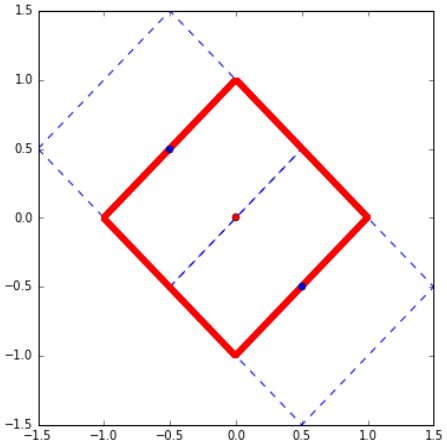


Figure 2.1 $a_1 = (-\frac{1}{2}, \frac{1}{2})$, $a_2 = (\frac{1}{2}, -\frac{1}{2})$, then $S_{|\cdot|_1}(0,1) \subset \bigcup_{i=1,2} \bar{B}_{|\cdot|_1}(a_i, 1)$

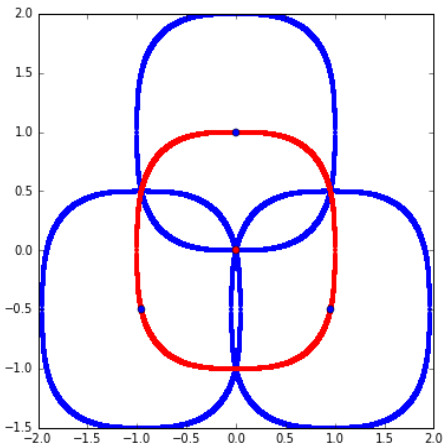


Figure 2.2 $c(2, |\cdot|_3) = 3$

(*iii*) Let $a_1 = (-1, 0, \ldots, 0)$ and $a_2 = (1, 0, \ldots, 0)$. We will show that $S_{|\cdot|_\infty}(0,1) \subset \bigcup_{i=1,2} \bar{B}_{|\cdot|_\infty}(a_i, 1)$.

Let $x = (x^1, \ldots, x^d) \in S_{|\cdot|_\infty}(0,1)$. There exists i_0 such that $\max_{1 \le i \le d} |x^i| \le |x^{i_0}| = 1$.

- If $i_0 = 1$, and $x^1 = -1$, then $\left|x - a_1\right|_\infty = \left|x^1 + 1\right| \vee \max_{i=\{2,\ldots,d\}} \left|x^i\right| \le 1$, that is, $x \in \bar{B}_{|\cdot|_\infty}(a_1, 1)$.
- If $i_0 = 1$, and $x^1 = 1$, then $\left|x - a_2\right|_\infty = \left|x^1 - 1\right| \vee \max_{i=\{2,\ldots,d\}} \left|x^i\right| \le 1$, that is, $x \in \bar{B}_{|\cdot|_\infty}(a_2, 1)$.
- If $i_0 \ge 2$, and $x^1 \le 0$, then $\left|x - a_1\right|_\infty = \left|x^1 + 1\right| \vee 1 \le 1$, that is, $x \in \bar{B}_{|\cdot|_\infty}(a_1, 1)$.

- If $i_0 \geq 2$, and $x^1 \geq 0$, then $\left|x - a_2\right|_\infty = \left|x^1 - 1\right| \vee 1 \leq 1$, that is, $x \in \bar{B}_{|\cdot|_\infty}(a_2, 1)$.

Consequently, we conclude that $S_{|\cdot|_\infty}(0,1) \subset \bigcup_{i=1,2} \bar{B}_{|\cdot|_\infty}(a_i, 1)$ and $c(d, |\cdot|_\infty) > 1$ is obvious.

(iv) Let $a_i = (0, \ldots, 1, \ldots, 0)$ - the i^{th} coordinate of a_i is equal to 1 and the others equal to 0. We will show that $S_{|\cdot|_r}(0,1) \subset \bigcup_{i=1}^d \left(\bar{B}_{|\cdot|_r}(a_i, 1) \cup \bar{B}_{|\cdot|_r}(-a_i, 1)\right)$.

For any $x = (x^1, \ldots, x^d) \in S_{|\cdot|_r}(0,1)$, then there exists $i_0 \in \{1, \ldots, d\}$ such that $\left|x^{i_0}\right| \geq \frac{1}{2}$. Otherwise $1 = \sum\limits_{1 \leq i \leq d} \left|x^i\right|^r < d \times 2^{-r} \leq 1$, which yields a contradiction.

• If $x^{i_0} \geq \frac{1}{2}$, then $\left|x - a_{i_0}\right|^r = (1 - x^{i_0})^r + \sum_{i \neq i_0} \left|x^i\right|^r = (1 - x^{i_0})^r + 1 - (x^{i_0})^r$. As $x^{i_0} \leq \frac{1}{2}$, we have $(1 - x^{i_0})^r - (x^{i_0})^r \leq 0$, so that $\left|x - a_{i_0}\right|^r \leq 1$, which implies that $x \in \bar{B}_{|\cdot|_r}(a_{i_0}, 1)$.

• If $x^{i_0} \leq -\frac{1}{2}$, one can similarly prove that $x \in \bar{B}_{|\cdot|_r}(-a_{i_0}, 1)$.

Consequently, we can conclude that $S_{|\cdot|_r}(0,1) \subset \bigcup_{i=1}^d \left(\bar{B}_{|\cdot|_r}(a_i, 1) \cup \bar{B}_{|\cdot|_r}(-a_i, 1)\right)$. $\square$

Chapter 3

Convergence Rate of the Optimal Quantizers and Application to the Clustering Performance of the Empirical Measure

This chapter corresponds to the arXiv preprint Liu and Pagès (2018), which is a joint work with Gilles Pagès.

Abstract: We study the convergence rate of optimal quantization for a probability measure sequence $(\mu_n)_{n\in\mathbb{N}^*}$ on $\mathbb{R}^d$ which converges in the Wasserstein distance in two aspects: the first one is the convergence rate of optimal quantizer $x^{(n)} \in (\mathbb{R}^d)^K$ of μ_n at level K; the other one is the convergence rate of the distortion function valued at $x^{(n)}$, called the "performance" of $x^{(n)}$. Moreover, we will study the mean performance of the optimal quantizer of the empirical measure of a distribution μ with finite second moment but possibly unbounded support. As an application, we show that the mean performance of the quantization of the empirical measure of the multidimensional normal distribution $\mathcal{N}(m, \Sigma)$ and of distributions with hyper-exponential tails behave like $\mathcal{O}(\frac{\log n}{\sqrt{n}})$. This extends the results from Biau et al. (2008) obtained for compactly supported distribution. We also derive a bound which is sharper in the quantization level K but suboptimal in n by applying results from Fournier and Guillin (2015).

Keyword: Clustering performance, Convergence rate of quantizers, Distortion function, Empirical measure, Optimal quantization.

3.1 Introduction

Let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^d$ induced by the canonical inner product $\langle\cdot|\cdot\rangle$ and the distance between a point ξ and a set A in $\mathbb{R}^d$ is defined by $d(\xi, A) = \min_{a\in A}|\xi - a|$.

For $p \in [1, +\infty)$, let $\mathcal{P}_p(\mathbb{R}^d)$ denote the set of all probability measures on $\mathbb{R}^d$ with a finite p^{th}-moment. Let X be an $\mathbb{R}^d$-valued random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with probability distribution $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. The (quadratic) *quantization procedure* of μ (or of X) at level $K \in \mathbb{N}^*$ consists in finding a discrete approximate quantizer $x = (x_1, ..., x_K) \in \left(\mathbb{R}^d\right)^K$ such that its quantization error

$$e_{K,\mu}(x) := \left[\mathbb{E}\min_{1\le i\le K}|X - x_i|^2\right]^{1/2}$$

achieves the *optimal quantization error* $e^*_{K,\mu}$ (or written $e^*_{K,X}$) for the distribution μ at level K, defined as follows,

$$e^*_{K,\mu} = \inf_{y=(y_1,...,y_K)\in(\mathbb{R}^d)^K}\left[\mathbb{E}\min_{1\le i\le K}|X - y_i|^2\right]^{\frac{1}{2}} = \inf_{y=(y_1,...,y_K)\in(\mathbb{R}^d)^K}\left[\int_{\mathbb{R}^d}\min_{1\le i\le K}|\xi - y_i|^2\,\mu(d\xi)\right]^{\frac{1}{2}}. \tag{3.1.1}$$

If $e_{K,\mu}(x) = e^*_{K,\mu}$, we call x an *optimal quantizer* (or called an *optimal cluster center*) of X (or *of* μ) at level K[(1)]. The function $x \in \left(\mathbb{R}^d\right)^K \mapsto e_{K,\mu}(x)$ is called the quantization error function. We denote by $G_K(\mu)$ the set of all optimal quantizers at level K of μ.

The distortion function is also often used to describe the quantization error of a quantizer $x \in (\mathbb{R}^d)^K$, defined as follows,

Definition 3.1.1 (Distortion function)**.** *Let $K \in \mathbb{N}^*$ be the quantization level. Let X be an $\mathbb{R}^d$-valued random variable with probability distribution $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. The (quadratic) distortion function $\mathcal{D}_{K,\mu}$ of μ at level K is defined on $(\mathbb{R}^d)^K \to \mathbb{R}_+$ by,*

$$x = (x_1, ..., x_K) \mapsto \mathcal{D}_{K,\mu}(x) = \mathbb{E}\min_{1\le k\le K}|X - x_k|^2 = \int_{\mathbb{R}^d}\min_{1\le i\le K}|\xi - x_i|^2\,\mu(d\xi). \tag{3.1.2}$$

(1) In many references, the quantizer at level K is defined by a set of points $\Gamma \subset \mathbb{R}^d$ with its cardinality $\mathrm{card}(\Gamma) \le K$ and the quadratic quantization error function is defined by $e_{K,\mu}(\Gamma) := \left[\mathbb{E}\,d(X,\Gamma)^2\right]^{1/2}$. However, for every $\Gamma = \{x_1, ..., x_{k'}\}$ with $k' \le K$, one can always find a K-tuple $x^\Gamma \in (\mathbb{R}^d)^K$ (by repeating some elements in Γ) such that $e_{K,\mu}(\Gamma) = e_{K,\mu}(x^\Gamma)$. For example, if $\Gamma = \{x_1, ..., x_{K-2}\}$ with $\mathrm{card}(\Gamma) = K - 2 \ge 1$ (the x_i are pointwise distinct), one may set $x^\Gamma = (x_1, x_1, x_1, x_2, ..., x_{K-2})$ or $(x_1, x_2, x_1, x_2, x_3..., x_{K-2})$ among many other possibilities.

In Graf and Luschgy (2000)[Theorem 4.12], the authors have proved that if the cardinality of the support of μ $\mathrm{card}\big(\mathrm{supp}(\mu)\big) \ge K$, an optimal quantizer Γ^* at quantization level K satisfies $\mathrm{card}\big(\mathrm{supp}(\Gamma^*)\big) = K$. Hence, $\inf_{\Gamma\subset\mathbb{R}^d,\,\mathrm{card}(\Gamma)\le K} e_{K,\mu}(\Gamma) = \inf_{x\in(\mathbb{R}^d)^K} e_{K,\mu}(x)$. Therefore, in this paper, with a slight abuse of notation, we will mostly use $x \in (\mathbb{R}^d)^K$ but also use (in Section 3.1.1) $\Gamma \subset \mathbb{R}^d$ with $\mathrm{card}(\Gamma) \le K$ to represent a quantizer at level K.

It is clear that for any quantizer $x \in (\mathbb{R}^d)^K$, $\mathcal{D}_{K,\mu}(x) = e^2_{K,\mu}(x)$. Hence, if $\mathrm{card}\big(\mathrm{supp}(\mu)\big) \geq K$, $G_K(\mu) = \mathrm{argmin}_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}$. Sometimes we withdraw the subscript K of $\mathcal{D}_{K,\mu}$ if the quantization level K is fixed in the context.

Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Let $\Pi(\mu,\nu)$ denote the set of all probability measures on $(\mathbb{R}^d \times \mathbb{R}^d, Bor(\mathbb{R}^d)^{\otimes 2})$ with marginals μ and ν. For $p \geq 1$, the L^p-Wasserstein distance $\mathcal{W}_p$ on $\mathcal{P}_p(\mathbb{R}^d)$ is defined by

$$\begin{aligned}\mathcal{W}_p(\mu,\nu) &= \Big(\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}^d\times\mathbb{R}^d} d(x,y)^p\pi(dx,dy)\Big)^{\frac{1}{p}} \\ &= \inf\Big\{\Big[\mathbb{E}\ |X-Y|^p\Big]^{\frac{1}{p}},\ X,Y:(\Omega,\mathcal{A},\mathbb{P})\to(\mathbb{R}^d,Bor(\mathbb{R}^d)) \text{ with } \mathbb{P}_X=\mu,\mathbb{P}_Y=\nu\Big\}.\end{aligned} \tag{3.1.3}$$

$\mathcal{P}_p(\mathbb{R}^d)$ equipped with Wasserstein distance $\mathcal{W}_p$ is a separable and complete space (see Bolley (2008)). If $\mu,\nu \in \mathcal{P}_p(\mathbb{R}^d)$, then for any $q \leq p$, $\mathcal{W}_q(\mu,\nu) \leq \mathcal{W}_p(\mu,\nu)$.

The target measure μ for the optimal quantization is sometimes unknown. In this case, in order to obtain the optimal quantizer of μ, we will implement the optimal quantization to a known distribution sequence $\mu_n, n \in \mathbb{N}^*$ which converges (in the Wasserstein distance) to μ and search the limiting point of optimal quantizers of μ_n. For $n \in \mathbb{N}^*$, let $x^{(n)}$ denote the optimal quantizer of μ_n. The consistency of $x^{(n)}$, i.e. $d\big(x^{(n)}, G_K(\mu)\big) \xrightarrow{n\to+\infty} 0$, has been proved by D. Pollard in Pollard (1982b)[see Theorem 9]. Therefore, a further question is, *at which rate the optimal quantizer $x^{(n)}$ of μ_n converges to an optimal quantizer x of μ?*

In the literature, there are two perspectives to study the convergence rate of optimal quantizers:

(i) The convergence rate of $d\big(x^n, G_K(\mu)\big)$;

(ii) The convergence rate of the distortion function of μ valued at $x^{(n)}$:

$$\mathcal{D}_{K,\mu}(x^{(n)}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x).$$

The latter quantity is also called the "quantization performance" (*performance* in short) at $x^{(n)}$ since this value describes how close between the optimal quantization error of μ and the quantization error of $x^{(n)}$, considered as a quantizer for μ (even $x^{(n)}$ is obviously not "optimal" for μ).

A typical example of what is described above is the quantization of the empirical measure. Let $X_1, ..., X_n, ...$ be i.i.d $\mathbb{R}^d$-valued observations of X with a unknown probability

distribution μ, then the empirical measure μ_n^ω is defined by:

$$\mu_n^\omega = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i(\omega)}, \tag{3.1.4}$$

where δ_a denotes the Dirac mass at a. The convergence of empirical measure $\mathcal{W}_p(\mu_n^\omega,\mu) \xrightarrow{a.s.} 0$ and $\mathbb{E}\,\mathcal{W}_2(\mu_n^\omega,\mu) \xrightarrow{n\to+\infty} 0$ have been proved in many reference, for example Pollard (1982b)[see Theorem 7] and Fournier and Guillin (2015)[see Theorem 1] so that we have the consistency for the optimal quantizers $x^{(n),\omega}$ of μ_n^ω. Moreover, most references about the convergence rate result for the optimal quantizers are concerning the empirical measure as far as we know: A first example is Pollard (1982a). In this paper, the author has proved that if x denotes the unique limiting point of $x^{(n),\omega}$, the convergence rate (convergence in law) of $\left|x^{(n),\omega}-x\right|$ is $\mathcal{O}(n^{-1/2})$ under appropriate conditions. For the second perspective, it is proved in a recent work Biau et al. (2008) that if μ has a support contained in B_R, where B_R denotes the ball in $\mathbb{R}^d$ centered at 0 with radius R, then $\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x) \le \frac{12K\cdot R^2}{\sqrt{n}}$.

In this paper, we will generalize these two precedent works.

In Section 3.2, we first establish a non-asymptotic upper bound for the convergence rate of the performance $\mathcal{D}_{K,\mu_\infty}(x^{(n)}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu_\infty}(x)$ for *any* probability distribution sequence μ_n converging in L^2-Wasserstein distance to μ_∞. We obtain for every $n\in\mathbb{N}^*$,

$$\mathcal{D}_{K,\mu_\infty}(x^{(n)}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu_\infty}(x) \le 4e^*_{K,\mu_\infty}\mathcal{W}_2(\mu_n,\mu_\infty) + 4\mathcal{W}_2^2(\mu_n,\mu_\infty). \tag{3.1.5}$$

Moreover, if $\mathcal{D}_{K,\mu_\infty}$ is twice differentiable on

$$F_K := \left\{x=(x_1,...,x_K)\in(\mathbb{R}^d)^K \;\middle|\; x_i\neq x_j,\ \text{if } i\neq j\right\} \tag{3.1.6}$$

and if the Hessian matrix $H_{\mathcal{D}_{K,\mu_\infty}}$ of $\mathcal{D}_{K,\mu_\infty}$ is positive definite in the neighboorhood of every optimal quantizer $x^{(\infty)}\in G_K(\mu_\infty)$ having the eigenvalues lower bounded by a $\lambda^*>0$, then for n large enough,

$$d\big(x^{(n)},G_K(\mu_\infty)\big)^2 \le \frac{8}{\lambda^*}e^*_{K,\mu_\infty}\cdot\mathcal{W}_2(\mu_n,\mu_\infty) + \frac{8}{\lambda^*}\cdot\mathcal{W}_2^2(\mu_n,\mu_\infty).$$

Several discussions around the Hessian matrix $H_{\mathcal{D}_{K,\mu}}$ of the distortion function $\mathcal{D}_{K,\mu}$ are established in Section 3.3. If $\mu\in\mathcal{P}_2(\mathbb{R}^d)$ with $\mathrm{card}\big(\mathrm{supp}(\mu)\big)\ge K$ and if μ is absolutely continuous with respect to Lebesgue measure having a continuous density function f, we prove in Section 3.3.1 that its distortion function $\mathcal{D}_{K,\mu}$ is twice differentiable in every $x\in F_K$ and give the exact formula of Hessian matrix. Moreover,

we also discuss several sufficient and necessary conditions for the positive definiteness of Hessian matrix in dimension $d \geq 2$ and in dimension 1.

Section 3.4 is devoted to the convergence rate of optimal quantization of the empirical measure. Let μ_n^ω be the empirical measure of μ defined in (3.1.4) and let $x^{(n),\omega}$ denote the optimal quantizer of μ_n^ω. In this section, we focus on the mean performance of $x^{(n),\omega}$, that is, $\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu}(x)$, which is also called the *clustering performance* in the field of unsupervised learning. If $\mu \in \mathcal{P}_q(\mathbb{R}^d)$ for some $q > 2$, the first result of Section 3.4 is Proposition 3.4.1, shown in the following formula, which is a direct application of the non-asymptotic upper bound (3.1.5) combined with the upper bound of the convergence rate (convergence in Wasserstein distance) of the empirical measure from Fournier and Guillin (2015).

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu}(x) \leq C_{d,q,\mu,K} \times \begin{cases} n^{-1/4} + n^{-(q-2)/2q} & \text{if } d < 4 \text{ and } q \neq 4 \\ n^{-1/4}\big(\log(1+n)\big)^{1/2} + n^{-(q-2)/2q} & \text{if } d = 4 \text{ and } q \neq 4 \\ n^{-1/d} + n^{-(q-2)/2q} & \text{if } d > 4 \text{ and } q \neq d/(d-2) \end{cases}.$$

where $C_{d,q,\mu,K}$ is a constant depending on d, q, μ and the quantization level K. Under certain conditions, this constant $C_{d,q,\mu,K}$ is roughly decreasing as $K^{-1/d}$ (see further Remark 3.4.1). This result is sharp in K but it suffers from the curse of dimensionality. Meanwhile, we establish another upper bound for the mean performance in Theorem 3.4.2, which is sharper in n, free from the curse of dimensionality but increasing faster than linearly in K. The main aim of this theorem is to generalize the mean performance result for the empirical measure of a distribution μ with bounded support established in Biau et al. (2008) to any distributions μ having simply a finite second moment. We obtain

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu}(x) \leq \frac{2K}{\sqrt{n}}\Big[r_{2n}^2 + \rho_K(\mu)^2 + 2r_1\big(r_{2n} + \rho_K(\mu)\big)\Big], \quad (3.1.7)$$

where $r_n := \big\|\max_{1\leq i\leq n} |X_i|\,\big\|_2$ and $\rho_K(\mu)$ is the maximum radius of $L^2(\mu)$-optimal quantizers, defined by

$$\rho_K(\mu) := \max\Big\{\max_{1\leq k\leq K} |x_k^*|\,,\ (x_1^*, ..., x_K^*) \text{ is an optimal quantizer of } \mu\Big\}. \quad (3.1.8)$$

Especially, we will give a precise upper bound for $\mu = \mathcal{N}(m, \Sigma)$, the multidimensionnal normal distribution

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu}(x) \leq C_\mu \cdot \frac{2K}{\sqrt{n}}\Big[1 + \log n + \gamma_K \log K\big(1 + \frac{2}{d}\big)\Big], \quad (3.1.9)$$

where $\limsup_K \gamma_K = 1$ and $C_\mu = 12 \cdot \Big[1 \vee \log\big(2 \int_{\mathbb{R}^d} \exp(\frac{1}{4}|\xi|^4)\mu(d\xi)\big)\Big]$. If $\mu = \mathcal{N}(0, \mathbf{I}_d)$, $C_\mu = 12(1 + \frac{d}{2}) \cdot \log 2$.

We will start our discussion with a brief review on the properties of optimal quantizer and the distortion function.

3.1.1 Properties of the Optimal quantizer and the Distortion Function

Let X be an $\mathbb{R}^d$-valued random variable with probability distribution μ such that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mathrm{card}\big(\mathrm{supp}(\mu)\big) \geq K$. Let $G_K(\mu)$ denote the set of all optimal quantizers at level K of μ and let $e^*_{K,\mu}$ denote the optimal quantization error of μ defined in (3.1.1). The properties below recall some classical background on optimal quantization of probability measure.

Proposition 3.1.1. *Let $K \in \mathbb{N}^*$. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mathrm{card}\big(\mathrm{supp}(\mu)\big) \geq K$.*

(i) (Decreasing of $K \mapsto e^*_{K,\mu}$) If $K \geq 2$, $e^*_{K,\mu} < e^*_{K-1,\mu}$.

(ii) (Existence and boundedness of optimal quantizers) The set

$$G'_K(\mu) := \big\{\Gamma^x = \{x_1, ..., x_K\} \;\big|\; x = (x_1, ..., x_K) \in \mathrm{argmin}\, \mathcal{D}_{K,\mu}\big\}$$

is nonempty and compact so that $\rho_K(\mu)$ defined in (3.1.8) is finite for any fixed K. Moreover, if $\Gamma^* \subset \mathbb{R}^d$ is an optimal quantizer of μ, then $\mathrm{card}(\Gamma^*) = K$. In particular, if $\Gamma^* = \{x_1, ..., x_K\}$, then $x^{\Gamma^*} := (x_1, ..., x_K) \in \mathrm{argmin}\, \mathcal{D}_{K,\mu} = G_K(\mu)$ and vice versa.

(iii) If μ has a compact support and if the norm $|\cdot|$ on $\mathbb{R}^d$ is Euclidean, drived by an inner product $\langle \cdot | \cdot \rangle$, then all the optimal quantizers $\Gamma^* = \{x_1, ..., x_K\}$ are contained in the closure of convex hull of $\mathrm{supp}(\mu)$, denoted by $\mathcal{H}_\mu := \overline{\mathrm{conv}\big(\mathrm{supp}(\mu)\big)}$.

For the proof of Proposition 3.1.1-(i) and (ii), we refer to Graf and Luschgy (2000)[see Theorem 4.12] and for the proof of (iii) to Appendix A.

Theorem 3.1.1. *(Non-asymptotic Zador's theorem) Let $\eta > 0$. If $\mu \in \mathcal{P}_{2+\eta}$, then for every quantization level K, there exists a constant $C_{d,\eta} \in (0, +\infty)$ which depends only on d and η such that*

$$e^*_{K,\mu} \leq C_{d,\eta} \cdot \sigma_{2+\eta}(\mu) K^{-1/d}, \tag{3.1.10}$$

where for $r \in (0, +\infty)$, $\sigma_r(\mu) = \min_{a \in \mathbb{R}^d} \big[\int_{\mathbb{R}^d} |\xi - a|^r \, \mu(d\xi)\big]^{1/r}$.

For the proof of non-asymptotic Zador's theorem, we refer to Luschgy and Pagès (2008) and Pagès (2018)[see Theorem 5.2]. Now we introduce some properties of $\rho_K(\mu)$ defined in (3.1.8). When μ has an unbounded support, we know from Pagès and Sagna (2012) that $\lim_K \rho_K(\mu) = +\infty$. The same paper also gives an asymptotic upper bound of ρ_K when μ has a polynomial tail or hyper-exponential tail. We first give the definitions of different tails of probability measure,

Definition 3.1.2. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ be absolutely continuous with respect to Lebesgue measure λ_d on $\mathbb{R}^d$ and let f denote its density function.*

(i) A distribution μ has a k-th radial-controlled tail if there exists $A > 0$ and a continuous and decreasing function $g : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\forall \xi \in \mathbb{R}^d, |\xi| \geq A, \qquad f(\xi) \leq g(|\xi|) \text{ and } \int_{\mathbb{R}_+} x^k g(x) dx < +\infty.$$

(ii) A distribution μ has a c-th polynomial tail if there exists $\tau > 0, \beta \in \mathbb{R}, c > d$ and $A > 0$ such that $\forall \xi \in \mathbb{R}^d, |\xi| \geq A \Longrightarrow f(\xi) = \frac{\tau}{|\xi|^c} (\log |\xi|)^\beta$.

(iii) A distribution μ has a (ϑ, κ)-hyper-exponential tail if there exists $\tau > 0, \kappa, \vartheta > 0, c > -d$ and $A > 0$ such that $\forall \xi \in \mathbb{R}^d, |\xi| \geq A \Longrightarrow f(\xi) = \tau |\xi|^c e^{-\vartheta |\xi|^\kappa}$.

The purpose of the definition of *radial-controlled tail* is to *control* the convergence rate of the density function $f(x)$ to 0 when x converges in every direction to infinity. Remark that the c-th polynomial tail with $c > k + 1$ and the hyper-exponential tail are sufficient conditions to k-th radial-controlled tail. A typical example of hyper-exponential tail is the multidimensional normal distribution $\mathcal{N}(m, \Sigma)$.

Theorem 3.1.2. *(Pagès and Sagna (2012)[see Theorem 1.2]) Assume that $\mu = f \cdot \lambda_d$*

(i) Polynomial tail. *For $p \geq 2$, if μ has a c-th polynomial tail with $c > d + p$, then*

$$\lim_K \frac{\log \rho_K}{\log K} = \frac{p + d}{d(c - p - d)}. \tag{3.1.11}$$

(ii) Hyper-exponential tail. *If μ has a (ϑ, κ)-hyper-exponential tail, then*

$$\limsup_K \frac{\rho_K}{(\log K)^{1/\kappa}} \leq 2\vartheta^{-1/\kappa} \left(1 + \frac{2}{d}\right)^{1/\kappa}. \tag{3.1.12}$$

Furthermore, if $d = 1$, $\lim_K \frac{\rho_K}{(\log K)^{1/\kappa}} = \left(\frac{3}{\vartheta}\right)^{1/\kappa}$.

Quantization theory has a close connection with Voronoï partitions. Let $x = (x_1, ..., x_K)$ be a quantizer at level K and let $|\cdot|$ be any norm on $\mathbb{R}^d$. The *Voronoï cell* (or *Voronoï region*) generated by x_i is defined by

$$V_{x_i}(x) = \big\{\xi \in \mathbb{R}^d : |\xi - x_i| = \min_{1\le j\le K} |\xi - x_j|\,\big\}, \tag{3.1.13}$$

and $\big(V_{x_i}(x)\big)_{1\le i\le K}$ is called the *Voronoï diagram* of Γ, which is a locally finite covering of $\mathbb{R}^d$. A Borel partition $\big(C_{x_i}(x)\big)_{1\le i\le K}$ is called a *Voronoï partition* of $\mathbb{R}^d$ induced by x if

$$\forall i \in \{1, ..., K\}, \qquad C_{x_i}(x) \subset V_{x_i}(x). \tag{3.1.14}$$

We also define the *open Voronoï cell* generated by x_i by

$$V^o_{x_i}(x) = \big\{\xi \in \mathbb{R}^d : |\xi - x_i| < \min_{1\le j\le K, j\ne i} |\xi - x_j|\,\big\}. \tag{3.1.15}$$

As $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$, we know from Graf and Luschgy (2000)[see Proposition 1.3] that $\mathrm{int}V_{x_i}(x) = V^o_{x_i}(x)$, where $\mathrm{int}A$ denotes the interior of a set A. Moreover, if we denote by λ_d the Lebesgue measure on $\mathbb{R}^d$, we have $\lambda_d\big(\partial V_{x_i}(x)\big) = 0$, where ∂A denotes the boundary of A (see Graf and Luschgy (2000)[Theorem 1.5]). If $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and x^* is an optimal quantizer of μ, even if μ is not absolutely continuous with the respect of λ_d, we have $\mu\big(\partial V_{x_i}(x^*)\big) = 0$ for all $i \in \{1, ..., K\}$ (see Graf and Luschgy (2000)[Theorem 4.2]).

For any K-tuple $x = (x_1, ..., x_K) \in (\mathbb{R}^d)^K$ such that $x_i \ne x_j$, $i \ne j$, one can rewrite the distortion function $\mathcal{D}_{K,\mu}$ with the definition of Voronoï partition $C_{x_i}(x)$ as follows,

$$\mathcal{D}_{K,\mu}(x) = \sum_{i=1}^{K} \int_{C_{x_i}(x)} |\xi - x_i|^2\, \mu(d\xi). \tag{3.1.16}$$

If $x^* = (x^*_1, ..., x^*_K) \in \mathrm{argmin}\, \mathcal{D}_{K,\mu}$, we know from Proposition 3.1.1 that $x^*_i \ne x^*_j$, $i \ne j$ and we have $\mu\big(\partial V_{x_i}(x^*)\big) = 0$. In this case, $\mathcal{D}_{K,\mu}$ is differentiable at x^* (see Pagès (2018)[Chapter 5]) and its gradient is given by

$$\nabla \mathcal{D}_{K,\mu}(x^*) = 2\Big[\int_{C_i(x^*)} (x^*_i - \xi)\mu(d\xi)\Big]_{i=1,...,K}. \tag{3.1.17}$$

For $\mu, \nu \in P_2(\mathbb{R}^d)$, if we denote by $\mathcal{D}_{K,\mu}$ the distortion function of μ and $\mathcal{D}_{K,\nu}$ the distortion function of ν. Then, for every $K \in \mathbb{N}^*$,

$$\Big\|\mathcal{D}^{1/2}_{K,\mu} - \mathcal{D}^{1/2}_{K,\nu}\Big\|_{\sup} := \sup_{x\in(\mathbb{R}^d)^K} \Big|\mathcal{D}^{1/2}_{K,\mu}(x) - \mathcal{D}^{1/2}_{K,\nu}(x)\Big| \le \mathcal{W}_2(\mu, \nu), \tag{3.1.18}$$

by a simple application of the triangle inequality for the L^2-norm (see Graf and Luschgy (2000) Formula (4.4) and Lemma 3.4). Hence, if $(\mu_n)_{n\geq 1}$ is a sequence in $\mathcal{P}_2(\mathbb{R}^d)$ converging for the $\mathcal{W}_2$-distance to $\mu_\infty \in \mathcal{P}_2(\mathbb{R}^d)$, then for every $K \in \mathbb{N}^*$

$$\left\| \mathcal{D}_{K,\mu_n}^{1/2} - \mathcal{D}_{K,\mu_\infty}^{1/2} \right\|_{\sup} \leq \mathcal{W}_2(\mu_n, \mu_\infty) \xrightarrow{n\to+\infty} 0. \tag{3.1.19}$$

Let $\mu_n, n \in \mathbb{N}^*, \mu_\infty \in \mathcal{P}_2(\mathbb{R})$ such that $\mathcal{W}_2(\mu_n, \mu_\infty) \xrightarrow{n\to+\infty} 0$. For a fixed quantization level $K \in \mathbb{N}^*$, the consistency of optimal quantizers is firstly established by D. Pollard by using

$$\mu_K \in \mathcal{P}(K) := \left\{ \nu \in \mathcal{P}_2(\mathbb{R}^d) \text{ such that } \mathrm{card}\big(\mathrm{supp}(\nu)\big) \leq K \right\}$$

to represent a quantization "quantizer" at level K and μ_K is called "optimal" for a probability mesure μ if $\mathcal{W}_2(\mu_K, \mu) = e^*_{K,\mu}(\mu)$. We will annonce differently the consistency theorem by letting $x^{(n)} = (x_1^{(n)}, ..., x_K^{(n)}) \in (\mathbb{R}^d)^K$ to represent the optimal quantizer of μ_n (of course we still call the theorem "Pollard's Theorem") and we will give the proof of Pollard's Theorem with respect of this representation to Annex B.

Theorem 3.1.3 (Pollard's Theorem)**.** *Let $K \in \mathbb{N}^*$ be the quantization level. Let $\mu_n, \mu_\infty \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\mathcal{W}_2(\mu_n, \mu_\infty) \to 0$. Assume $\mathrm{card}\big(\mathrm{supp}(\mu_n)\big) \geq K$, for $n \in \mathbb{N}^* \cup \{+\infty\}$. For $n \geq 1$, let $x^{(n)} = \big(x_1^{(n)}, ..., x_K^{(n)}\big)$ be a K-optimal quantizer for μ_n, then the quantizer sequence $(x^{(n)})_{n\geq 1}$ is bounded in $\mathbb{R}^d$ and any limiting point of $(x^{(n)})_{n\geq 1}$, denoted by $x^{(\infty)}$, is an optimal quantizer of μ_∞.*

3.2 General case

Let $\mu_n, n \in \mathbb{N}^*, \mu_\infty \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\mathcal{W}_2(\mu_n, \mu_\infty) \to 0$ as $n \to 0$. Fix a quantization level $K \in \mathbb{N}^*$ through this section. For every $n \in \mathbb{N}^*$, let $x^{(n)} = (x_1^{(n)}, ..., x_K^{(n)}) \in \mathrm{argmin}_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu_n}$ which is, after Proposition 3.1.1 - (ii), an optimal quantizer of μ_n at level K. In this section, we first establish a non-asymptotic upper bound of the convergence rate for the quantization performance $\mathcal{D}_{K,\mu_\infty}(x^{(n)}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu_\infty}(x)$. Then we discuss the convergence rate of $d(x^{(n)}, G_K(\mu))$.

Theorem 3.2.1 (Non-asymptotic convergence rate for the quantization performance)**.** *Let $K \in \mathbb{N}^*$ be the fixed quantization level. For every $n \in \mathbb{N}^* \cup \{\infty\}$, let $\mu_n \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mathrm{card}\big(\mathrm{supp}(\mu_n)\big) \geq K$ such that $\mathcal{W}_2(\mu_n, \mu_\infty) \to 0$ as $n \to +\infty$. For every $n \in \mathbb{N}^*$, let $x^{(n)}$ be an optimal quantizer of μ_n. Then*

(i) $e_{K,\mu_\infty}(x^{(n)}) - e^*_{K,\mu_\infty} \leq 2\mathcal{W}_2(\mu_n, \mu_\infty)$

$$(ii) \quad \mathcal{D}_{K,\mu_\infty}(x^{(n)}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu_\infty}(x) \le 4e^*_{K,\mu_\infty}\mathcal{W}_2(\mu_n,\mu_\infty) + 4\mathcal{W}_2^2(\mu_n,\mu_\infty),$$

where e^*_{K,μ_∞} *is the optimal quantization error of* μ_∞ *at level* K.

Proof of Theorem 3.2.1. Let $x^{(\infty)}$ be an optimal quantizer of μ_∞. Remark that we don't need to $\left|x^{(n)} - x^{(\infty)}\right| \xrightarrow{n\to+\infty} 0$. Then

$$\begin{aligned} e_{K,\mu_\infty}(x^{(n)}) - e^*_{K,\mu_\infty} &= e_{K,\mu_\infty}(x^{(n)}) - e_{K,\mu_n}(x^{(n)}) + e_{K,\mu_n}(x^{(n)}) - e_{K,\mu_\infty}(x^{(\infty)}) \\ &\le 2\left\|e_{K,\mu_\infty} - e_{K,\mu_n}\right\|_{\sup} \le 2\mathcal{W}_2(\mu_n,\mu_\infty), \end{aligned} \tag{3.2.1}$$

where the first inequality is due to the fact that for any $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^d)$ with respective K-level optimal quantizers x^μ and x^ν, if $e_{K,\mu}(x^\mu) \ge e_{K,\nu}(x^\nu)$, we have

$$|e_{K,\mu}(x^\mu) - e_{K,\nu}(x^\nu)| = e_{K,\mu}(x^\mu) - e_{K,\nu}(x^\nu) \le e_{K,\mu}(x^\nu) - e_{K,\nu}(x^\nu) \le \left\|e_{K,\mu_\infty} - e_{K,\mu_n}\right\|_{\sup}.$$

If $e_{K,\mu}(x^\mu) \le e_{K,\nu}(x^\nu)$, we have the same inequality by the same reasoning[(1)].

Moreover,

$$\begin{aligned} \mathcal{D}_{K,\mu_\infty}(x^{(n)}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu_\infty}(x) &= \mathcal{D}_{K,\mu_\infty}(x^{(n)}) - \mathcal{D}_{K,\mu_\infty}(x^{(\infty)}) \\ &\le \left[\mathcal{D}^{1/2}_{K,\mu_\infty}(x^{(n)}) + \mathcal{D}^{1/2}_{K,\mu_\infty}(x^{(\infty)})\right]\left(e_{K,\mu_\infty}(x^{(n)}) - e_{K,\mu_\infty}(x^{(\infty)})\right) \\ &\le 2\left[\mathcal{D}^{1/2}_{K,\mu_\infty}(x^{(n)}) - \mathcal{D}^{1/2}_{K,\mu_\infty}(x^{(\infty)}) + 2\mathcal{D}^{1/2}_{K,\mu_\infty}(x^{(\infty)})\right]\cdot\mathcal{W}_2(\mu_n,\mu_\infty) \quad \big(\text{by } (3.2.1)\big) \\ &\le 4\left[\mathcal{W}_2(\mu_n,\mu_\infty) + e^*_{K,\mu_\infty}\right]\cdot\mathcal{W}_2(\mu_n,\mu_\infty) \quad \big(\text{by } (3.2.1)\big) \\ &\le 4e^*_{K,\mu_\infty}\mathcal{W}_2(\mu_n,\mu_\infty) + 4\mathcal{W}_2^2(\mu_n,\mu_\infty). \end{aligned}$$

□

Before we establish the convergence rate of the optimal quantizer sequence $x^{(n)}, n \in \mathbb{N}$, we first discuss the differentiability of $\mathcal{D}_{K,\mu}$. Let $B(x,r)$ denote the ball centered at x with radius r. Remark that if $x \in F_K$, where F_K is defined in (3.1.6), then every $y \in B\big(x, \frac{1}{3}\min_{1\le i,j\le K, i\ne j}|x_i - x_j|\big)$ lies still in F_K (see Section 3.5.3 for the proof).

Lemma 3.2.1. *Let* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ *with* $\mathrm{card}\big(\mathrm{supp}(\mu)\big) \ge K$. *If the probability distribution* μ *is absolutely continuous with respect to Lebesgue measure and has a continuous density function* f, *written by* $\mu(d\xi) = f(\xi)\lambda_d(d\xi)$, *then its distortion function* $\mathcal{D}_{K,\mu}$ *is twice differentiable in every* $x \in F_K$.

(1) This part of preuve also appears in Linder (2002)[Corollary 4.1].

The proof of Lemma 3.2.1 is postponed in Section 3.3.1 in which we also give the exact formula of the Hessian matrix. In the following theorem we show the convergence rate of the optimal quantizer sequence $x^{(n)}, n \in \mathbb{N}^*$.

Theorem 3.2.2 (Convergence rate of optimal quantizers)**.** *Let $K \in \mathbb{N}^*$ be the fixed quantization level. For every $n \in \mathbb{N}^* \cup \{\infty\}$, let $\mu_n \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mathrm{card}\big(\mathrm{supp}(\mu_n)\big) \geq K$ such that $\mathcal{W}_2(\mu_n, \mu_\infty) \to 0$ as $n \to +\infty$. For every $n \in \mathbb{N}^*$, let $x^{(n)}$ be an optimal quantizer of μ_n and let $G_K(\mu_\infty)$ denote the set of all optimal quantizers of μ_∞. If*

(a) the probability distribution μ_∞ is absolutely continuous with respect to Lebesgue measure λ_d and has a continuous density function f, written by $\mu_\infty(d\xi) = f(\xi)\lambda_d(d\xi)$,

(b) for every $x^{(\infty)} \in G_K(\mu_\infty)$, the Hessian matrix of $\mathcal{D}_{K,\mu_\infty}$, denoted by $H_{\mathcal{D}_{K,\mu_\infty}}$, is positive definite in the neighbourhood of $x^{(\infty)}$ having eigenvalues lower bounded by some $\lambda^* > 0$,

then, for n large enough,

$$d\big(x^{(n)}, G_K(\mu_\infty)\big)^2 \leq \frac{8}{\lambda^*} e^*_{K,\mu_\infty} \cdot \mathcal{W}_2(\mu_n, \mu_\infty) + \frac{8}{\lambda^*} \cdot \mathcal{W}_2^2(\mu_n, \mu_\infty).$$

Remark 3.2.1. (1) Owing to Lemma 3.2.1 and Proposition 3.1.1-(ii), the Condition (a) in the above theorem implies that the distortion function $\mathcal{D}_{K,\mu_\infty}$ is twice differentiable in every $x^{(\infty)} \in G_K(\mu_\infty)$ and its neighbourhood so that the use of the Hessian matrix $H_{\mathcal{D}_\infty}$ in Condition (b) is permitted. However, the conditions (b) is not obvious to satisfy. In Section 3.3, we give an exact formula of the Hessian matrix $H_{\mathcal{D}_{K,\mu_\infty}}$. Thus, one may obtain the positive definiteness of Hessian matrix $H_{\mathcal{D}_{K,\mu_\infty}}$ (condition (b)) by an explicite computation or by numerical methods. Moreover, in Section 3.3, we also establish a sufficient condition for the continuity of every term in the Hessian matrix in dimension d and several sufficient conditions for the positive definiteness of the Hessian matrix $H_{\mathcal{D}_{K,\mu_\infty}}$ in the neighbourhood of $x^{(\infty)} \in G_K(\mu_\infty)$ in dimension 1.

(2) If the distribution μ_∞ is d-th radial-controlled, a necessary condition of Condition (b) is $\mathrm{card}\big(G_K(\mu_\infty)\big) < +\infty$ (we will prove this statement later in Lemma 3.3.3). Thus, if $\mathrm{card}\big(G_K(\mu_\infty)\big) = +\infty$, it is better to use the non-asymtotic upper bound of the performance (Theorem 3.2.1) as a tool to study the convergence rate of optimal quantization. A typical example is $\mu_\infty = \mathcal{N}(0, I_d)$, the standard multidimensional normal distribution: it is d-th radial-controlled and any rotation of an optimal quantizer x is still an optimal quantizer so that $\mathrm{card}\big(G_K(\mu_\infty)\big) = +\infty$.

Proof of Theorem 3.2.2. Since the quantization level K is fixed throughout the proof, we will drop the subscripts K and μ of the distortion function $\mathcal{D}_{K,\mu}$ and we will denote by $\mathcal{D}_n$ (*respectively,* $\mathcal{D}_\infty$) the distortion function of μ_n (*resp.* of μ_∞).

After Pollard's theorem in Section 3.1.1, $(x^{(n)})_{n\in\mathbb{N}^*}$ is bounded and any limiting point of $x^{(n)}$ is in $G_K(\mu_\infty)$. We may assume that, up to a subsequence of $x^{(n)}$, still denoted by $x^{(n)}$, we have $x^{(n)} \to x^{(\infty)} \in G_K(\mu_\infty)$. Hence $d\big(x^{(n)}, G_K(\mu_\infty)\big) \le \big|x^{(n)} - x^{(\infty)}\big|$.

By Lemma 3.2.1 and Proposition 3.1.1-(ii), Condition (a) implies that the distortion function $\mathcal{D}_\infty$ is twice-differentiable at $x^{(\infty)}$. Hence, the Taylor expansion of $\mathcal{D}_\infty$ at $x^{(\infty)}$ reads:

$$\mathcal{D}_\infty(x^{(n)}) = \mathcal{D}_\infty(x^{(\infty)}) + \big\langle \nabla\mathcal{D}_\infty(x^{(\infty)}) \mid x^{(n)} - x^{(\infty)} \big\rangle + \frac{1}{2} H_{\mathcal{D}_\infty}(\zeta^{(n)})(x^{(n)} - x^{(\infty)})^{\otimes 2},$$

where $H_{\mathcal{D}_\infty}$ denotes the Hessian matrix of $\mathcal{D}_\infty$, $\zeta^{(n)}$ lies in the geometric segment $(x^{(n)}, x^{(\infty)})$, and for a matrix A and a vector u, $Au^{\otimes 2}$ stands for $u^T Au$.

As $x^{(\infty)} \in G_K(\mu_\infty) = \operatorname{argmin} \mathcal{D}_\infty$ and $\operatorname{card}\big(\operatorname{supp}(\mu_\infty)\big) \ge K$, one has $\nabla\mathcal{D}_\infty(x^{(\infty)}) = 0$ by applying Fermat's theorem on stationary point. Hence

$$\mathcal{D}_\infty(x^{(n)}) - \mathcal{D}_\infty(x^{(\infty)}) = \frac{1}{2} H_{\mathcal{D}_\infty}(\zeta^{(n)})(x^{(n)} - x^{(\infty)})^{\otimes 2}. \tag{3.2.2}$$

It follows that

$$\begin{aligned} H_{\mathcal{D}_\infty}(\zeta^{(n)})(x^{(n)} - x^{(\infty)})^{\otimes 2} &= 2\big(\mathcal{D}_\infty(x^{(n)}) - \mathcal{D}_\infty(x^{(\infty)})\big) \\ &\le 8e^*_{K,\mu_\infty} \mathcal{W}_2(\mu_n, \mu_\infty) + 8\mathcal{W}_2^2(\mu_n, \mu_\infty). \end{aligned} \tag{3.2.3}$$

By condition (b), $H_{\mathcal{D}_\infty}(x)$ is assumed to be positive definite in the neighbourhood of all $x^{(\infty)} \in G_K(\mu_\infty)$ having eigenvalues lower bounded by some λ^*. Since $\zeta^{(n)}$ lies in the geometric segment $(x^{(n)}, x^{(\infty)})$ and $x^{(n)} \to x^{(\infty)}$, then there exists an $n_0(x^{(\infty)})$ such that for all $n \ge n_0$, $H_{\mathcal{D}_\infty}(\zeta^{(n)})$ is a positive definite matrix. It follows that for $n \ge n_0$,

$$\begin{aligned} \lambda^* \big|x^{(n)} - x^{(\infty)}\big|^2 &\le H_{\mathcal{D}_\infty}(\zeta^{(n)})(x^{(n)} - x^{(\infty)})^{\otimes 2} \\ &\le 8e^*_{K,\mu_\infty} \mathcal{W}_2(\mu_n, \mu_\infty) + 8\mathcal{W}_2^2(\mu_n, \mu_\infty). \end{aligned}$$

Thus, one can directly conclude by multiplying $\frac{1}{\lambda^*}$ at each side of the above inequality. □

Based on conditions in Theorem 3.2.2, if moreover, we know the exact limit of the optimal quantizer sequence $x^{(n)}$,we have the following result whose proof is similar to the proof of Theorem 3.2.2.

Corollary 3.2.1. *Let $\mu_n, \mu_\infty \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mathcal{W}_2(\mu_n, \mu_\infty) \to 0$ as $n \to +\infty$. Assume that* $\operatorname{card}\big(\operatorname{supp}(\mu_n)\big) \ge K$ *for every $n \in \mathbb{N}^* \cup \{\infty\}$. Let $x^{(n)} \in \operatorname{argmin} \mathcal{D}_{K,\mu_n}$ such that $\lim_n x^{(n)} \to x^{(\infty)}$. If the Hessian matrix of $\mathcal{D}_{K,\mu_\infty}$ is a positive definite matrix in the*

neighbourhood of $x^{(\infty)}$*, then for* n *large enough*

$$\left|x^{(n)} - x^{(\infty)}\right|^2 \leq C^{(1)}_{\mu_\infty} \cdot \mathcal{W}_2(\mu_n, \mu_\infty) + C^{(2)}_{\mu_\infty} \cdot \mathcal{W}_2^2(\mu_n, \mu_\infty),$$

where $C^{(1)}_{\mu_\infty}$ *and* $C^{(2)}_{\mu_\infty}$ *are constants only depending on* μ_∞*.*

3.3 Hessian matrix $H_{\mathcal{D}_{K,\mu}}$ of the distortion function $\mathcal{D}_{K,\mu}$

Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\text{card}\big(\text{supp}(\mu)\big) \geq K$ and let x^* be an optimal quantizer of μ at level K. In Section 3.3.1, we prove Lemma 3.2.1 by giving the exact formula for the Hessian matrix $H_{\mathcal{D}_{K,\mu}}$ of the distortion function $\mathcal{D}_{K,\mu}$ when μ is absolutely continuous with the respect of Lebesgue measure λ_d on $\mathbb{R}^d$, having a continuous density function f. Moreover, we also give a sufficient condition for the continuity of every term of the Hessian matrix $H_{\mathcal{D}_{K,\mu}}$ and a necessary condition for the positive definitiveness of the Hessian matrix $H_{\mathcal{D}_{K,\mu}}(x^*)$. Next, in Section 3.3.2, we give several sufficient conditions for the positive definiteness of the Hessian matrix $H_{\mathcal{D}_{K,\mu}}$ in the neighbourhood of x^* in dimension 1.

3.3.1 Hessian matrix $H_{\mathcal{D}_{K,\mu}}$ on $\mathbb{R}^d$

If μ is absolutely continuous with the respect of Lebesgue measure λ_d on $\mathbb{R}^d$ with the density function$f$, $\mathcal{D}_{K,\mu}$ is differentiable (see Pagès (1998)) and at all point $x = (x_1, ..., x_K) \in F_K$ with

$$\frac{\partial \mathcal{D}_{K,\mu}}{\partial x_i}(x) = 2\int_{V_i(x)} (x_i - \xi) f(\xi) \lambda_d(d\xi), \quad \text{for } i = 1, ..., K. \tag{3.3.1}$$

Now we use Lemma 11 in Fort and Pagès (1995) to compute the Hessian matrix $H_{\mathcal{D}_{K,\mu}}$ of $\mathcal{D}_{K,\mu}$.

Lemma 3.3.1 (Lemma 11 in Fort and Pagès (1995))**.** *Let* φ *be a countinous* $\mathbb{R}$*-valued function defined on* $[0,1]^d$*. For every* $x \in D_K := \big\{y \in \big([0,1]^d\big)^K \mid y_i \neq y_j \text{ if } i \neq j\big\}$*, let* $\Phi_i(x) := \int_{V_i(x)} \varphi(\omega) d\omega$*. Then* Φ_i *is continuously differentiable on* D_K *and*

$$\forall i \neq j, \ \frac{\partial \Phi_i}{\partial x_j}(x) = \int_{V_i(x) \cap V_j(x)} \varphi(\omega) \Big\{ \frac{1}{2} \overrightarrow{n}^{ij}_x + \frac{1}{|x_j - x_i|} \times \big(\frac{x_i + x_j}{2} - \omega\big) \Big\} \lambda^{ij}_x(d\omega) \tag{3.3.2}$$

$$\text{and } \frac{\partial \Phi}{\partial x_i}(x) = - \sum_{1 \leq j \leq K, j \neq i} \frac{\partial \Phi_j}{\partial x_i}(x), \tag{3.3.3}$$

where $\overrightarrow{n}_x^{ij} := \frac{x_j - x_i}{|x_j - x_i|}$,

$$M_{ij}^x := \Big\{ u \in \mathbb{R}^d \mid \langle u - \frac{x_i + x_j}{2} \mid x_i - x_j \rangle = 0 \Big\} \tag{3.3.4}$$

and $\lambda_x^{ij}(d\omega)$ *the Lebesgue measure on* M_{ij}^x.

One can simplify the result of Lemma 3.3.1 as follows,

$$\begin{aligned} \forall i \neq j, \quad \frac{\partial \Phi_i}{\partial x_j}(x) &= \int_{V_i(x) \cap V_j(x)} \varphi(\omega) \Big\{ \frac{1}{2} \frac{x_j - x_i}{|x_j - x_i|} + \frac{1}{|x_j - x_i|} \big(\frac{x_i + x_j}{2} - \omega \big) \Big\} \lambda_x^{ij}(d\omega) \\ &= \int_{V_i(x) \cap V_j(x)} \varphi(\omega) \frac{1}{|x_j - x_i|} \Big\{ \frac{x_j - x_i}{2} + \frac{x_i + x_j}{2} - \omega \Big\} \lambda_x^{ij}(d\omega) \\ &= \int_{V_i(x) \cap V_j(x)} \varphi(\omega) \frac{1}{|x_j - x_i|} (x_j - \omega) \lambda_x^{ij}(d\omega). \end{aligned} \tag{3.3.5}$$

Now we prove Lemma 3.2.1 and give the exact formula of the Hessian matrix $H_{\mathcal{D}_{K,\mu}}$ in the proof.

Proof of Lemma 3.2.1. Set $\varphi^i(\xi) = (x_i - \xi) f(\xi)$ and $\Phi_i(x) = \int_{V_i(x)} \varphi(\xi) d\xi = \frac{\partial \mathcal{D}_{K,\mu}}{\partial x_i}$ for $i = 1, ..., K$. It follows from Lemma 3.3.1 that for $j = 1, ..., K$ and $j \neq i$

$$\frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_j \partial x_i}(x) = 2 \frac{\partial \Phi_i(x)}{\partial x_j} = 2 \int_{V_i(x) \cap V_j(x)} (x_i - \xi) \otimes (x_j - \xi) \cdot \frac{1}{|x_j - x_i|} f(\xi) \lambda_x^{ij}(d\xi), \tag{3.3.6}$$

and for $i = 1, ..., K$,

$$\frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_i^2}(x) = \frac{\partial \Phi_i(x)}{\partial x_i} = 2 \Big[\mu(V_i(x)) \mathrm{I}_d - \sum_{\substack{i \neq j \\ 1 \leq j \leq K}} \int_{V_i(x) \cap V_j(x)} (x_i - \xi) \otimes (x_i - \xi) \cdot \frac{1}{|x_j - x_i|} f(\xi) \lambda_x^{ij}(d\xi) \Big], \tag{3.3.7}$$

where in (3.3.6) and (3.3.7), $u \otimes v := [u^i v^j]_{1 \leq i,j \leq d}$ for any two vectors $u = (u^1, ..., u^d)$ and $v = (v^1, ..., v^d)$ in $\mathbb{R}^d$. □

Next, we show in the following lemma a sufficient condition to the continuity of the Hessian matrix $H_{\mathcal{D}_{K,\mu}}$ in F_K so that under this condition, if the Hessian matrix $H_{\mathcal{D}_{K,\mu}}$ is positive definite in x^*, it is also positive definite in the neighbourhood of x^*. The proof of Lemma 3.3.2 is in Appendix C.

Lemma 3.3.2. *Let* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ *be absolutely continuous with the respect to Lebesgue measure* λ_d *on* $\mathbb{R}^d$ *with a continuous density function* f. *If* μ *has a* d*-th radial-controlled tail, then every element of the Hessian matrix* $H_{\mathcal{D}_{K,\mu}}$ *of the distortion function* $\mathcal{D}_{K,\mu}$ *is a continuous function.*

Under the condition of Lemma 3.3.2, we prove now that Condition (b) in Theorem 3.2.2 implies $\mathrm{card}\big(G_K(\mu_\infty)\big) < +\infty$.

Lemma 3.3.3. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ be absolutely continuous with the respect to Lebesgue measure λ_d on $\mathbb{R}^d$ with a continuous density function f. If μ_∞ has a d-th radial-controlled tail and $\mathrm{card}\big(G_K(\mu_\infty)\big) = +\infty$, then there exists a point $x \in G_K(\mu_\infty)$ such that the Hessian matrix $H_{\mathcal{D}_{K,\mu_\infty}}$ of $\mathcal{D}_{K,\mu_\infty}$ valued at x has an eigenvalue 0.*

Remark 3.3.1. If μ_∞ satisfies the condtions in Lemma 3.3.3 and if $\mathrm{card}\big(G_K(\mu_\infty)\big) < +\infty$, a sufficient condition of Condition (b) in Theorem 3.2.2 is that $H_{\mathcal{D}_{K,\mu_\infty}}$ is positive definite in every $x \in G_K(\mu_\infty)$. In this case, one can take $\lambda^* = \min_{x\in G_K(\mu_\infty)} \underline{\lambda}_{H_{\mathcal{D}_{K,\mu_\infty}}(x)} - \varepsilon$ for a $\varepsilon > 0$, where $\underline{\lambda}_A$ denotes the smallest eigenvalue of a matrix A.

Proof of Lemma 3.3.3. We denote by $H_{\mathcal{D}_\infty}$ instead of $H_{\mathcal{D}_{K,\mu_\infty}}$ to simplify the notation. Proposition 3.1.1 implies that $G_K(\mu_\infty)$ is a compact set. If $\mathrm{card}\big(G_K(\mu_\infty)\big) = +\infty$, there exists $x, x^{(k)} \in G_K(\mu_\infty), k \in \mathbb{N}^*$ such that $x^{(k)} \to x$ when $k \to +\infty$. Set $u_k := \frac{x^{(k)}-x}{|x^{(k)}-x|}$, $k \geq 1$, then we have $|u_k| = 1$ for all $k \in \mathbb{N}^*$. Hence, there exists a subsequence $\varphi(k)$ of k such that $u_{\varphi(k)}$ converges to some $\widetilde{u}$ with $|\widetilde{u}| = 1$.

The Taylor expansion of $\mathcal{D}_{K,\mu_\infty}$ at x reads:

$$\mathcal{D}_{K,\mu_\infty}(x^{\varphi(k)}) = \mathcal{D}_{K,\mu_\infty}(x) + \big\langle \nabla\mathcal{D}_{K,\mu_\infty}(x) \;\big|\; x^{\varphi(k)} - x\big\rangle + \frac{1}{2} H_{\mathcal{D}_\infty}(\zeta^{\varphi(k)})(x^{\varphi(k)} - x)^{\otimes 2},$$

where $\zeta^{\varphi(k)}$ lies in the geometric segment $(x^{\varphi(k)}, x)$. Since $x, x^{(k)}, k \in \mathbb{N}^* \in G_K(\mu_\infty)$, then $\nabla\mathcal{D}_{K,\mu_\infty}(x) = 0$ and for any $k \in \mathbb{N}^*$, $\mathcal{D}_{K,\mu_\infty}(x^{\varphi(k)}) = \mathcal{D}_{K,\mu_\infty}(x)$. Hence, for any $k \in \mathbb{N}^*$, $H_{\mathcal{D}_\infty}(\zeta^{\varphi(k)})(x^{\varphi(k)} - x)^{\otimes 2} = 0$. Consequently, for any $k \in \mathbb{N}^*$,

$$H_{\mathcal{D}_\infty}(\zeta^{\varphi(k)})\Big(\frac{x^{\varphi(k)} - x}{|x^{\varphi(k)} - x|}\Big)^{\otimes 2} = 0.$$

Thus we have $H_{\mathcal{D}_\infty}(x)\widetilde{u}^{\otimes 2} = 0$ by letting $k \to +\infty$, which implies that $H_{\mathcal{D}_\infty}(x)$ has an eigenvalue 0. □

3.3.2 A criterion for positive definiteness of $H_{\mathcal{D}_\infty}(x^*)$ in 1-dimension

Let X denote a real random variable with distribution μ satisfying $\mu \in \mathcal{P}_2(\mathbb{R})$. Assume that μ is absolutely continuous with the respect of the Lebesgue measure with a continuous density function f, written by $\mu(d\xi) = f(\xi)d\xi$. In the one-dimensional case, it is necessary to point out a sufficient condition for the uniqueness of optimal quantizer. A probability distribution μ is called *strongly unimodal* if its density function f satisfies

that $I = \{f > 0\}$ is an open (possibly unbounded) interval and $\log f$ is concave on I. Moreover, we also have the uniqueness of optimal quantizer for such distributions.

Lemma 3.3.4. *For $K \in \mathbb{N}^*$, if μ is strongly unimodal with $\mathrm{card}\big(\mathrm{supp}(\mu)\big) \geq K$, then there are only one stationary (then optimal) quantizer of level K.*

We refer to Kieffer (1983), Trushkin (1982) and Graf and Luschgy (2000)[see Theorem 5.1] for the proof of Lemma 3.3.4 and for more details.

Let $F_K^+ := \big\{x = (x_1, ..., x_K) \in \mathbb{R}^K \mid -\infty < x_1 < x_2 < ... < x_K < +\infty\big\}$. Given an K-tuple $x = (x_1, ..., x_K) \in F_K^+$, the Voronoi region $V_i(x)$ can be explicitly written: $V_1(x) = (-\infty, \frac{x_1+x_2}{2}]$, $V_K(x) = [\frac{x_{K-1}+x_K}{2}, +\infty)$ and $V_i(x) = [\frac{x_{i-1}+x_i}{2}, \frac{x_i+x_{i+1}}{2}]$ for $i = 2, ..., K-1$. For all $x \in F_K^+$, $\mathcal{D}_{K,\mu}$ is differentiable at x and from (3.3.1), and

$$\nabla \mathcal{D}_{K,\mu}(x) = \left[\int_{V_i(x)} 2(x_i - \xi) f(\xi) d\xi\right]_{i=1,...,K}. \tag{3.3.8}$$

Therefore, one can solve the optimal quantizer $x^* \in F_K^+$ from $\nabla \mathcal{D}_{K,\mu}(x^*) = 0$,

$$x_i^* = \frac{\int_{V_i(x^*)} \xi f(\xi) d\xi}{\mu\big(V_i(x^*)\big)}, \quad \text{for } i = 1, ..., K. \tag{3.3.9}$$

For any $x \in F_K^+$, the Hessian matrix $H_{\mathcal{D}_{K,\mu}}$ of $\mathcal{D}_{K,\mu}$ at x is a tridiagonal symmetry matrix and can be calculated as follows,

$$H_{\mathcal{D}_{K,\mu}}(x) = \begin{pmatrix} A_1 - B_{1,2} & -B_{1,2} & & & & \\ & \ddots & & & & \\ & -B_{i-1,i} & A_i - B_{i-1,i} - B_{i,i+1} & -B_{i,i+1} & & \\ & & & \ddots & & \\ & & & & -B_{K-1,K} & A_K - B_{K-1,K} \end{pmatrix}, \tag{3.3.10}$$

where $A_i = 2\mu\big(C_i(x)\big)$ for $1 \leq i \leq K$ and $B_{i,j} = \frac{1}{2}(x_j - x_i) f(\frac{x_i+x_j}{2})$ for $1 \leq i < j \leq K$. Let F_μ be the cumulative distribution function of μ, then

$$\begin{aligned} A_1 &= 2\mu\big(C_1(x)\big) = 2F_\mu\Big(\frac{x_1 + x_2}{2}\Big), \\ A_i &= 2\mu\big(C_i(x)\big) = 2\Big[F_\mu\Big(\frac{x_{i+1} + x_i}{2}\Big) - F_\mu\Big(\frac{x_{i-1} + x_i}{2}\Big)\Big], \quad \text{for } i = 2, ..., K-1, \\ A_K &= 2\mu\big(C_K(x)\big) = 2\Big[1 - F_\mu\Big(\frac{x_{K-1} + x_K}{2}\Big)\Big]. \end{aligned}$$

Then the continuity of each term in the matrix $H_{\mathcal{D}_{K,\mu}}(x)$ can be directly obtained by the continuity of F_μ and f.

For $1 \leq i \leq K$, we define $L_i(x) \coloneqq \sum_{j=1}^{K} \frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_i \partial x_j}(x)$. The following proposition gives sufficient conditions to obtain the positive definiteness of $H_{\mathcal{D}_{K,\mu}}(x^*)$.

Proposition 3.3.1. *Any of the following two conditions implies the positive definiteness of* $H_{\mathcal{D}_{K,\mu}}(x^*)$,

(i) μ *is the uniform distribution,*

(ii) f *is differentiable and* $\log f$ *is strictly concave.*

In particular, (ii) also implies that $L_i(x^*) > 0$, $i = 1, ..., K$.

Remark that, under the conditions of Proposition 3.3.1, μ is strongly unimodal so that, if $x^* = (x_1^*, ..., x_K^*) \in F_K^+ \cap \operatorname{argmin} \mathcal{D}_{K,\mu}$, then $\Gamma^* = \{x_1, ..., x_K\}$ is the unique optimal quantizer for μ at level K (viewed as a set). Proposition 3.3.1 is proved in Appendix D. The conditions in Proposition 3.3.1 directly imply the convergence rate results.

Theorem 3.3.1. *Let* μ_n, $\mu_\infty \in \mathcal{P}_2(\mathbb{R})$ *such that* $\mathcal{W}_2(\mu_n, \mu_\infty) \to 0$. *Let* $x^{(n)}$ *be the optimal quantizer of* μ_n *which converges to* $x^{(\infty)}$. *Suppose* μ_∞ *is absolutely continuous with the respect of Lebesgue measure, written* $\mu_\infty(d\xi) = f(\xi)d\xi$. *Any one of the following conditions implies the existence of a constant* C_{μ_∞} *only depending on* μ_∞ *such that* $\left|x^{(n)} - x^{(\infty)}\right|^2 \leq C_{\mu_\infty} \cdot \mathcal{W}_2(\mu_n, \mu_\infty)$.

(i) μ_∞ *is the uniform distribution,*

(ii) f *is differentiable and* $\log f$ *is strictly concave.*

Proof. Let $\mathcal{D}_{K,\mu_\infty}$ denote the distortion function of μ_∞ and let $H_{\mathcal{D}_\infty}$ denote the Hessian matrix of $\mathcal{D}_{K,\mu_\infty}$.

(i) Let $f_k(x)$ be the k-th leading principal minor of $H_{\mathcal{D}_\infty}(x)$ defined in (3.3.10), then $f_k(x), k = 1, ..., K$, are continuous functions in x since every element in this matrix is continuous. Proposition 3.3.1 implies $f_k(x^{(\infty)}) > 0$, thus there exists $r > 0$ such that for every $x \in B(x^{(\infty)}, r)$, $f_k(x^{(\infty)}) > 0$ so that $H_{\mathcal{D}_\infty}(x)$ is positive definite. What remains can be directly proved by Corollary 3.2.1.

(ii) $L_i(x) \coloneqq \sum_{j=1}^{K} \frac{\partial^2 \mathcal{D}_{K,\mu_\infty}}{\partial x_i \partial x_j}(x)$ is continuous on x and Proposition 3.3.1 implies that $L_i(x^{(\infty)}) > 0$. Hence, there exists $r > 0$ such that $\forall x \in B(x^{(\infty)}, r)$, $L_i(x) > 0$. From

(3.3.10), one can remark that the i-th diagonal elements in $H_{\mathcal{D}_\infty}(x)$ is always larger than $L_i(x)$ for any $x \in \mathbb{R}^K$, then after Gershgorin Circle theorem, we have $H_{\mathcal{D}_\infty}(x)$ is positive definite for every $x \in B(x^{(\infty)}, r)$. What remains can be directly proved by Corollary 3.2.1. □

3.4 Empirical measure case

Let $\mu \in \mathcal{P}_{2+\varepsilon}(\mathbb{R}^d)$ for some $\varepsilon > 0$ and $\mathrm{card}\big(\mathrm{supp}(\mu)\big) \geq K$. Let X be a random variable with distribution μ and let $(X_n)_{n\geq 1}$ be a sequence of independent identically distributed $\mathbb{R}^d$-valued random variables with probability distribution μ. The empirical measure is defined for every $n \in \mathbb{N}^*$ by

$$\mu_n^\omega := \frac{1}{n}\sum_{i=1}^n \delta_{X_i(\omega)}, \quad \omega \in \Omega, \tag{3.4.1}$$

where δ_a is the Dirac measure on a. Let $K \in \mathbb{N}^*$ be the quantization level . For $n \geq 1$, let $x^{(n),\omega}$ be an optimal quantizer of μ_n^ω. The superscript ω is to emphasize that both μ_n^ω and $x^{(n),\omega}$ are random and we will drop ω when there is no ambiguity. We will cite two results of the convergence of $\mathcal{W}_2(\mu_n^\omega, \mu)$ among so many researches in this topic: the a.s. convergence in Pollard (1982b)[see Theorem 7] studied by D. Pollard, and the L^p-convergence rate of $\mathcal{W}_p(\mu_n^\omega, \mu)$ studied in Fournier and Guillin (2015).

Theorem 3.4.1. *(Fournier and Guillin (2015)[see Theorem 1]) Let $p > 0$ and let $\mu \in \mathcal{P}_q(\mathbb{R}^d)$ for some $q > p$. Let μ_n^ω denote the empirical measure of μ defined in (3.4.1). There exists a constant C only depending on p, d, q such that, for all $n \geq 1$,*

$$\mathbb{E}\Big(\mathcal{W}_p^p(\mu_n^\omega, \mu)\Big) \leq C M_q^{p/q}(\mu) \times \begin{cases} n^{-1/2} + n^{-(q-p)/q} & \text{if } p > d/2 \text{ and } q \neq 2p \\ n^{-1/2}\log(1+n) + n^{-(q-p)/q} & \text{if } p = d/2 \text{ and } q \neq 2p \\ n^{-p/d} + n^{-(q-p)/q} & \text{if } p \in (0, d/2) \text{ and } q \neq d/(d-p) \end{cases}, \tag{3.4.2}$$

where $M_q(\mu) = \int_{\mathbb{R}^d} |\xi|^q \, \mu(d\xi)$.

As the empirical measure μ_n^ω is usually used as an estimator of μ, a natural estimator of the optimal quantizer of μ is $x^{(n),\omega}$, the optimal quantizer for μ_n^ω. Let $\mathcal{D}_{K,\mu}$ denote the distortion function of μ and let $\mathcal{D}_{K,\mu_n}$ denote the distortion fuction of μ_n^ω for any $n \in \mathbb{N}^*$. Recall by Definition 3.1.1 that for $c = (c_1, ..., c_K) \in (\mathbb{R}^d)^K$,

$$\mathcal{D}_{K,\mu}(c) = \mathbb{E}\min_{1\leq k\leq K} |X - c_k|^2 = \mathbb{E}\Big[\,|X|^2 + \min_{1\leq k\leq K}\big(-2\langle X|c_k\rangle + |c_k|^2\big)\Big],$$

$$\text{and } \mathcal{D}_{K,\mu_n}(c) = \frac{1}{n}\sum_{i=1}^n \min_{1\leq k\leq K} |X_i - c_k|^2 = \frac{1}{n}\sum_{i=1}^n |X_i|^2 + \min_{1\leq k\leq K}\left(-\frac{2}{n}\sum_{i=1}^n \langle X_i|c_k\rangle + |c_k|^2\right).$$

The a.s. convergence of optimal quantizers for the empirical measure has been proved in Pollard (1981). We have the following convergence rate result for the clustering performance by applying directly Theorem 3.2.1 and (3.4.2).

Proposition 3.4.1. *Let $\mu \in \mathcal{P}_q(\mathbb{R}^d)$ for some $q > 2$ with card(supp(μ)) $\geq K$ and let μ_n^ω be the empirical measure of μ defined in (3.4.1). Fix a quantization level $K \in \mathbb{N}^*$. Let $x^{(n),\omega}$ be an optimal quantizer at level K of μ_n^ω. Then for any $n > K$,*

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu}(x) \leq C_{d,q,\mu,K} \times \begin{cases} n^{-1/4} + n^{-(q-2)/2q} & \text{if } d < 4 \text{ and } q \neq 4 \\ n^{-1/4}\big(\log(1+n)\big)^{1/2} + n^{-(q-2)/2q} & \text{if } d = 4 \text{ and } q \neq 4 \\ n^{-1/d} + n^{-(q-2)/2q} & \text{if } d > 4 \text{ and } q \neq d/(d-2) \end{cases}, \tag{3.4.3}$$

where $C_{d,q,\mu,K}$ is a constant depending on d, q, μ and the quantization level K.

The reason why we only consider $n > K$ is that for a fixed $n \in \mathbb{N}^*$, the empirical measure μ_n defined in (3.4.1) is supported by n points, which implies that if $n \leq K$, the optimal quantizer of μ_n at level K, viewed as a set, is in fact supp(μ_n). This makes the above bound of no interest. Following the remark after Theorem 1 in Fournier and Guillin (2015), one can remark that if the probability distribution μ has sufficiently many moments (namely if $q > 4$ when $d \leq 4$ and $q > 2d/(d-2)$ when $d > 4$), then the term $n^{-(q-2)/2q}$ is small and can be removed.

Proof of Proposition 3.4.1. For every $\omega \in \Omega$ and for every $n > K$, Theorem 3.2.1 implies that

$$\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu}(x) \leq 4e^*_{K,\mu}\mathcal{W}_2(\mu_n^\omega, \mu) + 4\mathcal{W}_2^2(\mu_n^\omega, \mu).$$

Thus,

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K} \mathcal{D}_{K,\mu}(x) \leq 4e^*_{K,\mu}\mathbb{E}\,\mathcal{W}_2(\mu_n^\omega, \mu) + 4\,\mathbb{E}\,\mathcal{W}_2^2(\mu_n^\omega, \mu).$$

It follows from (3.4.2) that

$$\mathbb{E}\,\mathcal{W}_2^2(\mu_n^\omega, \mu) \leq C_{d,q,\mu} \times \begin{cases} n^{-1/2} + n^{-(q-2)/q} & \text{if } d < 4 \text{ and } q \neq 4 \\ n^{-1/2}\log(1+n) + n^{-(q-2)/q} & \text{if } d = 4 \text{ and } q \neq 4 \\ n^{-2/d} + n^{-(q-2)/q} & \text{if } d > 4 \text{ and } q \neq d/(d-2) \end{cases}, \tag{3.4.4}$$

where $C_{d,q,\mu} = C \cdot M_q^{2/q}(\mu)$ and C is the constant in 3.4.2. Moreover, as $\mathbb{E}\,\mathcal{W}_2(\mu_n^\omega, \mu) \leq \big(\mathbb{E}\mathcal{W}_2^2(\mu_n^\omega, \mu)\big)^{1/2}$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \in \mathbb{R}_+$, the inequality (3.4.2) also implies

$$\mathbb{E}\,\mathcal{W}_2(\mu_n^\omega, \mu) \leq C_{d,q,\mu}^{1/2} \times \begin{cases} n^{-1/4} + n^{-(q-2)/2q} & \text{if } d < 4 \text{ and } q \neq 4 \\ n^{-1/4}\big(\log(1+n)\big)^{1/2} + n^{-(q-2)/2q} & \text{if } d = 4 \text{ and } q \neq 4 \\ n^{-1/d} + n^{-(q-2)/2q} & \text{if } d > 4 \text{ and } q \neq d/(d-2) \end{cases}.$$

Consequently,

$$\begin{aligned} \mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x \in (\mathbb{R}^d)^K} \mathcal{D}_{K,\mu}(x) &\leq 4e_{K,\mu}^* \mathbb{E}\,\mathcal{W}_2(\mu_n^\omega, \mu) + 4\,\mathbb{E}\,\mathcal{W}_2^2(\mu_n^\omega, \mu). \\ &\leq 8(C_{d,q,\mu}^{1/2} e_{K,\mu}^* \vee C_{d,q,\mu}) \times \\ &\quad \begin{cases} n^{-1/4} + n^{-(q-2)/2q} & \text{if } d < 4 \text{ and } q \neq 4 \\ n^{-1/4}\big(\log(1+n)\big)^{1/2} + n^{-(q-2)/2q} & \text{if } d = 4 \text{ and } q \neq 4 \\ n^{-1/d} + n^{-(q-2)/2q} & \text{if } d > 4 \text{ and } q \neq d/(d-2) \end{cases}. \end{aligned} \tag{3.4.5}$$

One can conclude by letting $C_{d,q,\mu,K} \coloneqq 8(C_{d,q,\mu}^{1/2} e_{K,\mu}^* \vee C_{d,q,\mu})$. □

Remark 3.4.1. When $d \geq 4$, if $\frac{q-2}{q} > \frac{2}{d}$ i.e. $q > \frac{2d}{d-2}$, the inequality (3.4.4) can be upper bounded as follows,

$$\begin{aligned} \mathbb{E}\,\mathcal{W}_2^2(\mu_n^\omega, \mu) &\leq 2C_{d,q,\mu} n^{-1/d} \times \begin{cases} n^{-\frac{1}{4}} \log(1+n) & \text{if } d = 4 \text{ and } q \neq 4 \\ n^{-\frac{1}{d}} & \text{if } d > 4 \text{ and } q \neq d/(d-2) \end{cases} \\ &\leq 2C_{d,q,\mu} K^{-1/d} \times \begin{cases} n^{-\frac{1}{4}} \log(1+n) & \text{if } d = 4 \text{ and } q \neq 4 \\ n^{-\frac{1}{d}} & \text{if } d > 4 \text{ and } q \neq d/(d-2) \end{cases}, \end{aligned}$$

since we consider only $n \geq K$ and if $q > \frac{2d}{d-2}$, the term $n^{-(q-2)/2q}$ is smaller than the first term. Consequently, (3.4.5) can be bounded by

$$\begin{aligned} \mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x \in (\mathbb{R}^d)^K} \mathcal{D}_{K,\mu}(x) &\leq 4e_{K,\mu}^* \mathbb{E}\,\mathcal{W}_2(\mu_n^\omega, \mu) + 4\,\mathbb{E}\,\mathcal{W}_2^2(\mu_n^\omega, \mu). \\ &\leq 8(C_{d,q,\mu}^{1/2} e_{K,\mu}^* \vee 2C_{d,q,\mu} K^{-1/d}) \times \\ &\quad \begin{cases} n^{-\frac{1}{4}}\big[(\log(1+n))^{\frac{1}{2}} + \log(1+n)\big] & \text{if } d = 4 \text{ and } q \neq 4 \\ 2n^{-\frac{1}{d}} & \text{if } d > 4 \text{ and } q \neq d/(d-2) \end{cases}. \end{aligned} \tag{3.4.6}$$

By the non-asymptotic Zador theorem (3.1.10), one has

$$e_{K,\mu}^* \leq C_{d,q}(\mu)\sigma_q(\mu)K^{-1/d}$$

with $\sigma_q(\mu) = \min_{a\in\mathbb{R}^d}\left[\int_{\mathbb{R}^d}|\xi-a|^q\,\mu(d\xi)\right]^{1/q}$. Thus, the inequality (3.4.6) can be upper-bounded as follows,

$$\begin{aligned}\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x) &\le 4e^*_{K,\mu}\mathbb{E}\,\mathcal{W}_2(\mu_n^\omega,\mu) + 4\,\mathbb{E}\,\mathcal{W}_2^2(\mu_n^\omega,\mu).\\ &\le 8K^{-1/d}\big(C^{1/2}_{d,q,\mu}C_{d,q}(\mu)\sigma_q(\mu)\vee 2C_{d,q,\mu}\big)\times\\ &\begin{cases} n^{-\frac14}\left[(\log(1+n))^{\frac12}+\log(1+n)\right] & \text{if } d=4 \text{ and } q\neq 4\\ 2n^{-\frac1d} & \text{if } d>4 \text{ and } q\neq d/(d-2)\end{cases},\end{aligned}$$

from which one can remark that the right side of this inequality is strictly decreasing with respect to K.

Theorem 3.4.2. *Let $K\in\mathbb{N}^*$ be the quantization level. Let $\mu\in\mathcal{P}_2(\mathbb{R}^d)$ with card(supp(μ)) $\ge K$ and let μ_n^ω be the empirical measure of μ defined in (3.4.1), generated by i.i.d observation $X_1,\dots,X_n$. We denote by $x^{(n),\omega}\in(\mathbb{R}^d)^K$ an optimal quantizer of μ_n^ω at level K. Then,*

(a) General upper bound of the performance.

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x) \le \frac{2K}{\sqrt{n}}\Big[r_{2n}^2+\rho_K(\mu)^2+2r_1\big(r_{2n}+\rho_K(\mu)\big)\Big], \quad (3.4.7)$$

where $r_n := \big\|\max_{1\le i\le n}|X_i|\,\big\|_2$ and $\rho_K(\mu)$ is the maximum radius of optimal quantizers of μ, defined in (3.1.8).

(b) Asymptotic upper bound for measure with polynomial tail. *For $p>2$, if μ has a c-th polynomial tail with $c>d+p$, then*

$$\mathbb{E}\,\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x) \le \frac{K}{\sqrt{n}}\Big[C_{\mu,p}\,n^{2/p}+6K^{\frac{2(p+d)}{d(c-p-d)}}\gamma_K\Big],$$

where $C_{\mu,p}$ is a constant depending μ,p and $\lim_K\gamma_K=1$.

(c) Asymptotic upper bound for measure with hyper-exponential tail. *Recall that μ has a hyper-exponential tail if $\mu=f\cdot\lambda_d$ and there exists $\tau>0,\kappa,\vartheta>0,c>-d$ and $A>0$ such that $\forall\xi\in\mathbb{R}^d,|\xi|\ge A\Rightarrow f(\xi)=\tau\,|\xi|^c\,e^{-\vartheta|\xi|^\kappa}$. If $\kappa\ge 2$, we can obtain a more precise upper bound of the performance*

$$\mathbb{E}\big[\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x)\big] \le C_{\vartheta,\kappa,\mu}\cdot\frac{K}{\sqrt{n}}\Big[1+(\log n)^{2/\kappa}+\gamma_K(\log K)^{2/\kappa}\big(1+\frac{2}{d}\big)^{2/\kappa}\Big],$$

where $C_{\vartheta,\kappa,\mu}$ is a constant depending ϑ,κ,μ and $\limsup_K\gamma_K=1$.
In particular, if $\mu=\mathcal{N}(m,\Sigma)$, the multidimensional normal distribution, we have

$$\mathbb{E}\big[\mathcal{D}_{K,\mu}(x^{(n),\omega}) - \inf_{x\in(\mathbb{R}^d)^K}\mathcal{D}_{K,\mu}(x)\big] \le C_\mu\cdot\frac{K}{\sqrt{n}}\Big[1+\log n+\gamma_K\log K\big(1+\frac{2}{d}\big)\Big],$$

where $\limsup_K \gamma_K = 1$ *and* $C_\mu = 24 \cdot \left(1 \vee \log 2\mathbb{E}e^{|X|^2/4}\right)$ *where* X *is a random variable with distribution* μ. *Moreover, when* $\mu = \mathcal{N}(0, \mathbf{I}_d)$, $C_\mu = 24(1 + \frac{d}{2}) \cdot \log 2$.

The proof of Theorem 3.4.2 relies on the Rademacher process theory. A Rademacher sequence $(\sigma_i)_{i\in\{1,\dots,n\}}$ is a sequence of i.i.d random variables with a symmetric $\{\pm 1\}$-valued Bernoulli distribution, independent to $(X_1, \dots, X_n)$ and we define the Rademacher process $\mathcal{R}_n(f), f \in \mathcal{F}$ by $\mathcal{R}_n(f) := \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)$. Remark that the Rademacher process $\mathcal{R}_n(f)$ depends on the sample $\{X_1, \dots, X_n\}$ of probability measure μ.

Theorem 3.4.3 (Symmetrization inequalites)**.** *For any class* $\mathcal{F}$ *of* $\mathbb{P}$*-integrable functions, we have*

$$\mathbb{E}\left\|\mu_n - \mu\right\|_{\mathcal{F}} \leq 2\mathbb{E}\left\|\mathcal{R}_n\right\|_{\mathcal{F}},$$

where for a probability distribution ν, $\|\nu\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\nu(f)| := \sup_{f\in\mathcal{F}} |\int_{\mathbb{R}^d} f d\nu|$ *and* $\|\mathcal{R}_n\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\mathcal{R}_n(f)|$.

For the proof of the above theorem, we refer to Koltchinskii (2011)[see Theorem 2.1]. Another more detailed reference is Van der Vaart and Wellner (1996)[see Lemma 2.3.1]. We will also introduce the *Contraction principle* in the following theorem and we refer to Boucheron et al. (2013)[see Theorem 11.6] for the proof.

Theorem 3.4.4 (Contraction principle)**.** *Let* $x_1, \dots, x_n$ *be vectors whose real-valued components are indexed by* $\mathcal{T}$, *that is,* $x_i = (x_{i,s})_{s\in\mathcal{T}}$. *For each* $i = 1, \dots, n$ *let* $\varphi_i : \mathbb{R} \to \mathbb{R}$ *be a Lipschitz function such that* $\varphi_i(0) = 0$. *Let* $\sigma_1, \dots, \sigma_n$ *be independent Rademacher random variables and let* $c_L = \max_{1\leq i\leq n} \sup_{\substack{x,y\in\mathbb{R}\\ x\neq y}} \left|\frac{\varphi_i(x)-\varphi_i(y)}{x-y}\right|$ *be the Lipschitz constant. Then*

$$\mathbb{E}\Big[\sup_{s\in\mathcal{T}} \sum_{i=1}^n \sigma_i\varphi_i(x_{i,s})\Big] \leq c_L \cdot \mathbb{E}\Big[\sup_{s\in\mathcal{T}} \sum_{i=1}^n \sigma_i x_{i,s}\Big]. \tag{3.4.8}$$

Remark that if we consider random variables $(Y_{1,s}, \dots, Y_{n,s})_{s\in\mathcal{T}}$ independent of $(\sigma_1, \dots, \sigma_n)$ and for all $s \in \mathcal{T}$ and $i \in \{1, \dots, n\}$, $Y_{i,s}$ is valued in $\mathbb{R}$, then (3.4.8) implies that

$$\begin{aligned}\mathbb{E}\Big[\sup_{s\in\mathcal{T}} \sum_{i=1}^n \sigma_i\varphi_i(Y_{i,s})\Big] &= \mathbb{E}\Big\{\mathbb{E}\Big[\sup_{s\in\mathcal{T}} \sum_{i=1}^n \sigma_i\varphi_i(Y_{i,s}) \mid (Y_{1,s}, \dots, Y_{n,s})_{s\in\mathcal{T}}\Big]\Big\}\\ &\leq c_L \cdot \mathbb{E}\Big\{\mathbb{E}\Big[\sup_{s\in\mathcal{T}} \sum_{i=1}^n \sigma_i Y_{i,s} \mid (Y_{1,s}, \dots, Y_{n,s})_{s\in\mathcal{T}}\Big]\Big\} \leq c_L \cdot \mathbb{E}\Big[\sup_{s\in\mathcal{T}} \sum_{i=1}^n \sigma_i Y_{i,s}\Big]. \end{aligned} \tag{3.4.9}$$

The proof of Theorem 3.4.2 is principally inspired by the proof of Theorem 2.1 in Biau et al. (2008).

Proof of Theorem 3.4.2. (a) In order to simplify the notation, we will denote by $\mathcal{D}$ (*respectively* $\mathcal{D}_n$) instead of $\mathcal{D}_{K,\mu}$ (*resp.* $\mathcal{D}_{K,\mu_n}$) as the distortion function of μ (*resp.*

μ_n). For any $c = (c_1, ..., c_K) \in (\mathbb{R}^d)^K$, recall the distortion function $\mathcal{D}(c)$ of μ can be written as

$$\mathcal{D}(c) = \mathbb{E}\big[\min_{1\le k\le K} |X - c_k|^2\big] = \mathbb{E}\big[|X|^2 + \min_{1\le k\le K}(-2\langle X|c_k\rangle + |c_k|^2)\big].$$

We define $\overline{\mathcal{D}}(c) := \min_{1\le k\le K}\big(-2\langle X|c_k\rangle + |c_k|^2\big)$. Similarly, for the distortion function $\mathcal{D}_n$ of the empirical measure μ_n,

$$\mathcal{D}_n(c) = \frac{1}{n}\sum_{i=1}^n \min_{1\le k\le K} |X_i - c_k|^2 = \frac{1}{n}\sum_{i=1}^n |X_i|^2 + \min_{1\le k\le K}\big(-\frac{2}{n}\sum_{i=1}^n \langle X_i|c_k\rangle + |c_k|^2\big),$$

we define $\overline{\mathcal{D}}_n(c) := \min_{1\le k\le K}\big(-\frac{2}{n}\sum_{i=1}^n\langle X_i|c_k\rangle + |c_k|^2\big)$. We will drop ω in $x^{(n),\omega}$ to alleviate the notation throughout the proof. Let $x \in G_K(\mu)$. It follows that

$$\begin{aligned}\mathbb{E}\big[\mathcal{D}(x^{(n)}) - \mathcal{D}(x)\big] &= \mathbb{E}\big[\overline{\mathcal{D}}(x^{(n)}) - \overline{\mathcal{D}}(x)\big] = \mathbb{E}\big[\overline{\mathcal{D}}(x^{(n)}) - \overline{\mathcal{D}}_n(x^{(n)})\big] + \mathbb{E}\big[\overline{\mathcal{D}}_n(x^{(n)}) - \overline{\mathcal{D}}(x)\big]\\ &\le \mathbb{E}\big[\overline{\mathcal{D}}(x^{(n)}) - \overline{\mathcal{D}}_n(x^{(n)})\big] + \mathbb{E}\big[\overline{\mathcal{D}}_n(x) - \overline{\mathcal{D}}(x)\big]. \qquad (3.4.10)\end{aligned}$$

Define for $\eta, x \in \mathbb{R}^d$, $f_\eta(x) = -2\langle\eta|x\rangle + |\eta|^2$.

Part (i): Upper bound of $\mathbb{E}[\overline{\mathcal{D}}(x^{(n)}) - \overline{\mathcal{D}}_n(x^{(n)})]$. Let $R_n(\omega) := \max_{1\le i\le n}|X_i(\omega)|$. Remark that for every $\omega \in \Omega$, $R_n(\omega)$ is invariant with the respect to all permutation of the components of $(X_1, ..., X_n)$. Let B_R denote the ball centred at 0 with radius R. Then owing to Proposition 3.1.1-(iii), $x^{(n)} \in B^K_{R_n}$. Hence,

$$\begin{aligned}\mathbb{E}\,[\overline{\mathcal{D}}(x^{(n)}) - \overline{\mathcal{D}}_n(x^{(n)})] &\le \mathbb{E}\sup_{c\in B^K_{R_n}}\big(\overline{\mathcal{D}}(c) - \overline{\mathcal{D}}_n(c)\big)\\ &= \mathbb{E}\Big[\sup_{c\in B^K_{R_n}}\big(\mathbb{E}\min_{1\le k\le K} f_{c_k}(X) - \frac{1}{n}\sum_{i=1}^n \min_{1\le k\le K} f_{c_k}(X_i)\big)\Big]\\ &= \mathbb{E}\Big[\sup_{c\in B^K_{R_n}}\mathbb{E}\big[\frac{1}{n}\sum_{i=1}^n \min_{1\le k\le K} f_{c_k}(X'_i) - \frac{1}{n}\sum_{i=1}^n \min_{1\le k\le K} f_{c_k}(X_i)\big|X_1, ..., X_n\big]\Big],\end{aligned} \qquad (3.4.11)$$

where $X'_1, ..., X'_n$ are i.i.d random variable with the distribution μ, independent of $(X_1, ..., X_n)$. Let $R_{2n} := \max_{1\le i\le n}|X_i| \vee |X'_i|$, then (3.4.11) becomes

$$\begin{aligned}&\mathbb{E}\,[\overline{\mathcal{D}}(x^{(n)}) - \overline{\mathcal{D}}_n(x^{(n)})]\\ &\le \mathbb{E}\Big[\sup_{c\in B^K_{R_{2n}}}\mathbb{E}\big[\frac{1}{n}\sum_{i=1}^n \min_{1\le k\le K} f_{c_k}(X'_i) - \frac{1}{n}\sum_{i=1}^n \min_{1\le k\le K} f_{c_k}(X_i)\big|X_1, ..., X_n\big]\Big]\\ &\le \mathbb{E}\Big[\mathbb{E}\big[\sup_{c\in B^K_{R_{2n}}}\big(\frac{1}{n}\sum_{i=1}^n \min_{1\le k\le K} f_{c_k}(X'_i) - \frac{1}{n}\sum_{i=1}^n \min_{1\le k\le K} f_{c_k}(X_i)\big)\big|X_1, ..., X_n\big]\Big]\end{aligned}$$

$$=\mathbb{E}\Big[\sup_{c\in B^K_{R_{2n}}}\frac{1}{n}\sum_{i=1}^{n}\big(\min_{1\le k\le K} f_{c_k}(X'_i)-\min_{1\le k\le K} f_{c_k}(X_i)\big)\Big]. \tag{3.4.12}$$

The distribution of $(X_1,...,X_n,X'_1,...,X'_n)$ and that of R_{2n} are invariant with the respect to all permutation of the components in $(X_1,...,X_n,X'_1,...,X'_n)$. Hence,

$$\begin{aligned}\mathbb{E}\left[\overline{\mathcal{D}}(x^{(n)})-\overline{\mathcal{D}}_n(x^{(n)})\right] &= \mathbb{E}\Big[\sup_{c\in B^K_{R_{2n}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\big(\min_{1\le k\le K} f_{c_k}(X'_i)-\min_{1\le k\le K} f_{c_k}(X_i)\big)\Big]\\ &\le \mathbb{E}\Big[\sup_{c\in B^K_{R_{2n}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\min_{1\le k\le K} f_{c_k}(X'_i)\Big]+\mathbb{E}\Big[\sup_{c\in B^K_{R^K_{2n}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\min_{1\le k\le K} f_{c_k}(X_i)\Big]\\ &= 2\mathbb{E}\Big[\sup_{c\in B^K_{R_{2n}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\min_{1\le k\le K} f_{c_k}(X_i)\Big]. \end{aligned}\tag{3.4.13}$$

In the second line of (3.4.13), we can change the sign before the second term since $-\sigma_i$ has the same distribution of σ_i, and we will continue to use this property throughout the proof. Let $S_K=\mathbb{E}\Big[\sup_{c\in B^K_{R_{2n}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\min_{1\le k\le K} f_{c_k}(X_i)\Big]$.

▶ For $K=1$,

$$\begin{aligned}S_1 =&\mathbb{E}\Big[\sup_{c\in B_{R_{2n}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\min_{1\le k\le K} f_c(X_i)\Big]=\mathbb{E}\Big[\sup_{c\in B_{R_{2n}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\big(-2\langle c|X_i\rangle+|c|^2\big)\Big]\\ \le& 2\,\mathbb{E}\Big[\sup_{c\in B_{R_{2n}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\langle c|X_i\rangle\Big]+\mathbb{E}\Big[\sup_{c\in B_{R_{2n}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\,|c|^2\Big]\\ \le&\frac{2}{n}\mathbb{E}\Big[\sup_{c\in B_{R_{2n}}}\langle c|\sum_{i=1}^{n}\sigma_iX_i\rangle\Big]+\frac{1}{n}\mathbb{E}\Big[\Big|\sum_{i=1}^{n}\sigma_i\Big|\cdot|R_{2n}|^2\Big]\\ \le&\frac{2}{n}\mathbb{E}\Big[\sup_{c\in B_{R_{2n}}}\Big|\sum_{i=1}^{n}\sigma_iX_i\Big|\cdot|c|\Big]+\frac{1}{n}\mathbb{E}\Big|\sum_{i=1}^{n}\sigma_i\Big|\cdot\mathbb{E}\,|R_{2n}|^2\end{aligned}$$

(by Cauchy-Schwarz inequality and independence of σ_i and X_i)

$$\begin{aligned}\le&\frac{2}{n}\Big\|\sum_{i=1}^{n}\sigma_iX_i\Big\|_2\cdot\|R_{2n}\|_2+\frac{1}{n}\Big\|\sum_{i=1}^{n}\sigma_i\Big\|_2^2\cdot\|R_{2n}\|_2^2\\ \le&\frac{2}{n}\sqrt{n}\,\|X_1\|_2\cdot\|R_{2n}\|_2+\frac{1}{\sqrt{n}}\,\|R_{2n}\|_2^2\le\frac{\|R_{2n}\|_2}{\sqrt{n}}\big(2\,\|X_1\|_2+\|R_{2n}\|_2\big).\end{aligned}\tag{3.4.14}$$

The first inequality of the last line of (3.4.14) is due to $\mathbb{E}\,|\sum_{i=1}^{n}\sigma_iX_i|^2=\mathbb{E}\sum_{i=1}^{n}\sigma_i^2X_i^2=n\mathbb{E}X_1^2$ since the $(\sigma_1,...,\sigma_n)$ is independent of $(X_1,...,X_n)$ and $\mathbb{E}\,\sigma_i=0$. For $n\in\mathbb{N}^*$, define $r_n:=\|\max_{1\le i\le n}|Y_i|\|_2$, where $Y_1,...,Y_n$ are i.i.d random variable with probability distribution μ. Hence, $r_{2n}=\|R_{2n}\|_2$, since $(Y_1,...,Y_{2n})$ has the same distribution than

$(X_1, ..., X_n, X'_1, ..., X'_n)$. Therefore,

$$S_1 \leq \frac{r_{2n}}{\sqrt{n}}\big(2\,\|X_1\|_2 + r_{2n}\big).$$

► For $K = 2$,

$$\begin{aligned}
S_2 =& \mathbb{E}\Big[\sup_{c=(c_1,c_2)\in B^2_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(f_{c_1}(X_i) \wedge f_{c_2}(X_i)\big)\Big] \\
=& \frac{1}{2}\mathbb{E}\Big[\sup_{c\in B^2_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(f_{c_1}(X_i) + f_{c_2}(X_i) - |f_{c_1}(X_i) - f_{c_2}(X_i)|\,\big)\Big] \\
& \quad (\text{as } a\wedge b = \frac{a+b}{2} - \frac{|a-b|}{2}) \\
\leq& \frac{1}{2}\Big\{\mathbb{E}\Big[\sup_{c\in B^2_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(f_{c_1}(X_i) + f_{c_2}(X_i)\big)\Big] + \mathbb{E}\Big[\sup_{c\in B^2_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i\,|f_{c_1}(X_i) - f_{c_2}(X_i)|\,\Big]\Big\} \\
\leq& \frac{1}{2}\Big\{2S_1 + \mathbb{E}\Big[\sup_{c\in B^2_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(f_{c_1}(X_i) - f_{c_2}(X_i)\big)\Big]\Big\} \quad \big(\text{by } (3.4.9)\big) \\
\leq& \frac{1}{2}\Big\{2S_1 + \mathbb{E}\Big[\sup_{c_1\in B_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i f_{c_1}(X_i)\Big] + \mathbb{E}\Big[\sup_{c_2\in B_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i f_{c_2}(X_i)\Big]\Big\} \leq 2S_1.
\end{aligned} \tag{3.4.15}$$

► Next, we will show by recurrence that $S_K \leq KS_1$ for every $K \in \mathbb{N}^*$. Assume that $S_K \leq KS_1$, for $K+1$,

$$\begin{aligned}
S_{K+1} &= \mathbb{E}\Big[\sup_{c\in B^{K+1}_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i \min_{1\leq k\leq K+1} f_{c_k}(X_i)\Big] \\
&= \mathbb{E}\Big[\sup_{c\in B^{K+1}_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(\min_{1\leq k\leq K} f_{c_k}(X_i) \wedge f_{c_{K+1}}(X_i)\big)\Big] \\
&\leq \frac{1}{2}\mathbb{E}\Big\{\sup_{c\in B^{K+1}_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i\Big[\big(\min_{1\leq k\leq K} f_{c_k}(X_i) + f_{c_{K+1}}(X_i)\big) - \Big|\min_{1\leq k\leq K} f_{c_k}(X_i) - f_{c_{K+1}}(X_i)\Big|\,\Big]\Big\} \\
&\leq \frac{1}{2}\mathbb{E}\Big\{\sup_{c\in B^{K+1}_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(\min_{1\leq k\leq K} f_{c_k}(X_i) + f_{c_{K+1}}(X_i)\big) \\
&\quad + \sup_{c\in B^{K+1}_{R_{2n}}} \frac{1}{n}\sum_{i=1}^n \sigma_i\Big|\min_{1\leq k\leq K} f_{c_k}(X_i) - f_{c_{K+1}}(X_i)\Big|\Big\} \\
&\leq \frac{1}{2}(S_K + S_1 + S_K + S_1) \leq S_K + S_1 \leq (K+1)S_1.
\end{aligned} \tag{3.4.16}$$

Hence,

$$\mathbb{E}\left[\overline{\mathcal{D}}(x^{(n)}) - \overline{\mathcal{D}}_n(x^{(n)})\right] \le 2S_K \le 2KS_1 \le \frac{2K \cdot r_{2n}}{\sqrt{n}}\big(2\,\|X_1\|_2 + r_{2n}\big). \tag{3.4.17}$$

Part (ii): Upper bound of $\mathbb{E}\left[\overline{\mathcal{D}}_n(x) - \overline{\mathcal{D}}(x)\right]$. As $x = (x_1, ..., x_K)$ is an optimal quantizer of μ, we have $\max_{1\le k\le K} |x_k| \le \rho_K(\mu)$ owing to the definition of $\rho_K(\mu)$ in (3.1.8). Consequently,

$$\mathbb{E}\big[\overline{\mathcal{D}}_n(x) - \overline{\mathcal{D}}(x)\big] \le \mathbb{E} \sup_{c\in B^K_{\rho_K(\mu)}} \big[\overline{\mathcal{D}}_n(c) - \overline{\mathcal{D}}(c)\big]$$

By the same reasoning of Part (I), we have

$$\mathbb{E}\big[\overline{\mathcal{D}}_n(x) - \overline{\mathcal{D}}(x)\big] \le \frac{2K}{\sqrt{n}}\rho_K(\mu)\big(2\,\|X_1\|_2 + \rho_K(\mu)\big).$$

Hence

$$\begin{aligned}\mathbb{E}\big[\mathcal{D}(x^{(n)}) - \mathcal{D}(x)\big] &\le \frac{2K}{\sqrt{n}} r_{2n}\big(2\,\|X_1\|_2 + r_{2n}\big) + \frac{2K}{\sqrt{n}}\rho_K(\mu)\big(2\,\|X_1\|_2 + \rho_K(\mu)\big)\\ &\le \frac{2K}{\sqrt{n}}\Big[r_{2n}^2 + \rho_K^2(\mu) + 2r_1\big(r_{2n} + \rho_K(\mu)\big)\Big].\end{aligned} \tag{3.4.18}$$

(b) If μ has a c-th polynomial tail with $c > d + p$, then $\mu \in \mathcal{P}_p(\mathbb{R}^d)$. Let $X, X_1, ..., X_n$ be i.i.d random variable with probability distribution μ. Then,

$$\begin{aligned}r_n = \|R_n\|_2^2 &= \mathbb{E}\big[\max\left(|X_1|, ..., |X_n|\right)^2\big] = \mathbb{E}\big[\max(|X_1|^p, ..., |X_n|^p)^{2/p}\big]\\ \le &\mathbb{E}\Big(\big[\sum_{i=1}^n |X_i|^p\big]^{2/p}\Big) \le \Big[\mathbb{E}\big(\sum_{i=1}^n |X_i|^p\big)\Big]^{2/p} = \Big[n\,\mathbb{E}\,|X|^p\Big]^{2/p} = n^{2/p}\,\|X\|_p^2,\end{aligned} \tag{3.4.19}$$

where the last line is due to the fact that $X_1, ..., X_n$ have the same distribution as X. Moreover, we have

$$\rho_K(\mu) = K^{\frac{p+d}{d(c-p-d)}\gamma_K} \quad \text{with} \quad \lim_{K\to+\infty} \gamma_K = 1 \tag{3.4.20}$$

owing to (3.1.11). It follows from (3.4.18) that

$$\mathbb{E}\big[\mathcal{D}(x^{(n)}) - \mathcal{D}(x)\big] \le \frac{2K}{\sqrt{n}}\Big[3r_{2n}^2 + \big((2m_2) \vee \rho_K(\mu)\big) \cdot \rho_K(\mu)\Big]$$

since $r_{2n} \ge m_2$ after the definitions of r_{2n} and m_2. In addition, (3.4.20) implies that $\rho_K(\mu) \to +\infty$ as $K \to +\infty$ and, for large enough K, $\rho_K(\mu) \ge 2m_2$. Therefore,

$$\mathbb{E}\big[\mathcal{D}(x^{(n)}) - \mathcal{D}(x)\big] \le \frac{2K}{\sqrt{n}}\Big(3 \cdot (2n)^{2/p}\,\|X\|_p^2 + 3K^{\frac{p+d}{d(c-p-d)}\gamma_K}\Big)$$

$$= \frac{K}{\sqrt{n}} \Big(C_{\mu,p}\, n^{2/p} + 6K^{\frac{p+d}{d(c-p-d)}\gamma_K} \Big),$$

where $C_{\mu,p} = 6 \cdot 2^{2/p} \|X\|_p^2$ and $\lim_K \gamma_K = 1$.

(c) μ is assumed to have a hyper-exponential tail, that is, $\mu = f \cdot \lambda_d$ and $f(\xi) = \tau\, |\xi|^c\, e^{-\vartheta|\xi|^\kappa}$ with $c > -d$ for $|\xi|$ large enough. The real constant κ is assumed to be greater than or equal to 2. Let X be a random variable with probability distribution μ. Therefore, for every $\lambda \in (0, \vartheta)$, $\mathbb{E}\, e^{\lambda|X|^\kappa} < +\infty$, and

$$\begin{aligned}
r_n &= \|R_n\|_2^2 = \mathbb{E}\big[\max(|X_1|, ..., |X_n|)^2\big] = \mathbb{E}\big[\max(|X_1|^\kappa, ..., |X_n|^\kappa)^{2/\kappa}\big] \\
&= \mathbb{E}\Big(\Big[\frac{1}{\lambda} \log\big(\max(e^{\lambda|X_1|^\kappa}, ..., e^{\lambda|X_n|^\kappa})\big)\Big]^{2/\kappa}\Big) \le \Big(\frac{1}{\lambda}\Big)^{2/\kappa} \Big[\log \mathbb{E} \max(e^{\lambda|X_1|^\kappa}, ..., e^{\lambda|X_n|^\kappa})\Big]^{2/\kappa} \\
&\le \Big(\frac{1}{\lambda}\Big)^{2/\kappa} \Big\{\log \mathbb{E}\Big[\sum_{i=1}^n e^{\lambda|X_i|^\kappa}\Big]\Big\}^{2/\kappa} = \Big(\frac{1}{\lambda}\Big)^{2/\kappa} \big\{\log(n\mathbb{E}\, e^{\lambda|X|^\kappa})\big\}^{2/\kappa} \\
&= \Big(\frac{1}{\lambda}\Big)^{2/\kappa} \big(\log \mathbb{E}\, e^{\lambda|X|^\kappa} + \log n\big)^{2/\kappa}, \qquad (3.4.21)
\end{aligned}$$

where the last line of (3.4.21) is due to the fact that $X_1, ..., X_n$ have the same distribution than X. Under the same assumption as before,

$$\rho_K(\mu) \le \gamma_K (\log K)^{1/\kappa} \cdot 2\vartheta^{-1/\kappa} \big(1 + \frac{2}{d}\big)^{1/\kappa} \quad \text{with} \quad \limsup_{K \to +\infty} \gamma_K \le 1 \qquad (3.4.22)$$

by applying (3.1.12). Moreover, it follows from (3.4.18) that

$$\mathbb{E}\big[\mathcal{D}(x^{(n)}) - \mathcal{D}(x)\big] \le \frac{2K}{\sqrt{n}} \Big[3r_{2n}^2 + \big((2m_2) \vee \rho_K(\mu)\big) \cdot \rho_K(\mu)\Big]$$

since $r_{2n} \ge m_2$ after the definitions of r_{2n} and m_2. In addition, (3.4.22) implies that $\rho_K(\mu) \to +\infty$ as $K \to +\infty$ and, for large enough K, $\rho_K(\mu) \ge 2m_2$. Therefore,

$$\begin{aligned}
\mathbb{E}\big[\mathcal{D}(x^{(n)}) - \mathcal{D}(x)\big] \le& \frac{2K}{\sqrt{n}} \Big\{3 \cdot \big(1 \vee \log 2\mathbb{E}\, e^{\lambda|X|^\kappa}\big)^{2/\kappa} \big(\frac{1}{\lambda}\big)^{2/\kappa} \big[(\log n)^{2/\kappa} + 1\big]\Big\} \\
&+ 4\vartheta^{-2/\kappa} \gamma_K (\log K)^{2/\kappa} \big(1 + \frac{2}{d}\big)^{2/\kappa}. \qquad (3.4.23)
\end{aligned}$$

The inequality (3.4.23) is true for all $\lambda \in (0, \vartheta)$. We may take $\lambda = \frac{\vartheta}{2}$. It follows that

$$\mathbb{E}\big[\mathcal{D}(x^{(n)}) - \mathcal{D}(x)\big] \le C_{\vartheta,\kappa,\mu} \cdot \frac{K}{\sqrt{n}} \Big[1 + (\log n)^{2/\kappa} + \gamma_K (\log K)^{2/\kappa} \big(1 + \frac{2}{d}\big)^{2/\kappa}\Big], \quad (3.4.24)$$

where $C_{\vartheta,\kappa,\mu} = \Big[6\big(\frac{2}{\vartheta}\big)^{2/\kappa} \cdot (1 \vee \log 2\mathbb{E}\, e^{\vartheta|X|^\kappa/2})\Big] \vee 8\vartheta^{-2/\kappa}$ and $\limsup_K \gamma_K = 1$.

Multi-dimensional normal distribution is a special case of hyper-exponential tail

distribution, i.e. if $\mu = \mathcal{N}(m, \Sigma)$, we have $\kappa = 2, \vartheta = \frac{1}{2}$ and $c = 0$. By the same reasoning as before,

$$\mathbb{E}\big[\mathcal{D}(x^{(n)}) - \mathcal{D}(x)\big] \leq C_\mu \cdot \frac{K}{\sqrt{n}}\Big[1 + \log n + \gamma_K \log K\big(1 + \frac{2}{d}\big)\Big],$$

where $C_\mu = 24 \cdot \big(1 \vee \log 2\mathbb{E}\, e^{|X|^2/4}\big)$. When $\mu = \mathcal{N}(0, \mathbf{I}_d)$, $C_\mu = 24(1 + \frac{d}{2}) \cdot \log 2$, since $\mathbb{E}\, e^{|X|^2/4} = 2^{d/2}$ by the moment-generating function of χ^2 distribution. □

3.5 Appendix

3.5.1 Appendix A: Proof of Proposition 3.1.1 - (iii)

Proof. Assume that there exists a $x^* = (x_1^*, ..., x_K^*) \in G_K(\mu)$ in which there exists $k \in \{1, ..., K\}$ such that $x_k^* \notin \mathcal{H}_\mu$.

Case (I): $\mu\big(V_{x_k^*}^o(\Gamma^*) \cap \text{supp}(\mu)\big) = 0$. After (3.1.16), the distortion function can be written as

$$\mathcal{D}_{K,\mu}(x^*) = \sum_{i=1}^K \int_{C_{x_i}(x)} |\xi - x_i^*|^2 \, \mu(d\xi) = \sum_{i=1}^K \int_{V_{x_i}^o(x)} |\xi - x_i^*|^2 \, \mu(d\xi)$$

(Since x^* is optimal and $|\cdot|$ is Euclidean, $\mu\big(\partial V_{x_i}(\Gamma^*)\big) = 0$ and $\text{int}V_{x_i}(\Gamma) = V_{x_i}^o(\Gamma)$)

$$= \sum_{i=1, i\neq k}^K \int_{V_{x_i}^o(x)} |\xi - x_i^*|^2 \, \mu(d\xi) = \mathcal{D}_{K,\mu}(\widetilde{x}), \tag{3.5.1}$$

where $\widetilde{x} = (x_1^*, ..., x_{k-1}^*, x_{k+1}^*, ..., x_K^*)$. Therefore, $\widetilde{\Gamma} = \{x_1^*, ..., x_{k-1}^*, x_{k+1}^*, ..., x_K^*\}$ is also a K-level optimal quantizer with $\text{card}(\widetilde{\Gamma}) < K$, contradictory to Proposition 3.1.1 - (i).

Case (II): $\mu\big(V_{x_k^*}^o(\Gamma^*) \cap \text{supp}(\mu)\big) > 0$. Since $x_k^* \neq \mathcal{H}_\mu$, there exists a hyperplane H strictly separating x_k^* and $\mathcal{H}_\mu$. Let $\hat{x}_k^*$ be the orthogonal projection of x_k^* on H. For any $z \in \mathcal{H}_\mu$, let b denote the point in the segment joining z and x_k^* which lies on H, then $\langle b - \hat{x}_k^* | x_k^* - \hat{x}_k^* \rangle = 0$. Hence,

$$|x_k^* - b|^2 = |\hat{x}_k^* - b|^2 + |x_k^* - \hat{x}_k^*|^2 > |\hat{x}_k^* - b|^2 .$$

Therefore, $|z - \hat{x}_k^*| \leq |z - b| + |b - \hat{x}_k^*| < |z - b| + |x_k^* - b| = |z - x_k^*|$.

Let $B(x, r)$ denote the ball on $\mathbb{R}^d$ centered at x with radius r. Since $\mu\big(V_{x_k^*}^o(\Gamma^*) \cap \text{supp}(\mu)\big) > 0$, there exists $\alpha \in V_{x_k^*}^o(\Gamma^*) \cap \text{supp}(\mu)$ such that $\exists r \geq 0$, $\mu\big(B(\alpha, r)\big) > 0$

(when $r = 0$, $B(\alpha, r) = \{r\}$). Moreover,

$$\forall \beta \in B(\alpha, r), \quad |\beta - \hat{x}_k^*| < |\beta - x_k^*| < \min_{i \neq k} |\beta - \hat{x}_i^*| \,. \tag{3.5.2}$$

Let $\hat{x} := (x_1^*, ..., x_{k-1}^*, \hat{x}_k^*, x_{k+1}^*, ..., x_K^*)$, (3.5.2) implies $\mathcal{D}_{K,\mu}(\hat{x}) < \mathcal{D}_{K,\mu}(x^*)$. This is contradictory with the fact that x^* is an optimal quantizer. Hence, $x^* \in \mathcal{H}_\mu$. □

3.5.2 Appendix B: Proof of Pollard's Theorem

Proof of Pollard's Theorem. Since the quantization level K is fixed, in this proof, we will withdraw K in the subscript of the distortion function $\mathcal{D}_{K,\mu}$ and denote by $\mathcal{D}_n$ (*respectively,* $\mathcal{D}_\infty$) as the distortion function of μ_n (*resp.* μ_∞).

We know $x^{(n)} \in \operatorname{argmin} \mathcal{D}_n$ owing to Proposition 3.1.1, that is, for all $y \in (y_1, ..., y_K) \in (\mathbb{R}^d)^K$, we have $\mathcal{D}_n(x^{(n)}) \leq \mathcal{D}_n(y)$. For every fixed $y = (y_1, ..., y_K)$, we have $\mathcal{D}_n(y) \to \mathcal{D}_\infty(y)$ after (3.1.19), then

$$\limsup_n \mathcal{D}_n(x^{(n)}) \leq \inf_{y \in (\mathbb{R}^d)^K} \mathcal{D}_\infty(y) \tag{3.5.3}$$

We assume that there exists an index set $\mathcal{I} \subset \{1, ..., K\}$ and $\mathcal{I}^c \neq \emptyset$ such that $(x_i^{(n)})_{i \in \mathcal{I}, n \geq 1}$ is bounded and $(x_i^{(n)})_{i \in \mathcal{I}^c, n \geq 1}$ is not bounded. Then there exists a subsequence $\psi(n)$ of n such that

$$\begin{cases} x_i^{\psi(n)} \to \widetilde{x}_i^{(\infty)} & i \in \mathcal{I} \\ \left| x_i^{\psi(n)} \right| \to +\infty & i \in \mathcal{I}^c \end{cases}$$

After (3.1.19), we have $\mathcal{D}_{\psi(n)}(x^{(\psi(n))})^{1/2} \geq \mathcal{D}_\infty(x^{(\psi(n))})^{1/2} - \mathcal{W}_2(\mu_{\psi(n)}, \mu_\infty)$. Hence,

$$\liminf_n \mathcal{D}_{\psi(n)}(x^{(\psi(n))})^{1/2} \geq \liminf_n \mathcal{D}_\infty(x^{(\psi(n))})^{1/2},$$

so that

$$\begin{aligned} \liminf_n \mathcal{D}_{\psi(n)}(x^{(\psi(n))})^{1/2} &\geq \liminf_n \mathcal{D}_\infty(x^{(\psi(n))})^{1/2} \\ &= \Big[\liminf_n \int \min_{i \in \{1,...,K\}} \left| x_i^{(\psi(n))} - \xi \right|^2 \mu_\infty(d\xi) \Big]^{1/2} \\ &\geq \Big[\int \liminf_n \min_{i \in \{1,...,K\}} \left| x_i^{(\psi(n))} - \xi \right|^2 \mu_\infty(d\xi) \Big]^{1/2} \\ &= \Big[\int \min_{i \in \mathcal{I}} \left| x_i^{(\infty)} - \xi \right|^2 \mu_\infty(d\xi) \Big]^{1/2}. \end{aligned} \tag{3.5.4}$$

Thus, (3.5.3) and (3.5.4) imply that

$$\int \min_{i\in\mathcal{I}} \left|x_i^{(\infty)} - \xi\right|^2 \mu_\infty(d\xi) \le \inf_{y\in(\mathbb{R}^d)^K} \mathcal{D}_\infty(y). \tag{3.5.5}$$

This implies that $\mathcal{I} = \{1, ..., K\}$ after Proposition 3.1.1 (otherwise, (3.5.5) implies that $e^{|\mathcal{I}|,*}(\mu_\infty) \le e^{K,*}(\mu_\infty)$ with $|\mathcal{I}| < K$, which is contradictory to Proposition 3.1.1-(i)). Therefore, $(x^{(n)})$ is bounded and any limiting point $x^{(\infty)} \in \operatorname{argmin}_{x\in(\mathbb{R}^d)^K} \mathcal{D}_\infty(x)$. □

3.5.3 Appendix C: Proof of Lemma 3.3.2

Recall that $F_K := \{x = (x_1, ..., x_K) \in (\mathbb{R}^d)^K \mid x_i \ne x_j, i \ne j\}$. We first prove that if $x \in F_K$ and y is close enough to x, then $y \in F_K$.

Lemma 3.5.1. *If $x \in F_K$, then any point y such that $y \in B\big(x, \frac{1}{3}\min_{1\le i,j\le K, i\ne j} |x_i - x_j|\big)$ lies still in F_K.*

Proof of Lemma 3.5.1. If there exist $i, j \in \{1, ..., K\}, i \ne j$ such that $y_i = y_j$, then

$$|x_i - x_j| \le |x_i - y_i| + |y_j - x_j| \le \frac{2}{3} \min_{1\le i,j\le K, i\ne j} |x_i - x_j|,$$

which is contradictory. Hence, $y \in F_K$. □

Now we prove Lemma 3.3.2.

Proof of Lemma 3.3.2. We will only prove the continuity of $\frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_1 \partial x_2}$ and $\frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_1^2}$ in a point $x \in F_K$. For the continuity of $\frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_i \partial x_j}$ for any others $i, j \in \{1, ..., K\}$ the proof is similar.

Let

$$\alpha(x, \xi) := (x_1 - \xi) \otimes (x_2 - \xi) \cdot \frac{1}{|x_2 - x_1|} f(\xi).$$

Then

$$\frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_1 \partial x_2}(x) = 2 \int_{V_1(x)\cap V_2(x)} \alpha(x, \xi) \lambda_x^{12}(d\xi).$$

Let $(e_1, ..., e_d)$ denote the canonical basis of $\mathbb{R}^d$. Set $u^x = \frac{x_1 - x_2}{|x_1 - x_2|}$. As $x_1 \ne x_2$, if we write the coordinate of u^x by $u^x = \sum_{i=1}^d u_i e_i$, then there exists at least one $i_0 \in \{1, ..., d\}$ s.t. $u_{i_0} \ne 0$. Then $(u^x, e_i, 1 \le i \le d, i \ne i_0)$ forms a new basis of $\mathbb{R}^d$. Applying the Gram-Schmidt orthonormalization procedure, we derive the existence of a new orthonormal basis $(u_1^x, ..., u_d^x)$ of $\mathbb{R}^d$ such that $u_1^x = u^x$. Moreover, the Gram-Schmidt orthonormalization procedure also implies that $u_i^x, 1 \le i \le d$ is continuous in x. With

respect to the new basis $(u_1^x, ..., u_d^x)$, the hyperplan M_{12}^x defined in (3.3.4) can be written by

$$M_{12}^x = \frac{x_1 + x_2}{2} + \text{span}\big(u_i^x,\ i = 2, ..., d.\big),$$

where $\text{span}(S)$ denotes the subspace of $\mathbb{R}^d$ spanned by a set S. Moreover, remark that

$$V_1(x) \cap V_2(x) = \big\{\xi \in M_{12}^x \ \big|\ \min_{k=3,...,K} |x_k - \xi| \geq |x_1 - \xi| = |x_2 - \xi|\ \big\}.$$

Then for every fixed $\xi \notin \partial\big(V_1(x) \cap V_2(x)\big)$, the function $x \mapsto \mathbb{1}_{V_1(x)\cap V_2(x)}(\xi)$ is continuous in $x \in F_K$ and

$$\lambda_x^{12}\Big(\partial\big(V_1(x) \cap V_2(x)\big)\Big) = 0 \tag{3.5.6}$$

since $V_1(x) \cap V_2(x)$ is a polyhedral convex set in M_{12}^x.

Now by a change of variable $\xi = \frac{x_1+x_2}{2} + \sum_{i=2}^d r_i u_i^x$,

$$\frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_1 \partial x_2}(x) = 2 \int_{\mathbb{R}^{d-1}} \mathbb{1}_{V_{12}(x)}\big((r_2, ..., r_d)\big)\alpha\Big(x, \frac{x_1 + x_2}{2} + \sum_{i=2}^d r_i u_i^x\Big) dr_2 ... dr_d,$$

where

$$V_{12}(x) \coloneqq \Big\{(r_2, ..., r_d) \in \mathbb{R}^{d-1} \ \Big|\ \min_{3\leq k\leq K} \Big|x_k - \frac{x_1 + x_2}{2} - \sum_{i=2}^d r_i u_i^x\Big| \geq \Big|\frac{x_1 - x_2}{2} - \sum_{i=2}^d r_i u_i^x\Big|\Big\}.$$

Let

$$\partial V_{12}(x) \coloneqq \Big\{(r_2, ..., r_d) \in \mathbb{R}^{d-1} \ \Big|\ \min_{3\leq k\leq K} \Big|x_k - \frac{x_1 + x_2}{2} - \sum_{i=2}^d r_i u_i^x\Big| = \Big|\frac{x_1 - x_2}{2} - \sum_{i=2}^d r_i u_i^x\Big|\Big\}.$$

Then (3.5.6) implies that $\lambda_{\mathbb{R}^{d-1}}\big(\partial V_{12}(x)\big) = 0$ where $\lambda_{\mathbb{R}^{d-1}}$ denotes the Lebesgue measure of the subspace $\text{span}\big(u_i^x, i = 2, ..., d\big)$.

Let us now consider a sequence $x^{(n)} = (x_1^{(n)}, ..., x_K^{(n)}) \in (\mathbb{R}^d)^K$ converging to a point $x = (x_1, ..., x_K) \in F_K$. By lemma 3.5.1, for n large enough, we have $x^{(n)} \in F_K$. For a fixed $(r_2, ..., r_d) \in \mathbb{R}^{d-1}$, the continuity of $x \mapsto \alpha(x, \frac{x_1+x_2}{2} + \sum_{i=2}^d r_i u_i^x)$ in F_K can be obtained by the continuity of $(x, \xi) \mapsto \alpha(x, \xi)$ and the continuity of Gram-Schmidt orthonormalization procedure.

Moreover, it is obvious that for any $a = (a_1, ..., a_d), b = (b_1, ..., b_d) \in \mathbb{R}^d$, we have $|a_i b_j| \leq |a|\,|b|, 1 \leq i, j \leq d$. Thus the absolute value of every term in the matrix

$$\alpha(x, \frac{x_1 + x_2}{2} + \sum_{i=2}^d r_i u_i^x)$$

$$= \frac{\left(\frac{x_1-x_2}{2} - \sum_{i=2}^d r_i u_i^x\right) \otimes \left(\frac{x_2-x_1}{2} - \sum_{i=2}^d r_i u_i^x\right)}{|x_2 - x_1|} f\Big(\frac{x_1+x_2}{2} + \sum_{i=2}^d r_i u_i^x\Big) \tag{3.5.7}$$

can be upper-bounded by

$$\begin{aligned}
&\frac{\left|\frac{x_1-x_2}{2} - \sum_{i=2}^d r_i u_i^x\right| \left|\frac{x_2-x_1}{2} - \sum_{i=2}^d r_i u_i^x\right|}{|x_2 - x_1|} f\Big(\frac{x_1+x_2}{2} + \sum_{i=2}^d r_i u_i^x\Big) \\
&\leq \frac{\Big(\left|\frac{x_1-x_2}{2}\right| + \left|\sum_{i=2}^d r_i u_i^x\right|\Big)^2}{|x_2 - x_1|} f\Big(\frac{x_1+x_2}{2} + \sum_{i=2}^d r_i u_i^x\Big) \\
&\leq C_x(1 + \sum_{i=2}^d r_i^2) f\Big(\frac{x_1+x_2}{2} + \sum_{i=2}^d r_i u_i^x\Big)
\end{aligned} \tag{3.5.8}$$

where $C_x > 0$ is a constant depending x.

The distribution μ is assumed to have a d-the radial-controlled tail. Recall that this means there exist a constant $A > 0$ and a continuous and decreasing function $g : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\forall \xi \in \mathbb{R}^d, |\xi| \geq A, \qquad f(\xi) \leq g(|\xi|) \text{ and } \int_{\mathbb{R}_+} x^d g(x) dx < +\infty. \tag{3.5.9}$$

Now let $K \coloneqq \frac{1}{2} \sup_n \left|x_1^{(n)} + x_2^{(n)}\right| \vee A$ and let $r \coloneqq \sum_{i=2}^d r_i u_i^x$. As g is a decreasing function, it follows that

$$\begin{aligned}
&C_x(1 + \sum_{i=2}^d r_i^2) f\Big(\frac{x_1^{(n)} + x_2^{(n)}}{2} + \sum_{i=2}^d r_i u_i^x\Big) \\
&\leq C_x(1+|r|^2) \sup_{\xi \in B(\mathbf{0}, 3K)} f(\xi) \mathbb{1}_{\{|r| \leq 2K\}} + C_x(1+|r|^2) g\Big(\Big|\frac{x_1^{(n)} + x_2^{(n)}}{2} + \sum_{i=2}^d r_i u_i^x\Big|\Big) \mathbb{1}_{\{|r| \geq 2K\}}. \\
&\leq C_x(1+|r|^2) \sup_{\xi \in B(\mathbf{0}, 3K)} f(\xi) \mathbb{1}_{\{|r| \leq 2K\}} + C_x(1+|r|^2) g\big(|r| - K\big) \mathbb{1}_{\{|r| \geq 2K\}}.
\end{aligned} \tag{3.5.10}$$

By a change of variable to polar coordinate system, one obtains by letting $\gamma = |r|$

$$\begin{aligned}
&\int_{\mathbb{R}^{d-1}} C_x |r|^2 g\big(|r| - K\big) \mathbb{1}_{\{|r| \geq 2K\}} dr_2 ... dr_d \\
&\leq C_{x,d} \int_{\mathbb{R}_+} \gamma^2 g(\gamma - K) \mathbb{1}_{\{\gamma \geq 2K\}} \gamma^{d-2} d\gamma \leq C_{x,d} \int_K^\infty (\gamma + K)^d g(\gamma) d\gamma \\
&\leq 2^d C_{x,d} \int_K^\infty (K^d + \gamma^d) g(\gamma) d\gamma < +\infty,
\end{aligned}$$

where the last inequality is owing to (3.5.9). Thus one obtains

$$\int_{\mathbb{R}^{d-1}} \Big[C_x(1+|r|^2) \sup_{\xi \in B(\mathbf{0},3K)} f(\xi) \mathbb{1}_{\{|r| \le 2K\}} + C_x(1+|r|^2) g\big(|r|-K\big) \mathbb{1}_{\{|r| \ge 2K\}} \Big] dr_2 ... dr_d < +\infty,$$

which implies $\frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_1 \partial x_2}(x^{(n)}) \to \frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_1 \partial x_2}(x)$ as $n \to +\infty$ by applying Lebesgue's dominated convergence theorem. Thus $\frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_1 \partial x_2}$ is continuous in $x \in F_K$.

It remains to prove the continuity of $x \mapsto \mu\big(V_1(x)\big) = \int_{\mathbb{R}^d} \mathbb{1}_{V_1(x)}(\xi) f(\xi) \lambda_d(d\xi)$ to obtain the continuity of $\frac{\partial^2 \mathcal{D}_{K,\mu}}{\partial x_1^2}$ defined in (3.3.7). Remark that

$$V_1(x) = \Big\{ \xi \in \mathbb{R}^d \;\Big|\; |\xi - x_1| \le \min_{1 \le j \le K} |\xi - x_j| \Big\},$$

and by Graf and Luschgy (2000)[Proposition 1.3],

$$\partial V_1(x) = \Big\{ \xi \in \mathbb{R}^d \;\Big|\; |\xi - x_1| = \min_{1 \le j \le K} |\xi - x_j| \Big\}.$$

Then for any $\xi \notin \partial V_1(x)$, the function $x \mapsto \mathbb{1}_{V_1(x)}(\xi)$ is continuous. As the norm $|\cdot|$ is the Euclidean norm, then $\lambda_d(\partial V_i(x)) = 0$ (see Graf and Luschgy (2000)[Proposition 1.3 and Theorem 1.5]). For any $x \in F_K$ and a sequence $x^{(n)}$ converging to x, we have $\mathbb{1}_{V_1(x^{(n)})}(\xi) f(\xi) \le f(\xi) \in L^1(\lambda_d)$. Thus the continuity of $x \mapsto \mu\big(V_1(x)\big) = \int_{\mathbb{R}^d} \mathbb{1}_{V_1(x)}(\xi) f(\xi) \lambda_d(d\xi)$ is a direct application of Lebesgue's dominated convergence theorem. □

3.5.4 Appendix D: Proof of Proposition 3.3.1

Proof. (i) We will only prove for the uniform distribution $U([0,1])$. The proof is similar for other uniform distributions.

In Graf and Luschgy (2000)[see Example 4.17 and 5.5], the authors show that $\Gamma^* = \{\frac{2i-1}{2K} : i - 1, ..., K\}$ is the unique optimal quantizers of $U([0,1])$. Let $x^* = (\frac{1}{2K}, ..., \frac{2i-1}{2K}, ..., \frac{2K-1}{2K})$, then one can compute explicitly $H_{\mathcal{D}}(x^*)$:

$$H_{\mathcal{D}}(x^*) = \begin{bmatrix} \frac{3}{2K} & -\frac{1}{2K} & & & & & 0 \\ & \ddots & \ddots & \ddots & & & \\ & & -\frac{1}{2K} & \frac{1}{K} & -\frac{1}{2K} & & \\ & & & \ddots & \ddots & \ddots & \\ 0 & & & & & -\frac{1}{2K} & \frac{3}{2K} \end{bmatrix}, \tag{3.5.11}$$

The matrix $H_{\mathcal{D}}(x^*)$ is tridiagonal. If we denote by $f_k(x^*)$ its k-th leading principal minor and we define $f_0(x^*) = 1$, then

$$f_k(x^*) = \frac{1}{K} f_{k-1}(x^*) - \frac{1}{4K^2} f_{k-2}(x^*) \quad \text{for } k = 2, ..., K-1, \tag{3.5.12}$$

and $f_1(x^*) = \frac{3}{2K}$ and $f_K(x^*) = |H_{\mathcal{D}}(x^*)| = \frac{3}{K} f_{K-1}(x^*) - \frac{1}{4K^2} f_{K-2}(x^*)$ (see El-Mikkawy (2003)). One can solve from the three-term recurrence relation that

$$f_k(x^*) = \frac{2k+1}{2^k K^k}, \quad \text{for } k = 1, ..., K-1 \tag{3.5.13}$$

$$\text{And} \quad f_K(x^*) = \frac{2K+1}{2^K K^K} + \frac{1}{2K} f_{K-1}. \tag{3.5.14}$$

In fact, (3.5.13) is true for $k = 1$. Suppose (3.5.13) holds for $k \le K-2$, then owing to (3.5.12)

$$f_{k+1}(x^*) = \frac{1}{K} \cdot \frac{2k+1}{2^k K^k} - \frac{1}{4K^2} \cdot \frac{2(k-1)+1}{2^{k-1} K^{k-1}} = \frac{2(k+1)+1}{2^{k+1} K^{k+1}}.$$

Then it is obvious that $f_k(x^*) > 0$ for $k = 1, ..., K$. Thus, $H_{\mathcal{D}}(x^*)$ is positive definite.

(ii) We define for $i = 2, ..., K$, $\widetilde{x}_i^* = \frac{x_{i-1}^* + x_i^*}{2}$, then the Voronoi region $V_i(x^*) = [\widetilde{x}_i^*, \widetilde{x}_{i+1}^*]$ for $i = 2, ..., K-1$, $V_1(x^*) = (-\infty, \widetilde{x}_2^*]$ and $V_K(x^*) = [\widetilde{x}_K^*, +\infty)$.

For $2 \le i \le K-1$,

$$\begin{aligned}
L_i(x^*) &= A_i - 2B_{i-1,i} - 2B_{i,i+1} \\
&= 2\mu\big(V_i(x^*)\big) - (x_i^* - x_{i-1}^*) f\Big(\frac{x_{i-1}^* + x_i^*}{2}\Big) - (x_{i+1}^* - x_i^*) f\Big(\frac{x_i^* + x_{i+1}^*}{2}\Big) \\
&= 2\mu\big(V_i(x^*)\big) - 2(x_i^* - \widetilde{x}_i^*) f(\widetilde{x}_i^*) - 2(\widetilde{x}_{i+1}^* - x_i^*) f(\widetilde{x}_{i+1}^*) \\
&= \frac{2}{\mu\big(V_i(x^*)\big)} \Big\{ \mu\big(V_i(x^*)\big)^2 - [x_i^* \mu\big(V_i(x^*)\big) \\
&\quad - \widetilde{x}_i^* \mu\big(V_i(x^*)\big)] f(\widetilde{x}_i^*) - [\widetilde{x}_{i+1}^* \mu\big(V_i(x^*)\big) - x_i^* \mu\big(V_i(x^*)\big)] f(\widetilde{x}_{i+1}^*) \Big\} \\
&= \frac{2}{\mu\big(V_i(x^*)\big)} \Big\{ \mu\big(V_i(x^*)\big)^2 - [\int_{V_i(x^*)} \xi f(\xi) d\xi - \widetilde{x}_i^* \int_{V_i(x^*)} f(\xi) d\xi] f(\widetilde{x}_i^*) \\
&\quad - [\widetilde{x}_{i+1}^* \int_{V_i(x^*)} f(\xi) d\xi - \int_{V_i(x^*)} \xi f(\xi) d\xi] f(\widetilde{x}_{i+1}^*) \Big\} \quad \big(\text{owing to (3.3.9)}\big) \\
&= \frac{2}{\mu\big(V_i(x^*)\big)} \underbrace{\Big\{ \mu\big(V_i(x^*)\big)^2 - f(\widetilde{x}_i^*) \int_{V_i(x^*)} (\xi - \widetilde{x}_i^*) f(\xi) d\xi + f(\widetilde{x}_{i+1}^*) \int_{V_i(x^*)} (\xi - \widetilde{x}_{i+1}^*) f(\xi) d\xi \Big\}}_{=: D_i(x^*)}.
\end{aligned} \tag{3.5.15}$$

For $u = (u_1, ..., u_{K+1}) \in F_{K+1}^+$, we define a function $\varphi_i(u)$ in order to study the

positivity of $D_i(x^*)$, for any $i \in \{1, ..., K\}$,

$$\varphi_i(u) \coloneqq \Big[\int_{u_i}^{u_{i+1}} f(\xi)d\xi\Big]^2 - f(u_i)\int_{u_i}^{u_{i+1}}(\xi - u_i)f(\xi)d\xi + f(u_{i+1})\int_{u_i}^{u_{i+1}}(\xi - u_{i+1})f(\xi)d\xi, \tag{3.5.16}$$

Lemma 3.5.2. *If f is positive and differentiable and if $\log f$ is strictly concave, then for all $u = (u_1, ..., u_{K+1}) \in F^+_{K+1}$, we have the following results for $\varphi_i(u)$ defined in (3.5.16),*

(a) for every $i = 1, ..., K$, $\varphi_i(u) > 0$;

(b) $\frac{\partial \varphi_1}{\partial u_1}(u) < 0$;

(c) $\frac{\partial \varphi_K}{\partial u_{K+1}}(u) > 0$.

Proof of lemma 3.5.2. For a fixed $i \in \{1, ..., K\}$, the partial derivatives of φ_i are

$$\begin{aligned}
\frac{\partial \varphi_i}{\partial u_i}(u) &= -2\Big[\int_{u_i}^{u_{i+1}} f(\xi)d\xi\Big]f(u_i) - f'(u_i)\int_{u_i}^{u_{i+1}}(\xi - u_i)f(\xi)d\xi + f(u_i)f(u_{i+1})(u_{i+1} - u_i)\\
\frac{\partial \varphi_i}{\partial u_{i+1}}(u) &= 2\Big[\int_{u_i}^{u_{i+1}} f(\xi)d\xi\Big]f(u_{i+1}) + f'(u_{i+1})\int_{u_i}^{u_{i+1}}(\xi - u_{i+1})f(\xi)d\xi\\
&\qquad - f(u_i)f(u_{i+1})(u_{i+1} - u_i)\\
\frac{\partial \varphi_i}{\partial u_l}(u) &= 0, \quad \text{for all } l \neq i \text{ and } l \neq i+1.
\end{aligned} \tag{3.5.17}$$

The second derivatives of φ_i are

$$\begin{aligned}
\frac{\partial^2 \varphi_i}{\partial u_{i+1}\partial u_i}(u) &= \frac{\partial^2 \varphi_i}{\partial u_i \partial u_{i+1}}(u) = -f(u_{i+1})f(u_i) + (u_{i+1} - u_i)\big(f(u_i)f'(u_{i+1}) - f'(u_i)f(u_{i+1})\big)\\
\frac{\partial^2 \varphi_i}{\partial u_l \partial u_i}(u) &= \frac{\partial^2 \varphi_i}{\partial u_i \partial u_l}(u) = 0 \quad \text{for all } l \neq i \text{ and } l \neq i+1.
\end{aligned} \tag{3.5.18}$$

If $\log f$ is strictly concave, then $(\log f)' = \dfrac{f'}{f}$ is strictly decreasing. For $u \in F^+_{K+1}$, we have $u_{i+1} > u_i$, then

$$\frac{f'(u_{i+1})}{f(u_{i+1})} - \frac{f'(u_i)}{f(u_i)} = \frac{f'(u_{i+1})f(u_i) - f(u_{i+1})f'(u_i)}{f(u_i)f(u_{i+1})} < 0.$$

Thus $f'(u_{i+1})f(u_i) - f(u_{i+1})f'(u_i) < 0$ and from which one can get $\dfrac{\partial^2 \varphi_i}{\partial u_{i+1}\partial u_i}(u) < 0$.

In fact, φ_i, $\frac{\partial \varphi_i}{\partial u_i}$, $\frac{\partial \varphi_i}{\partial u_{i+1}}$ and $\frac{\partial^2 \varphi_i}{\partial u_{i+1}\partial u_i}$ are functions of only (u_i, u_{i+1}).

(a) For $1 \leq i \leq K$, $\varphi_i(u_{i+1}, u_{i+1}) = 0$. After the Mean value theorem, there exists a $\gamma \in (u_i, u_{i+1})$ such that

$$\frac{1}{u_i - u_{i+1}}\big(\varphi_i(u_i, u_{i+1}) - \varphi_i(u_{i+1}, u_{i+1})\big) = \frac{\partial \varphi_i}{\partial u_i}(\gamma, u_{i+1}). \tag{3.5.19}$$

Moreover, there exists a $\zeta \in (\gamma, u_{i+1})$ such that

$$\frac{1}{u_{i+1} - \gamma}\Big(\frac{\partial \varphi_i}{\partial u_i}(\gamma, u_{i+1}) - \frac{\partial \varphi_i}{\partial u_i}(\gamma, \gamma)\Big) = \frac{\partial^2 \varphi_i}{\partial u_{i+1} \partial u_i}(\gamma, \zeta).$$

As $\gamma < \zeta$, $\dfrac{\partial^2 \varphi_i}{\partial u_{i+1} \partial u_i}(\gamma, \zeta) < 0$. Thus $\dfrac{\partial \varphi_i}{\partial u_i}(\gamma, u_{i+1}) < 0$, since $\dfrac{\partial \varphi_i}{\partial u_i}(\gamma, \gamma) = 0$. Then $\varphi_i(u_i, u_{i+1}) > 0$ by applying $\dfrac{\partial \varphi_i}{\partial u_i}(\gamma, u_{i+1}) < 0$ in (3.5.19).

(b) After the Mean value theorem, there exists a $\gamma' \in (u_1, u_2)$ such that

$$\frac{\partial^2 \varphi_1}{\partial u_1 \partial u_2}(u_1, \gamma') = \frac{1}{u_2 - u_1}\Big(\frac{\partial \varphi_1}{\partial u_1}(u_1, u_2) - \frac{\partial \varphi_1}{\partial u_1}(u_1, u_1)\Big).$$

As $\dfrac{\partial^2 \varphi_1}{\partial u_1 \partial u_2}(u_1, \gamma') < 0$ and $\dfrac{\partial \varphi_1}{\partial u_1}(u_1, u_1) = 0$, one can get $\dfrac{\partial \varphi_1}{\partial u_1}(u_1, u_2) < 0$.

(c) In the same way, there exists a $\zeta' \in (u_K, u_{K+1})$ such that

$$\frac{\partial^2 \varphi_K}{\partial u_K \partial u_{K+1}}(\zeta', u_{K+1}) = \frac{1}{u_K - u_{K+1}}\Big(\frac{\partial \varphi_K}{\partial u_{K+1}}(u_K, u_{K+1}) - \frac{\partial \varphi_K}{\partial u_{K+1}}(u_{K+1}, u_{K+1})\Big).$$

As $\dfrac{\partial^2 \varphi_K}{\partial u_K \partial u_{K+1}}(\zeta', u_{K+1}) < 0$ and $\dfrac{\partial \varphi_K}{\partial u_{K+1}}(u_{K+1}, u_{K+1}) = 0$, one can get $\dfrac{\partial \varphi_K}{\partial u_{K+1}}(u_K, u_{K+1}) > 0$. □

Proof of Proposition 3.3.1, continuation. We set $\widetilde{x}^{*,M} := (-M, \widetilde{x}_2^*, ..., \widetilde{x}_K^*, M)$ with a M large enough such that $\widetilde{x}^{*,M} \in F_{K+1}^+$, then for $2 \leq i \leq K-1$, $L_i(x^*) = \frac{2}{\mu(V_i(x^*))}\varphi_i(\widetilde{x}^{*,M})$. Thus $L_i(x^*) > 0$, $i = 2, ..., K-1$ owing to Lemma 3.5.2 (i).

For $i = 1$,

$$\begin{aligned} L_1(x^*) &= A_1(x^*) - 2B_{1,2}(x^*) \\ &= \frac{2}{\mu\big(V_1(x^*)\big)}\Big\{\mu\big(V_1(x^*)\big)^2 - f(\widetilde{x}_2^*)\int_{V_1(x^*)}(\widetilde{x}_2^* - \xi) f(\xi) d\xi\Big\}. \end{aligned}$$

If we denote $D_1(x^*) \coloneqq \mu\big(V_1(x^*)\big)^2 - f(\widetilde{x}_2^*)\int_{V_1(x^*)}(\widetilde{x}_2^* - \xi)f(\xi)d\xi$, then

$$D_1(x^*) = \lim_{M\to+\infty}\varphi_1(\widetilde{x}^{*,M}) + f(-M)\int_{V_1^M(x^*)}\big(\xi-(-M)\big)f(\xi)d\xi,$$

where $V_1^M(x^*) = [-M, \widetilde{x}_2^*]$.

For all M such that $-M < \widetilde{x}_2^*$, $f(-M)\int_{V_1^M(x^*)}\big(\xi-(-M)\big)f(\xi)d\xi > 0$, then

$$\lim_{M\to+\infty} f(-M)\int_{V_1^M(x^*)}\big(\xi-(-M)\big)f(\xi)d\xi \geq 0.$$

After Lemma 3.5.2 (ii), $\dfrac{\partial\varphi_1}{\partial u_1}(u) < 0$ for $u \in F_{K+1}^+$, so that for a fixed M_1 such that $\widetilde{x}^{*,M_1} \in F_{K+1}^+$, we have $\varphi_1(\widetilde{x}^{*,M_1}) \leq \lim\limits_{M\to+\infty}\varphi_1(\widetilde{x}^{*,M})$. We also have $\varphi_1(\widetilde{x}^{*,M_1}) > 0$ by applying Lemma 3.5.2 (1). It follows that

$$\begin{aligned} D_1(x^*) &= \lim_{M\to+\infty}\varphi_1(\widetilde{x}^{*,M}) + \lim_{M\to+\infty} f(-M)\int_{V_1^M(x^*)}\big(\xi-(-M)\big)f(\xi)d\xi \\ &\geq \varphi_1(\widetilde{x}^{*,M_1}) + \lim_{M\to+\infty} f(-M)\int_{V_1^M(x^*)}\big(\xi-(-M)\big)f(\xi)d\xi \\ &> 0. \end{aligned}$$

Then $L_1(x^*) = \dfrac{2}{\mu\big(V_1(x^*)\big)}D_1(x^*) > 0$.

The proof of $L_K(x^*)$ is similar by applying Lemma 3.5.2 (iii). Thus $H_{\mathcal{D}}(x^*)$ is positive definite owing to Gershgorin circle theorem. □

Part II:
McKean-Vlasov Equation: Particle Method, Quantization Based and Hybrid Scheme, Application to the Convex Ordering

Chapter 4

Introduction of Part II

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a filtered probability space and let $(E, \|\cdot\|_E)$ be a separable Banach space. For any random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \|\cdot\|_E)$, we denote by $P_X = \mathbb{P} \circ X^{-1}$ its probability distribution on $(E, \|\cdot\|_E)$ and denote by $\|X\|_p$ its L^p-norm defined by $\|X\|_p = \left[\mathbb{E}\,\|X\|_E^p\right]^{1/p}$.

Let $(B_t)_{t\geq 0}$ be an $(\mathcal{F}_t)-$standard Brownian motion defined on the probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ and valued in $\mathbb{R}^q$. Let $\mathbb{M}_{d,q}(\mathbb{R})$ denote the set of matrices with d rows and q columns, equipped with an operator norm $|||A||| \coloneqq \sup_{|z|\leq 1} |Az|_q$, where $|\cdot|_d$ denotes the norm on $\mathbb{R}^d$ (we drop the subscript d when there is no ambiguity). We consider an $\mathbb{R}^d$-valued *McKean-Vlasov Equation* defined by

$$\begin{cases} dX_t = b(t, X_t, \mu_t)dt + \sigma(t, X_t, \mu_t)dB_t \\ \forall t \geq 0,\ \mu_t \text{ denotes the probability distribution of } X_t, \end{cases} \tag{4.0.1}$$

where X_0 is an $\mathbb{R}^d$-valued random variable defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ and independent to Brownian motion $(B_t)_{t\geq 0}$, b, σ are Borel functions defined on $[0, T] \times \mathbb{R}^d \times \mathcal{P}_p(\mathbb{R}^d)$ having values in $\mathbb{R}^d$ and $\mathbb{M}_{d,q}(\mathbb{R})$ respectively.

For $p \in [1, +\infty)$, let $\mathcal{P}_p(\mathbb{R}^d)$ denote the set of probability distributions on $\mathbb{R}^d$ with p-th finite moment. For any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the *Wasserstein distance* $\mathcal{W}_p$ on $\mathcal{P}_p(\mathbb{R}^d)$ is defined by

$$\begin{aligned} \mathcal{W}_p(\mu, \nu) &= \Big(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} d(x, y)^p \pi(dx, dy) \Big)^{\frac{1}{p}} \\ &= \inf \Big\{ \Big[\mathbb{E}\,|X - Y|^p \Big]^{\frac{1}{p}},\ X, Y : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^d, Bor(\mathbb{R}^d)) \text{ with } P_X = \mu, P_Y = \nu \Big\}, \end{aligned} \tag{4.0.2}$$

where in the first ligne of (4.0.2), $\Pi(\mu,\nu)$ denotes the set of all probability measures on $(\mathbb{R}^d\times\mathbb{R}^d, Bor(\mathbb{R}^d)^{\otimes 2})$ with marginals μ and ν. For two $\mathbb{R}^d$-valued random variables X and Y with respective probability distributions μ and ν in $\mathcal{P}_p(\mathbb{R}^d)$, with an obvious abuse of notation, we will also denote by $\mathcal{W}_p(X,Y)$ to represent the L^p-Wasserstein distance between μ and ν.

We suppose throughout Part II:

Assumption (I): *There exists $p\in[2,+\infty)$ such that $\|X_0\|_p<+\infty$. Moreover, b,σ are continuous in t, Lipschitz continuous in x and in μ with Lipschitz constant L uniformly with respect to $t\in[0,T]$, i.e.*

$$\forall t\in[0,T],\forall x,y\in\mathbb{R}^d \text{ and } \forall\mu,\nu\in\mathcal{P}_p(\mathbb{R}^d),$$
$$|b(t,x,\mu)-b(t,y,\nu)|\vee|||\sigma(t,x,\mu)-\sigma(t,y,\nu)|||\le L\big[\,|x-y|+\mathcal{W}_p(\mu,\nu)\big].$$

In the so-called Vlasov case, that is, there exist $\beta:[0,T]\times\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}^d$ and $a:[0,T]\times\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{M}_{d,q}(\mathbb{R})$ such that

$$b(t,x,\mu)=\int_{\mathbb{R}^d}\beta(t,x,u)\mu(du)\ \text{ and }\ \sigma(t,x,\mu)=\int_{\mathbb{R}^d}a(t,x,u)\mu(du),\tag{4.0.3}$$

a sufficient condition to fulfill Assumption (I) is to assume β and a continuous in t, Lipschitz continuous in x and u uniformly with respect to $t\in[0,T]$, i.e.

$$\forall t\in[0,T],\forall x_1,x_2,u_1,u_2\in\mathbb{R}^d,$$
$$|\beta(t,x_1,u_1)-\beta(t,x_2,u_2)|\vee|a(t,x_1,u_1)-a(t,x_2,u_2)|\le L\big(\,|x_1-x_2|+|u_1-u_2|\,\big).$$

Chapter 5 is devoted to the proof of existence and uniqueness of a strong solution of the McKean-Vlasov equation and the convergence of theoretical Euler scheme. Our proof of existence and uniqueness of a strong solution of the McKean-Vlasov equation (4.0.1) under Assumption (I) is based on Feyel's method (see Bouleau (1988)[Section 7]). The idea is to define an application Φ_C depending on some constant $C\in\mathbb{R}_+$ on a product space, namely, "path space $\times$ path distribution space" as follows

$$\begin{aligned}(Y,P_Y)&\mapsto\Phi_C(Y,P_Y)\\ &:=\Big(\underbrace{\big(X_0+\int_0^t b(s,Y_s,\nu_s)ds+\int_0^t\sigma(s,Y_s,\nu_s)dB_s\big)_{t\in[0,T]}}_{=:\Phi_C^{(1)}(Y,P_Y)},P_{\Phi_C^{(1)}(Y,P_Y)}\Big)\end{aligned}\tag{4.0.4}$$

then to show that the product space is complete and that Φ_C is a contraction mapping

by controlling the value of C. Thus the existence and uniqueness of a strong solution of the McKean-Vlasov equation is a direct result by applying the fixed-point theorem. During the proof, we also give a rigorous definition of such "path space" and "path distribution space" which will be also used in the sections devoted to numerical schemes.

Once obtained the existence and uniqueness of a strong solution, we show in Section 5.2 the convergence rate of Euler scheme of the McKean-Vlasov equation (4.0.1). Let $M \in \mathbb{N}^*$ and let $h = \frac{T}{M}$. For $m = 0, ..., M$, define $t_m = t_m^M := m \cdot h = m \cdot \frac{T}{M}$. The Euler scheme of the McKean-Vlasov equation (4.0.1) is defined as follows,

$$\begin{cases} \bar{X}^M_{t_{m+1}} = \bar{X}^M_{t_m} + h \cdot b(t_m, \bar{X}^M_{t_m}, \bar{\mu}^M_{t_m}) + \sqrt{h}\,\sigma(t_m, \bar{X}^M_{t_m}, \bar{\mu}^M_{t_m}) Z_{m+1} \\ \bar{X}_0 = X_0 \end{cases}, \tag{4.0.5}$$

where $\bar{\mu}^M_{t_m}$ denotes the probability distribution of $\bar{X}^M_{t_m}$ and $Z_m, m = 0, ..., M$ are i.i.d random variables having an $\mathbb{R}^q$-standard normal distribution $\mathcal{N}(0, \mathbf{I}_q)$. When there is no ambiguity, we will omit the superscript M and use $\bar{X}_{t_m}$ and $\bar{\mu}_{t_m}$ instead of $\bar{X}^M_{t_m}$ and $\bar{\mu}^M_{t_m}$ in the following discussion.

We call (4.0.5) the "theoretical" Euler scheme since it does not directly indicate how to simulate $\bar{\mu}_{t_m}$ and we will propose several spatial discretizations later in Chapter 7 to simulate $\bar{\mu}_{t_m}$. In Section 5.2, we establish the following convergence rate of the theoretical Euler scheme

$$\sup_{0 \le m \le M} \mathcal{W}_p(\bar{\mu}_{t_m}, \mu_{t_m}) \le \left\| \sup_{0 \le m \le M} \left| X_{t_m} - \bar{X}_{t_m} \right| \right\|_p \le C_e h^{\frac{1}{2} \wedge \gamma}, \tag{4.0.6}$$

with C_e a constant depending on $b, \sigma, L, T, \tilde{L}$ and $\|X_0\|_p$, under Assumption (I) and the following condition

$$\begin{aligned} &\forall t, s \in [0, T],\ s < t,\ \forall x \in \mathbb{R}^d,\ \forall \mu \in \mathcal{P}(\mathbb{R}^d),\ \text{ there exist } \tilde{L}, \gamma \in \mathbb{R}_+ \text{ s.t.} \\ &\quad |b(t, x, \mu) - b(s, x, \mu)| \vee |||\sigma(t, x, \mu) - \sigma(s, x, \mu)||| \le \tilde{L}\big(1 + |x| + \mathcal{W}_p(\mu, \delta_0)\big)(t - s)^\gamma. \end{aligned} \tag{4.0.7}$$

In Chapter 6, we establish the functional convex order result for the scaled McKean-Vlasov equation. For any two random variables X, Y valued in a Banach space $(E, \|\cdot\|_E)$, if for any convex function $\varphi : E \to \mathbb{R}$ such that

$$\mathbb{E}\,\varphi(X) \le \mathbb{E}\,\varphi(Y) \ \text{ as soon as these two expectations make sense,}$$

then we call X is dominated by Y for the *convex order* and denote by $X \preceq_{cv} Y$. Let

$(X_t)_{t\in[0,T]}$, $(Y_t)_{t\in[0,T]}$ be two processes defined by

$$\begin{aligned} dX_t &= (\alpha X_t + \beta)dt + \sigma(t, X_t, \mu_t)dB_t, \quad X_0 \in L^p(\mathbb{R}^d), \\ dY_t &= (\alpha Y_t + \beta)\,dt + \theta(t,\, Y_t,\, \nu_t)\,dB_t, \quad Y_0 \in L^p(\mathbb{R}^d), \end{aligned} \tag{4.0.8}$$

where $\alpha, \beta \in \mathbb{R}$ and for any $t \in [0,T]$, $\mu_t = P_{X_t}$, $\nu_t = P_{Y_t}$. We first prove that the theoretical Euler scheme (4.0.5) of the McKean-Vlasov equation propagates the convex order of random variables. Let $\bar{X}_{t_m}, \bar{Y}_{t_m}, m = 0, ..., M$ respectively denote the theoretical Euler scheme of $(X_t)_{t\in[0,T]}$, $(Y_t)_{t\in[0,T]}$ defined by (4.0.5). If $X_0 \preceq_{cv} Y_0$ and the coefficient functions σ, θ are ordered by a matrix order in the sense that

$$\begin{aligned} &\forall t \in [0,T],\ \forall x \in \mathbb{R}^d,\ \forall \mu \in \mathcal{P}(\mathbb{R}^d), \\ &\qquad \theta(t,x,\mu)\theta(t,x,\mu)^* - \sigma(t,x,\mu)\sigma(t,x,\mu)^* \text{ is a positive semi-definite matrix,} \end{aligned}$$

and σ is convex in x and non-decreasing in μ with respect to the convex order, then for any $m = 0, ..., M$, $\bar{X}_{t_m} \preceq_{cv} \bar{Y}_{t_m}$. Moreover, owing to the convergence result of the theoretical Euler scheme (4.0.6), we derive a functional convex order result for the processes $X = (X_t)_{t\in[0,T]}$ and $Y = (Y_t)_{t\in[0,T]}$, i.e. for any convex function $F : \mathcal{C}([0,T], \mathbb{R}^d) \to \mathbb{R}$ having an r-polynomial growth with respect to the sup-norm, $1 \le r \le p$, in the sense that

$$\forall \alpha \in \mathcal{C}([0,T], \mathbb{R}^d), \exists C \ge 0 \ \ s.t. \ \ |F(\alpha)| \le C(1 + \|\alpha\|_{\sup}^r)$$

we have

$$\mathbb{E}\, F(X) \le \mathbb{E}\, F(Y). \tag{4.0.9}$$

This result generalizes the functional convex order results in Pagès (2016) established for the one dimensional martingale diffusion, which is the solution of stochastic differential equation $dX_t = \sigma(t, X_t)dB_t$. Furthermore, we generalize the above functional convex result (4.0.9) to a function

$$G : (\alpha, (\gamma_t)_{t\in[0,T]}) \in \mathcal{C}([0,T], \mathbb{R}^d) \times \mathcal{C}([0,T], \mathcal{P}_p(\mathbb{R}^d)) \to G(\alpha, (\gamma_t)_{t\in[0,T]}) \in \mathbb{R}$$

convex in α, non-decreasing for the convex order in $(\gamma_t)_{t\in[0,T]}$ and having a r-polynomial growth, $1 \le r \le p$ and obtain a new convex order result for X and Y and its probability distributions defined in (4.0.8) as follows,

$$\mathbb{E}\, G(X, (\mu_t)_{t\in[0,T]}) \le \mathbb{E}\, G(Y, (\nu_t)_{t\in[0,T]}).$$

Chapter 7 is devoted to the study of (several) simulable discretization schemes for the McKean-Vlasov equation. In order to simplify the notation, the discussion of Chapter 7 is based on the *homogeneous* McKean-Vlasov equation which means that the coefficient

functions b and σ do not depend on t, i.e.

$$(A): \quad \begin{cases} dX_t = b(X_t, \mu_t)dt + \sigma(X_t, \mu_t)dB_t \\ X_0 : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right) \text{ random variable indenpendent to } (B_t)_{t\in[0,T]} \\ \forall t \geq 0,\ \mu_t \text{ denotes the probability distribution of } X_t \end{cases}.$$

In the homogeneous setting, b and σ automatically satisfy the condition in (4.0.7). Let $X_0^{1,N}, ..., X_0^{N,N}$ be i.i.d random variables with the same distribution as X_0 in (A) and let $(B_t^n)_{t\geq 0}, n = 1, ..., N$ be i.i.d $\mathcal{F}_t$-standard Brownian motions independent to $(X_0^{1,N}, ..., X_0^{N,N})$. The N-*particle system* associated to the McKean-Vlasov equation (A) is defined by

$$(B): \quad \begin{cases} \forall n \in \{1, ..., N\}, \\ dX_t^{n,N} = b(X_t^{n,N}, \mu_t^N)dt + \sigma(X_t^{n,N}, \mu_t^N)dB_t^n, \\ \text{for any } t \in [0,T],\ \mu_t^N := \frac{1}{N}\sum_{n=1}^N \delta_{X_t^{n,N}}, \end{cases}$$

where δ_x denotes the Dirac mass at x. The convergence of μ_t^N to μ_t and the asymptotic mutual independence of the components $X_t^{n,N}$ as $n \to +\infty$ is usually called by *propagation of chaos* in the literature (see for example Gärtner (1988) and Lacker (2018), also Chassagneux et al. (2019) for a detailed analysis of the weak error).

We rewrite the theoretical Euler scheme in the homogeneous case,

$$(C): \quad \begin{cases} \bar{X}_{t_{m+1}} = \bar{X}_{t_m} + h \cdot b(\bar{X}_{t_m}, \bar{\mu}_{t_m}) + \sqrt{h}\,\sigma(\bar{X}_{t_m}, \bar{\mu}_{t_m})Z_{m+1} \\ \bar{X}_0 = X_0,\ \ \bar{\mu}_{t_m} = P_{\bar{X}_{t_m}} \end{cases},$$

and we propose several spatial discretizations in Chapter 7.

A first method of the spatial discretization is the *particle method* inspired by the N-particle system (B), which is the Euler scheme of the N-particle system (B). Let $\bar{X}_0^{1,N}, ..., \bar{X}_0^{N,N}$ be i.i.d copies of X_0 in (A). We take the same M and h as in (C) and the *particle method* is defined by

$$(D): \quad \begin{cases} \forall n \in \{1, ..., N\}, \\ \bar{X}_{t_{m+1}}^{n,N} = \bar{X}_{t_m}^{n,N} + hb(\bar{X}_{t_m}^{n,N}, \bar{\mu}_{t_m}^N) + \sqrt{h}\,\sigma(\bar{X}_{t_m}^{n,N}, \bar{\mu}_{t_m}^N)Z_{m+1}^n \\ \bar{\mu}_{t_m}^N := \frac{1}{N}\sum_{n=1}^N \delta_{\bar{X}_{t_m}^{n,N}} \end{cases},$$

where $Z_m^n, n = 1, ..., N, m = 0, ..., M \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \mathbf{I}_q)$. In the particle method, we use $\bar{\mu}_{t_m}^N$ as an estimator of $\bar{\mu}_{t_m}$ for each time step. In one dimensional setting, the convergence rate of $\bar{\mu}_{t_m}^N$ to $\bar{\mu}_m$ as $N \to +\infty$ has been established in Bossy and Talay (1997). For the

convergence rate in high dimension ($d \geq 2$), we obtain in Section 7.1 that

$$\left\| \sup_{1\leq m\leq M} \mathcal{W}_p(\bar{\mu}^N_{t_m}, \bar{\mu}_{t_m}) \right\|_p \leq C_{d,p,L,T} \left\| \mathbb{W}_p(\bar{\mu}, \nu^N) \right\|_p ,$$

where $\bar{\mu}$ denotes the probability distribution of $\bar{X} = (\bar{X}_t)_{t\in[0,T]}$ defined further in (5.2.3) and ν^N denotes the empirical measure of $\bar{\mu}$. Moreover, if $\|X_0\|_{p+\varepsilon} < +\infty$ for some $\varepsilon > 0$, then we also derive in Section 7.1 from recent results on empirical measures (see Fournier and Guillin (2015)) that

$$\left\| \sup_{1\leq m\leq M} \mathcal{W}_p(\bar{\mu}^N_{t_m}, \bar{\mu}_m) \right\|_p \leq \widetilde{C} \times \begin{cases} N^{-\frac{1}{2p}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p > d/2 \text{ and } \varepsilon \neq p \\ N^{-\frac{1}{2p}} \left[\log(1+N) \right]^{\frac{1}{p}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p = d/2 \text{ and } \varepsilon \neq p \\ N^{-\frac{1}{d}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p \in (0, d/2) \text{ and } p + \varepsilon \neq \frac{d}{(d-p)} \end{cases},$$

where $\widetilde{C}$ is a constant depending on $p, \varepsilon, d, b, \sigma, L, T$.

Another method to approximate $\bar{\mu}_{t_m}, m = 0, ..., M$ in the theoretical Euler scheme (C) is the *quadratic optimal quantization method*, which is also known as *K-means method*. Now we recall some definitions and properties of this method.

Let $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^d, |\cdot|)$ be a random variable with probability distribution $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, where $|\cdot|$ is the Euclidean norm on $\mathbb{R}^d$. The *quadratic quantization error function* at level K of Y (or *of* ν), denoted by $e_{K,\nu}$ (or $e_{K,Y}$), is defined by

$$y = (y_1, ..., y_K) \in (\mathbb{R}^d)^K \mapsto e_{K,\nu}(y) := \Big[\int_{\mathbb{R}^d} \min_{1\leq k\leq K} |\xi - y_k|^2 \, \nu(d\xi) \Big]^{1/2}. \tag{4.0.10}$$

Moreover, the *L^2-distortion function* of ν (or *of* Y) at level K, denoted by $\mathcal{D}_{K,\nu}$, is defined by $\mathcal{D}_{K,\nu} = e^2_{K,\nu}$.

In the framework of the optimal quantization, the variable $y \in (\mathbb{R}^d)^K$ of the quantization error function $e_{K,\nu}$ is called a *quantizer*. A K-tuple $y^* = (y^*_1, ..., y^*_K) \in (\mathbb{R}^d)^K$ is called an *optimal quantizer* of Y (or *of* ν) at level K if

$$y^* \in \operatorname{argmin} e_{K,\nu} \qquad (\text{ or equivalently, } \quad y^* \in \operatorname{argmin} \mathcal{D}_{K,\nu}). \tag{4.0.11}$$

For the proof of the existence of an optimal quantizer, we refer to Graf and Luschgy (2000)[Theorem 4.12] among other references.

Quantization theory has a close connection with the Voronoï partition. Let $y =$

$(y_1, ..., y_K)$ be a quantizer at level K. The *Voronoï cell* (or *Voronoï region*) generated by y_k is defined by

$$V_k(y) = V_{y_k}(y) := \big\{\xi \in \mathbb{R}^d \ : \ |\xi - y_k| = \min_{1\le j\le K} |\xi - y_j| \,\big\}, \tag{4.0.12}$$

and $\big(V_k(y)\big)_{1\le k\le K}$ is called the *Voronoï diagram* of y which is a finite covering of $\mathbb{R}^d$. A Borel partition $\big(C_k(y)\big)_{1\le k\le K}$ is called a *Voronoï partition* of $\mathbb{R}^d$ generated by y if

$$\forall\, k \in \{1, ..., K\}, \qquad C_k(y) \subset V_k(y). \tag{4.0.13}$$

The boundary of a Voronoï cell $V_k(y)$, denoted by $\partial V_k(y)$, is contained in $\cup_{j\neq k} H_{k,j}$, where $H_{k,j}$ is the median hyperplane of y_k and y_j

$$H_{k,j} := \big\{\, \xi \in \mathbb{R}^d \ : \ |\xi - y_k| = |\xi - y_j| \,\big\}.$$

For a fixed quantizer $y = (y_1, ..., y_K) \in (\mathbb{R}^d)^K$ and a Voronoï partition $\big(C_k(y)\big)_{1\le k\le K}$ generated by y, we can define a projection function Proj_y by

$$\xi \in \mathbb{R}^d \mapsto \mathrm{Proj}_y(\xi) = \sum_{k=1}^{K} y_k \mathbb{1}_{C_k(y)}(\xi). \tag{4.0.14}$$

Then for an $\mathbb{R}^d$-valued variable Y with probability distribution $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, we define its projection on y by

$$\widehat{Y}^y := \mathrm{Proj}_y(Y). \tag{4.0.15}$$

When there is no ambiguity, we write $\widehat{Y}$ instead of $\widehat{Y}^y$. If $y^* = (y_1^*, ..., y_K^*)$ is an optimal quantizer of ν and if $\widehat{\nu}^*$ denotes the probability distribution of $\mathrm{Proj}_{y^*}(Y)$, we have

$$e_{K,\nu}(y^*) = \left\| Y - \widehat{Y}^{y^*} \right\|_2 = \mathcal{W}_2(\nu, \widehat{\nu}^*) \tag{4.0.16}$$

and $\nu\big(\partial V_k(y^*)\big) = 0$ for every $k = 1, ..., K$ (see Graf and Luschgy (2000)[Lemma 3.4 and Theorem 4.2] for the proof of (4.0.16)).

The optimal quantizer has the following properties,

Proposition 4.0.1. (*a*) *(Stationary of optimal quantization) Let X be a random variable with probability distribution $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and assume that* $\mathrm{card}(\mathrm{supp}(\nu)) \ge K$. *Any quadratic optimal quantizer $x = (x_1^*, ..., x_K^*) \in (\mathbb{R}^d)^K$ of X is stationary in the following sense,*

$$\mathbb{E}\,(X \mid \widehat{X}^{x^*}) = \widehat{X}^{x^*},$$

where $\widehat{X}^{x^*}$ is defined in (4.0.15).

(*b*) *(Non-asymptotic Zador's theorem) For every* $\nu \in \mathcal{P}_{2+\varepsilon}(\mathbb{R}^d)$ *with* $\varepsilon > 0$ *and for every quantization level* K*, there exists a constant* $C_{d,\varepsilon} \in (0, +\infty)$ *which depends on* d *and* ε *such that*

$$e_{K,\nu}(y^*) \leq C_{d,\varepsilon} \cdot \sigma_{2+\varepsilon}(\nu) K^{-1/d}, \tag{4.0.17}$$

where y^* *is an optimal quantizer of* ν *and for* $r \in (0, +\infty)$*,*

$$\sigma_r(\nu) := \min_{a \in \mathbb{R}^d} \Big[\int_{\mathbb{R}^d} |\xi - a|^r \, \nu(d\xi) \Big]^{1/r}.$$

(*c*) *(Consistency of the optimal quantization) If* $\nu_n \in \mathcal{P}_2(\mathbb{R}^d), n \in \mathbb{N}^* \cup \{\infty\}$*, such that* $\mathcal{W}_2(\nu_n, \nu_\infty) \xrightarrow{n \to +\infty} 0$ *and* $\operatorname{card}\big(\operatorname{supp}(\nu_n)\big) \geq K$*,* $n \in \mathbb{N}^* \cup \{\infty\}$*, then any limiting point of* K*-level quadratic optimal quantizer* $y^{(n)}$ *of* ν_n *is an optimal quantizer of* ν_∞*, and*

$$\mathcal{D}_{K,\nu_\infty}(y^{(n)}) - \inf_{y \in (\mathbb{R}^d)^K} \mathcal{D}_{K,\nu_\infty}(y) \leq 4 e^*_{K,\nu_\infty} \mathcal{W}_2(\nu_n, \nu_\infty) + 4 \mathcal{W}_2^2(\nu_n, \nu_\infty), \tag{4.0.18}$$

where e^*_{K,ν_∞} *is the optimal quadratic quantization error for* ν_∞ *at level* K*.*

We refer to Pagès (2018)[Proposition 5.1] for the proof of Proposition 4.0.1-(*a*), to Luschgy and Pagès (2008) and Pagès (2018)[Theorem 5.2] for the proof of (*b*) and refer to Liu and Pagès (2018) for the proof of (*c*).

Quadratic optimal quantizer can be computed by several numerical methods, for example the CLVQ algorithm and the Lloyd I algorithm presented in the introduction of this thesis (Section 1.1.3.2). In Chapter 7, we will use Lloyd I algorithm to find the optimal quantizer, but we could also use the CLVQ algorithm as well.

The idea of applying the optimal quantization method to the simulation of the McKean-Vlasov equation was firstly introduced in Gobet et al. (2005)[Section 4] in a slightly different framework. Let $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$, $m = 1, ..., M$ be the quantizer of $\bar{X}_{t_m}$ in the m-th Euler step. The quantization based Euler scheme of the McKean-Vlasov equation (A) is defined by

$$(E): \quad \begin{cases} \widetilde{X}_0 = X_0, \quad \widehat{X}_0 = \operatorname{Proj}_{x^{(0)}}(\widetilde{X}_0) \\ \widetilde{X}_{t_{m+1}} = \widehat{X}_{t_m} + h \cdot b(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) + \sqrt{h}\, \sigma(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) Z_{m+1}, \quad m = 0, ..., M-1 \\ \text{where } h = \frac{T}{M} \text{ and } \widehat{\mu}_{t_m} = P_{\widehat{X}_{t_m}} \\ \widehat{X}_{t_{m+1}} = \operatorname{Proj}_{x^{(m+1)}}(\widetilde{X}_{t_{m+1}}), \end{cases}.$$

Such quantization based Euler schemes have been introduced in Pagès and Sagna (2015) for standard Brownian diffusions. They also appear in a somewhat hidden way in Pages

et al. (2004) and Gobet et al. (2005). Same as the theoretical Euler scheme, (E) does not indicate how to explicitly express $\widehat{\mu}_{t_m}$, so we call (E) the *theoretical quantization procedure*. We propose the following solutions to explicitly express $\widehat{\mu}_{t_m}$.

(1) In the Vlasov case (4.0.3), we can use the recursive quantization method, which is firstly introduced in Pagès and Sagna (2015) and Gobet et al. (2006) for the stochastic differential equation $dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t$. By the recursive quantization method, we obtain the Markovian transitions of $(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})$ based on the quantized scheme (E). Let $p^{(m)} = (p_1^{(m)}, ..., p_K^{(m)})$ denote the corresponding weight of quantizer $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$. Thus $\widehat{\mu}_{t_m} = \sum_{k=1}^K \delta_{x_k^{(m)}} p_k^{(m)}$. The Markovian transition of $(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})$ that we propose in Section 7.3 can be written as (with an obvious slight abuse of notation)

$$\begin{aligned}
&\mathbb{P}\big(\widehat{X}_{t_{m+1}} = x_j^{(m+1)} \mid \widehat{X}_{t_m} = x_i^{(m)}, p^{(m)}\big) \\
&\quad = \mathbb{P}\Big[\big(x_i^{(m)} + h\sum_{k=1}^K p_k^{(m)} \beta(x_i^{(m)}, x_k^{(m)}) + \sqrt{h}\sum_{k=1}^K p_k^{(m)} a(x_i^{(m)}, x_k^{(m)}) Z_{m+1}\big) \in C_j(x^{(m+1)})\Big]
\end{aligned}$$

so that given $p^{(m)}$, we can compute $p_j^{(m+1)}$ for every $j = 1, ..., K$ by

$$\begin{aligned}
p_j^{(m+1)} &= \mathbb{P}\big(\widehat{X}_{t_{m+1}} = x_j^{(m+1)} \mid p^{(m)}\big) \\
&= \sum_{i=1}^K \mathbb{P}\big(\widehat{X}_{t_{m+1}} = x_j^{(m+1)} \mid \widehat{X}_{t_m} = x_i^{(m)}, p^{(m)}\big) \cdot \mathbb{P}(\widehat{X}_{t_m} = x_i^{(m)}).
\end{aligned}$$

A proof of the above equalities is provided in Section 7.3, where we also explain in the same section how to combine this scheme with the Lloyd I algorithm to optimize the quantizer $x^{(m)}$ at each time step, as proposed in Pagès and Sagna (2015).

(2) The second solution to simulate $\widehat{\mu}_{t_m}$ is to use the optimal quantizer of the normal distribution $\mathcal{N}(0, \mathbf{I}_q)$ and its weight, which can be downloaded from the website

$$\text{www.quantize.maths} - \text{fi.com/gaussian}\_\text{database}$$

for dimensions $q = 1, ..., 10$. Let $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$ denote the quantizer of $\bar{X}_{t_m}$ in m-th Euler step. Let $z = (z_1, ..., z_J)$ be the optimal quantizer of $\mathcal{N}(0, \mathbf{I}_q)$ with $J > K$ and let $w = (w_1, ..., w_J)$ be the corresponding weight vector of the quantizer z. This simulation method by using the optimal quantizer of $\mathcal{N}(0, \mathbf{I}_q)$[(1)], that is,

(1) By a slight abus of notation, we use here the same notation as in (E).

replacing Z_{m+1} by $\widehat{Z}^z_{m+1}$, reads

$$(H): \quad \begin{cases} \widetilde{X}_0 = X_0, \ \ \widehat{X}_0 = \mathrm{Proj}_{x^{(0)}}(\widetilde{X}_0) \\ \widetilde{X}_{t_{m+1}} = \widehat{X}_{t_m} + h \cdot b(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) + \sqrt{h}\,\sigma(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})\widehat{Z}^z_{m+1}, \ \ m = 0, ..., M-1 \\ \text{where } h = \frac{T}{M} \ \text{ and } \ \widehat{\mu}_{t_m} = P_{\widehat{X}_{t_m}} \\ \widehat{X}_{t_{m+1}} = \mathrm{Proj}_{x^{(m+1)}}(\widetilde{X}_{t_{m+1}}), \end{cases},$$

where $\widehat{Z}^z_m \overset{i.i.d}{\sim} \sum_{j=1}^J \delta_{z_j} w_j$ and $(\widehat{Z}^z_m)_{m=1,...,M}$ are independent to X_0. This new scheme, denoted by (H), will be called the *doubly quantized scheme*. We will show in Section 7.4 the error analysis of this scheme.

(3) Once we obtain the convergence of $\mathcal{W}_p(\bar{\mu}^N_{t_m}, \bar{\mu}_{t_m})$ in Section 7.1, it follows from Proposition 4.0.1-(c) that we may use the optimal quantizer of $\bar{\mu}^N_{t_m}$ as a quasi-optimal quantizer of $\bar{\mu}_{t_m}$. Let $x^{(m)} = (x^{(m)}_1, ..., x^{(m)}_K)$, $m = 0, 1, ..., M$, be the quantizer for the empirical measure $\bar{\mu}^N_{t_m}$ in (D). We implement the optimal quantization method for the particle system (D) as follows:

$$(F): \quad \begin{cases} \forall n \in \{1, ..., N\}, \\ \widetilde{X}^{n,N}_{t_{m+1}} = \widetilde{X}^{n,N}_{t_m} + h \cdot b(\widetilde{X}^{n,N}_{t_m}, \widehat{\mu}^K_{t_m}) + \sqrt{h}\,\sigma(\widetilde{X}^{n,N}_{t_m}, \widehat{\mu}^K_{t_m}) Z^n_{m+1} \\ \widehat{\mu}^K_{t_m} = \left(\frac{1}{N}\sum_{n=1}^N \delta_{\widetilde{X}^{n,N}_{t_m}}\right) \circ \mathrm{Proj}^{-1}_{x^{(m)}} = \sum_{k=1}^K \left[\delta_{x^{(m)}_k} \cdot \sum_{n=1}^N \mathbb{1}_{V_k(x^{(m)})}(\widetilde{X}^{n,N}_{t_m})\right] \\ \bar{X}^{n,N}_0 \overset{i.i.d}{\sim} X_0, \ \ Z^n_m \overset{i.i.d}{\sim} \mathcal{N}(0, I_q). \end{cases}$$

We call (F) the *hybrid particle-quantization scheme*.

Chapter 7 is displayed as Figure 4.1 in which we also briefly mention the convergence rate of the different methods.

At the end of Chapter 7, we give two examples of simulation where we test the above numerical methods. The first one is the simulation of a one-dimensional Burgers equation introduced in Sznitman (1991) and Bossy and Talay (1997). The solution of this Burgers equation admits a closed form so that we can compare the accuracy of different methods. The second example is 3-dimensional which was firstly introduced and simulated in Baladron et al. (2012) and also simulated in Reis et al. (2018).

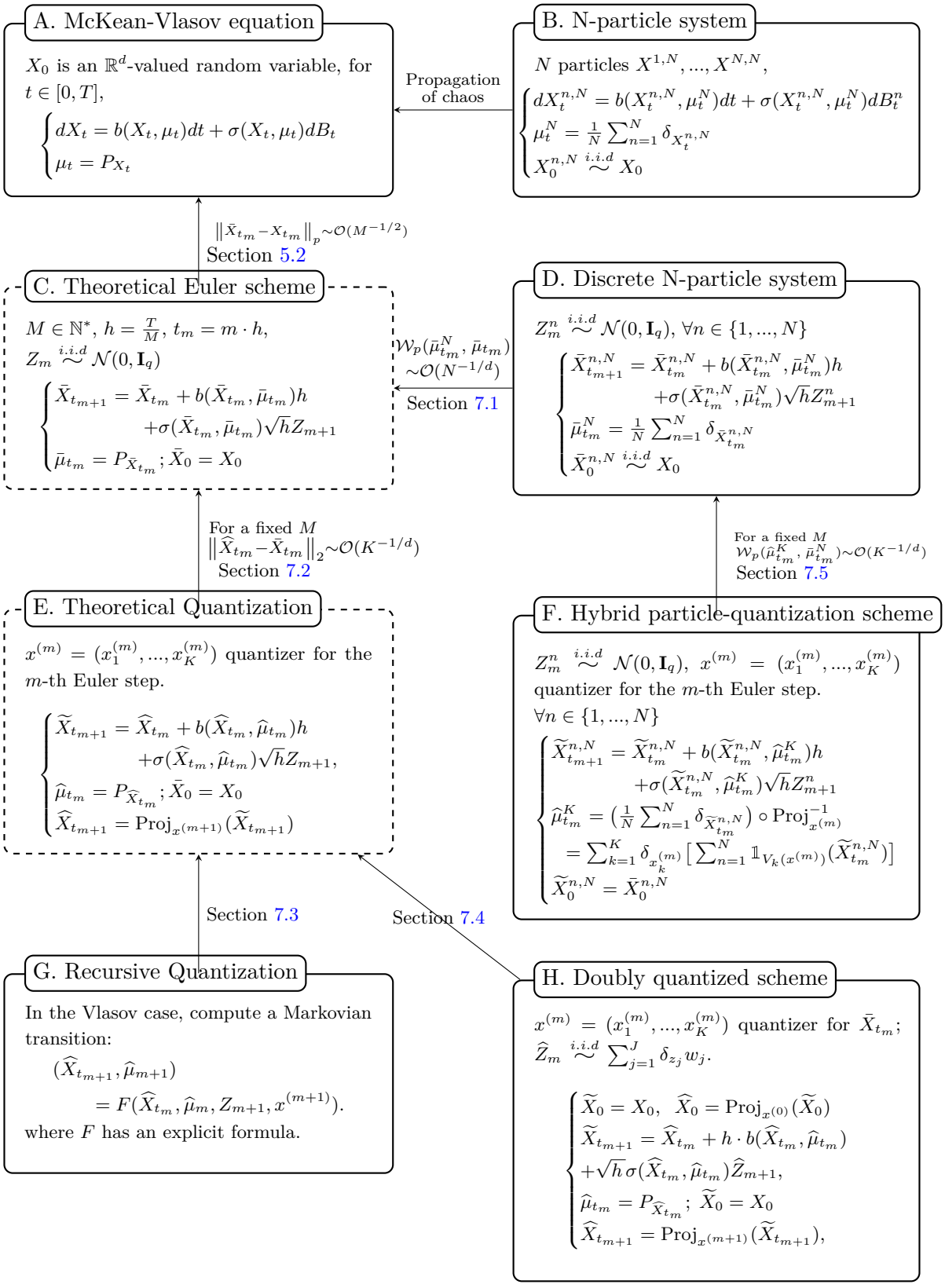


Figure 4.1 Structure of this paper

Main algorithms

4.0.0.1 A review of the Lloyd I algorithm

Lloyd I algorithm, described as follows, is an efficient way to numerically find a quadratic optimal quantizer for a probability distribution $\nu \in \mathcal{P}_2(\mathbb{R}^d)$.

Algorithm 0: Lloyd I algorithm

Set $K \in \mathbb{N}^*$

Input : $y^{[0]} = (y_1^{[0]}, ..., y_K^{[0]})$ such that $y_k^{[0]} \subset \text{supp}(\nu),\ k = 1, ..., K$

repeat

$$y_k^{[l+1]} := \frac{\int_{C_k(y^{[l]})} \xi \nu(d\xi)}{\nu\big(C_k(y^{[l]})\big)}, \quad k = 1, ..., K, \tag{4.0.19}$$

until $\{y_1^{[l+1]}, ..., y_K^{[l+1]}\} = \{y_1^{[l]}, ..., y_K^{[l]}\}$ *or other stopping criterion occurs*

Output : $y^{[l]} = (y_1^{[l]}, ..., y_K^{[l]})$

4.0.0.2 Algorithm based on the particle method (D)

Assume that $b(x, \mu)$ and $\sigma(x, \mu)$ are calculable for a countable sum of weighted dirac measures $\mu = \sum_{i=1}^N p_i \delta_{y_i}$.

Algorithm 1: Particle method

Set $N, M \in \mathbb{N}^*$

begin Euler step 0

Simulate N random variables $X_0^{1,N}, ..., X_0^{N,N} \overset{i.i.d}{\sim} X_0$

repeat

Compute for every $n \in \{1, ..., N\}$,

$$\bar{X}_{t_{m+1}}^{n,N} = \bar{X}_{t_m}^{n,N} + b(\bar{X}_{t_m}^{n,N}, \bar{\mu}_{t_m}^N)h + \sigma(\bar{X}_{t_m}^{n,N}, \bar{\mu}_{t_m}^N)\sqrt{h}Z_{m+1}^n, \tag{4.0.20}$$

where $\bar{\mu}_{t_m}^N := \frac{1}{N}\sum_{n=1}^N \delta_{\bar{X}_{t_m}^{n,N}}$.

until $m + 1 > M$

4.0.0.3 Algorithm based on the recursive quantization method (G)

The algorithm based on the recursive quantization method is:

Algorithm 2: Recursive quantization method-Part 1

Function *Euler*(x, p)**:**

Input : $x = (x_1, ..., x_K) \in (\mathbb{R}^d)^K,\ p = (p_1, ..., p_K) \in [0,1]^K$

Output : $y = (y_1, ..., y_K) \in (\mathbb{R}^d)^K$

Simulate $Z \sim \mathcal{N}(0, \mathbf{I}_q)$

Compute $y_i = x_i + h\sum_{k=1}^{K} \beta(x_i, x_k)p_k + \sqrt{h}\sum_{k=1}^{K} a(x_i, x_k)p_k \cdot Z, i = 1, ..., K.$

Function $f(\xi : m, \Sigma)$**:** `/* density function of` $\mathcal{N}(m, \Sigma^2)$ `*/`

Input : $m = (m_1, ..., m_d) \in \mathbb{R}^d,\ \Sigma \in \mathbb{M}_{d,d}$

Output : function f

$$f(\xi) = \frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}}\exp\Big(-\frac{1}{2}(\xi - m)^\top \Sigma^{-1}(\xi - m)\Big)$$

Function *Transition*(x, p, A)**:**

Input : $x = (x_1, ..., x_K) \in (\mathbb{R}^d)^K,\ p = (p_1, ..., p_K) \in [0,1]^K, A \in \mathcal{B}(\mathbb{R}^d)$

Output : $e = (e_1, ..., e_K) \in (\mathbb{R}^d)^K,\ p = (p_1, ..., p_K) \in [0,1]^K$

Compute $m = (m_1, ..., m_K) \in (\mathbb{R}^d)^K$ and $\Sigma = (\Sigma_1, ..., \Sigma_K) \in (\mathbb{M}_{d,d})^K$ by

$$m_i = x_i + h\sum_{k=1}^{K} \beta(x_i, x_k)p_k,\ \Sigma_i = h\Big[\sum_{k=1}^{K} a(x_i, x_k)p_k\Big]^\top \Big[\sum_{k=1}^{K} a(x_i, x_k)p_k\Big], i = 1, ..., K.$$

Compute $e = (e_1, ..., e_K)$ by $e_i = \sum_{i=1}^{K}\big[\int_A \xi f(\xi : m_i, \Sigma_i)\lambda_d(d\xi)\big]p_i, i = 1, ..., K.$

Compute $p = (p_1, ..., p_K)$ by $p_i = \sum_{i=1}^{K}\big[\int_A f(\xi : m_i, \Sigma_i)\lambda_d(d\xi)\big]p_i, i = 1, ..., K.$

Algorithm 2: Recursive quantization method-Part 2

Set $K, M \in \mathbb{N}^*$

begin Euler step 0

Choose $x^{(0)} = (x_1^{(0)}, ..., x_k^{(0)}) \subset \text{supp}(\mu_0)$[(a)].

begin Lloyd iteration

Define $\Upsilon^{[0]} = (y_1^{[0]}, ..., y_K^{[0]})$ by letting $y_k^{[0]} \leftarrow x_k^{(0)}, k = 1, ..., K$.

repeat

Compute $y_k^{[l+1]} = \dfrac{\int_{C_k(\Upsilon^{[l]})} \xi \mu_0(d\xi)}{\mu_0\big(C_k(\Upsilon^{[l]})\big)}, k = 1, ..., K.$

until $\Upsilon^{[l+1]} = \Upsilon^{[l]}$ *or some stopping criterion occurs*

Set $x^{(0)} = (x_1^{(0)}, ..., x_k^{(0)}) \leftarrow (y_1^{[l]}, ..., y_K^{[l]})$,

Compute $p_k^{(0)} = \mu_0\big(C_k(x^{(0)})\big), k = 1, ..., K.$

Euler step $m \longrightarrow$ Euler step $m+1$:

repeat

Input : $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)}) \in (\mathbb{R}^d)^K$, $p^{(m)} = (p_1^{(m)}, ..., p_K^{(m)}) \in [0,1]^K$

Compute $x^{(m+1)} = (x_1^{(m+1)}, ..., x_k^{(m+1)})$ by $x_k^{(m+1)} = Euler(x^{(m)}, p^{(m)})[k]$.

begin Lloyd iteration

Define $\Upsilon^{[0]} = (y_1^{[0]}, ..., y_K^{[0]})$ by letting $y_k^{[0]} \leftarrow x_k^{(m+1)}, k = 1, ..., K$.

repeat

Compute $y_k^{[l+1]} = \dfrac{Transition\big(x^{(m)}, p^{(m)}, C_k(y^{[l]})\big)[e]}{Transition\big(x^{(m)}, p^{(m)}, C_k(y^{[l]})\big)[p]}, \ k = 1, ..., K.$

until $\Upsilon^{[l+1]} = \Upsilon^{[l]}$ *or some stopping criterion occurs*

(b)

Set $x^{(m+1)} = (x_1^{(m+1)}, ..., x_k^{(m+1)}) \leftarrow (y_1^{[l]}, ..., y_K^{[l]})$,

Compute $p_k^{(m+1)} = Transition\big(x^{(m)}, p^{(m)}, C_k(x^{(m+1)})\big)[p], k = 1, ..., K.$

Output : $x^{(m+1)} = (x_1^{(m+1)}, ..., x_k^{(m+1)})$, $p^{(m+1)} = (p_1^{(m+1)}, ..., p_K^{(m+1)})$

until $m + 1 > M$

(a) $x^{(0)}$ can be obtained by sampling K random variables with the probability distribution μ_0 or the *self-quantization* method.

(b) In the Lloyd iteration, we need to compute the integral of the density function $f(\xi)$ and $\xi \cdot f(\xi)$ over a Voronoï cell. In dimension 1, there exists a close formula to compute them (see further (7.3.8)). In low dimension, we recommend the package *Qhull* (*http://www.qhull.org*) or package *pysdot* (*https://github.com/sd-ot/pysdot).* In high dimension, we recommend to use other algorithms proposed in this chapter.

4.0.0.4 Algorithm based on the doubly quantized scheme (H)

Assume that $b(x,\mu)$ and $\sigma(x,\mu)$ are calculable for a countable sum of weighted dirac measures $\mu=\sum_{i=1}^{N} p_i\delta_{y_i}$. Assume that we have already the optimal quantizer $z=(z_1,...,z_J)$ of $\mathcal{N}(0,\mathbf{I}_q)$ and its corresponding weight $w=(w_1,...,w_J)$ with J large enough.

Algorithm 3: Doubly quantized scheme

Function $f(x,\mu,z)$**:**

Input : $x\in\mathbb{R}^d,\ \mu=\sum_{k=1}^{K}\delta_{x_k}w_k,\ z\in\mathbb{R}^q$.

Output : $x+h\cdot b(x,\mu)+\sqrt{h}\sigma(x,\mu)z$

Set $K,M\in\mathbb{N}^*$

begin Euler step 0

Choose $x^{(0)}=(x_1^{(0)},...,x_k^{(0)})\subset \mathrm{supp}(\mu_0)$.

begin Lloyd iteration

Define $\Upsilon^{[0]}=(y_1^{[0]},...,y_K^{[0]})$ by letting $y_k^{[0]}\leftarrow x_k^{(0)}, k=1,...,K$.

repeat

Compute $y_k^{[l+1]}=\dfrac{\int_{C_k(\Upsilon^{[l]})}\xi\mu_0(d\xi)}{\mu_0\big(C_k(\Upsilon^{[l]})\big)}, k=1,...,K.$

until $\Upsilon^{[l+1]}=\Upsilon^{[l]}$ *or some stopping criterion occurs*

Set $x^{(0)}=(x_1^{(0)},...,x_k^{(0)})\leftarrow(y_1^{[l]},...,y_K^{[l]})$,

Compute $p_k^{(0)}=\mu_0\big(C_k(x^{(0)})\big), k=1,...,K.$

Euler step $m\longrightarrow$ Euler step $m+1$:

repeat

Input : $x^{(m)}=(x_1^{(m)},...,x_K^{(m)})\in(\mathbb{R}^d)^K,\ p^{(m)}=(p_1^{(m)},...,p_K^{(m)})\in[0,1]^K$.

Thus $\widehat{\mu}_{t_m}=\sum_{k=1}^{K}\delta_{x_k^{(m)}}p_k^{(m)}$.

Compute $f_{kj}=f(x_k^{(m)},\widehat{\mu}_{t_m},z_j),\ k=1,...,K,\ j=1,...,J.$

begin Lloyd iteration

Define $\Upsilon^{[0]}=(y_1^{[0]},...,y_K^{[0]})$ by letting $y_k^{[0]}\leftarrow x_k^{(m+1)}, k=1,...,K$.

repeat

Compute $y_i^{[l+1]}=\dfrac{\sum_{kj}p_k^{(m)}w_jf_{kj}\mathbb{1}_{\{f_{kj}\in C_i(y^{[l]})\}}}{\sum_{kj}p_k^{(m)}w_j\mathbb{1}_{\{f_{kj}\in C_i(y^{[l]})\}}},\ i=1,...,K.$

until $\Upsilon^{[l+1]}=\Upsilon^{[l]}$ *or some stopping criterion occurs*

Set $x^{(m+1)}=(x_1^{(m+1)},...,x_k^{(m+1)})\leftarrow(y_1^{[l]},...,y_K^{[l]})$,

Set $p_k^{(m+1)}=\sum_{kj}p_k^{(m)}w_j\mathbb{1}_{\{f_{kj}\in C_k(y^{[l]})\}}, k=1,...,K.$

Output : $x^{(m+1)}=(x_1^{(m+1)},...,x_k^{(m+1)}),\ p^{(m+1)}=(p_1^{(m+1)},...,p_K^{(m+1)})$.

until $m+1>M$

4.0.0.5 Algorithm based on the hybrid particle-quantization scheme (F)

Assume that $b(x,\mu)$ and $\sigma(x,\mu)$ are calculable for a countable sum of weighted dirac measures $\mu=\sum_{i=1}^{N} p_i\delta_{y_i}$. The algorithm based on the hybrid scheme (F) is:

Algorithm 4: Hybrid particle-quantization scheme

Set $K, M, N \in \mathbb{N}^*$ with $K \leq N$.

begin Euler step 0

Simulate $X_0^{1,N}, ..., X_0^{N,N} \overset{i.i.d}{\sim} X_0$.

Choose $x^{(0)} = (x_1^{(0)}, ..., x_k^{(0)}) \subset \text{supp}(P_{X_0})$.

begin Lloyd iteration

Define $\Upsilon^{[0]} = (y_1^{[0]}, ..., y_K^{[0]})$ by letting $y_k^{[0]} \leftarrow x_k^{(0)}, k = 1, ..., K$.

repeat

Compute $y_k^{[l+1]} = \dfrac{\sum_{n=1}^{N} X_0^{n,N} \mathbb{1}_{C_k(y^{[l]})}(X_0^{n,N})}{\sum_{n=1}^{N} \mathbb{1}_{C_k(y^{[l]})}(X_0^{n,N})}, k = 1, ..., K.$

until $\Upsilon^{[l+1]} = \Upsilon^{[l]}$ *or some stopping criterion occurs*

Set $x^{(0)} = (x_1^{(0)}, ..., x_k^{(0)}) \leftarrow (y_1^{[l]}, ..., y_K^{[l]})$,

Compute $p_k^{(0)} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{C_k(y^{[l]})}(X_0^{n,N}), k = 1, ..., K$.

Euler step $m \longrightarrow$ Euler step $m+1$:

repeat

Input : $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)}) \in (\mathbb{R}^d)^K$, $p^{(m)} = (p_1^{(m)}, ..., p_K^{(m)}) \in [0,1]^K$

Simulate N-particle $X_{m+1}^{1,N}, ..., X_{m+1}^{N,N}$ by

$$X_{m+1}^{n,N} = X_m^{n,N} + h \cdot b\Big(X_m^{n,N}, \sum_{k=1}^{K} \delta_{x_k^{(m)}} p_k^{(m)}\Big) + \sqrt{h} \cdot \sigma\Big(X_m^{n,N}, \sum_{k=1}^{K} \delta_{x_k^{(m)}} p_k^{(m)}\Big) Z_{m+1},\ n = 1, ..., N.$$

Compute the initial quantizer $x^{(m+1)} = (x_1^{(m+1)}, ..., x_k^{(m+1)})$ by

$$x_j^{(m+1)} = x_j^{(m)} + h \cdot b\Big(x_j^{(m)}, \sum_{k=1}^{K} \delta_{x_k^{(m)}} p_k^{(m)}\Big) + \sqrt{h} \cdot \sigma\Big(x_j^{(m)}, \sum_{k=1}^{K} \delta_{x_k^{(m)}} p_k^{(m)}\Big) Z_{m+1}, j = 1, ..., K.$$

begin Lloyd iteration

Define $\Upsilon^{[0]} = (y_1^{[0]}, ..., y_K^{[0]})$ by letting $y_k^{[0]} \leftarrow x_k^{(m+1)}, k = 1, ..., K$.

repeat

Compute $y_k^{[l+1]} = \dfrac{\sum_{n=1}^{N} X_{m+1}^{n,N} \mathbb{1}_{C_k(y^{[l]})}(X_{m+1}^{n,N})}{\sum_{n=1}^{N} \mathbb{1}_{C_k(y^{[l]})}(X_{m+1}^{n,N})}, k = 1, ..., K.$

until $\Upsilon^{[l+1]} = \Upsilon^{[l]}$ *or some stopping criterion occurs*

Set $x^{(m+1)} = (x_1^{(m+1)}, ..., x_k^{(m+1)}) \leftarrow (y_1^{[l]}, ..., y_K^{[l]})$,

Compute $p_k^{(m+1)} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{C_k(y^{[l]})}(X_{m+1}^{n,N}), k = 1, ..., K$.

until $m+1 > M$

Frequently used notation

$(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$	filtered probability space
$(E, \|\cdot\|_E)$	Banach space with norm $\|\cdot\|_E$
P_X	the probability distribution of the random variable X, i.e. $P_X = \mathbb{P} \circ X^{-1}$
$\|\cdot\|_p$	L^p-norm of the random variable
$\|\cdot\|$	norm on $\mathbb{R}^d$, Euclidean norm in Section 7.2-7.5
$(B_t)_{t\geq 0}$	$\mathcal{F}_t$-standard Brownian motion, valued in $\mathbb{R}^q$
$\mathbb{M}_{d,q}(\mathbb{R})$	set of matrices with size $d \times q$
$\|\|\|\cdot\|\|\|$	norm on $\mathbb{M}_{d,q}(\mathbb{R})$, defined by $\|\|\|A\|\|\| \coloneqq \sup_{\|z\|_q \leq 1} \|Az\|$
δ_x	Dirac measure on x
$\mathcal{P}(E)$	set of probability distributions on E
$\mathcal{P}_p(E)$	set of probability distributions on E with p-th finite moment
$\mathcal{W}_p$	Wasserstein distance on $\mathcal{P}_p(\mathbb{R}^d)$
L	Lipschitz constant in Assumption (I)
$\mathbf{I}_q$	$q \times q$ identity matrix
$\mathcal{N}(0, \mathbf{I}_q)$	$\mathbb{R}^q$-standard normal distribution
card	cardinality
$\text{supp}(\mu)$	support of a probability distribution μ
$V_k(x)$	Voronoï cell generated by $x \in (\mathbb{R}^d)^K$, defined in (4.0.12)
$(C_k(x))_{1\leq k\leq K}$	Voronoï partition generated by $x \in (\mathbb{R}^d)^K$, defined in (4.0.13)
$e_{K,\nu}$	quadratic quantization error function, defined in (4.0.10)
$\mathcal{D}_{K,\nu}$	quadratic distortion function, $\mathcal{D}_{K,\nu} = e_{K,\nu}^2$
Proj_x	projection function on x, defined in (4.0.14)
$\mathcal{C}([0,T], \mathbb{R}^d)$	the space of $\mathbb{R}^d$-valued continuous applications defined on $[0,T]$
$\|\cdot\|_{\sup}$	sup norm on $\mathcal{C}([0,T], \mathbb{R}^d)$, defined by $\|\alpha\|_{\sup} = \sup_{t\in[0,T]} \|\alpha_t\|$
$L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega, \mathcal{F}, \mathbb{P})$	L^p-space of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and valued in $\mathcal{C}([0,T], \mathbb{R}^d)$
$\|\cdot\|_{p,C,T}$	norm on $L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega, \mathcal{F}, \mathbb{P})$, defined in (5.1.1)
$\mathcal{H}_{p,C,T}$	space of $\mathcal{F}_t$-adapted process in $L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega, \mathcal{F}, \mathbb{P})$
$\mathcal{P}_p\big(\mathcal{C}([0,T], \mathbb{R}^d)\big)$	probability distribution μ on $\mathcal{C}([0,T], \mathbb{R}^d)$ s.t. $\int_{\mathcal{C}([0,T],\mathbb{R}^d)} \|\xi\|^p_{\sup} \mu(d\xi) < +\infty$
$\mathbb{W}_p$	Wasserstein distance on $\mathcal{P}_p\big(\mathcal{C}([0,T], \mathbb{R}^d)\big)$
$\Pi(\mu, \nu)$	set of all probability distribution with marginals μ and ν
$\mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big)$	$(\mu_t)_{t\in[0,T]}$ s.t. $t \mapsto \mu_t$ is continuous, and $\mu_t \in \mathcal{P}(\mathbb{R}^d)$ for every $t \in [0,T]$
$d_{\mathcal{C}}$	distance on $\mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big)$, defined in (5.1.5)
π_t	marginal projection on $\mathcal{C}([0,T], \mathbb{R}^d) \to \mathbb{R}^t$: $\alpha \mapsto \pi_t(\alpha) = \alpha_t$
$\mathbb{W}_{p,t}$	truncated Wasserstein distance defined in (5.1.6)

$d_{\mathcal{H}\times\mathcal{P}}$	distance on $\mathcal{H}_{p,C,T} \times \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big)$ defined in (5.1.7)
ι	application defined on $\mathcal{P}_p\big(\mathcal{C}([0,T], \mathbb{R}^d)\big) \to \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big)$ by $\mu \mapsto \iota(\mu) \coloneqq (\mu \circ \pi_t^{-1})_{t\in[0,T]} = (\mu_t)_{t\in[0,T]}$
$\perp\!\!\!\perp$	independence of two random variables
$\preceq_{cv}$	convex order between two random variables or two probability distributions, see Definition 6.0.1
$\preceq$	partial matrix order in $\mathbb{M}_{d\times q}$, defined in (6.0.3)

Chapter 5

Existence and Uniqueness of a Strong Solution of the McKean-Vlasov Equation, Convergence of the Theoretical Euler Scheme

In this chapter, we first discuss the existence and uniqueness of a strong solution of the McKean-Vlasov equation

$$\begin{cases} dX_t = b(t, X_t, \mu_t)dt + \sigma(t, X_t, \mu_t)dB_t, \\ X_0 : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right) \text{ random variable, } X_0 \perp\!\!\!\perp (B_t)_{t\in[0,T]}, \\ \forall t \geq 0, \ \mu_t \text{ denotes the probability distribution of } X_t, \end{cases} \tag{5.0.1}$$

under Assumption (I). Furthermore, in Section 5.2, we establish the L^p-convergence rate of its theoretical Euler scheme:

$$\begin{cases} \bar{X}_{t_{m+1}} = \bar{X}_{t_m} + h \cdot b(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m}) + \sqrt{h}\,\sigma(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m})Z_{m+1}, \\ \bar{\mu}_{t_m} = P_{\bar{X}_{t_m}}, Z_m \overset{i.i.d}{\sim} \mathcal{N}(0, \mathbf{I}_q), \\ \bar{X}_0 = X_0, \end{cases} \tag{5.0.2}$$

where $M \in \mathbb{N}$ is the chosen number of Euler steps, $h = \frac{T}{M}$ and $t_m = m \cdot h = m \cdot \frac{T}{M}$, $m = 0, ..., M$.

5.1 Existence, uniqueness and properties of a strong solution of the McKean-Vlasov equation under Lipschitz condition

Let $\big(\mathcal{C}\big([0,T],\mathbb{R}^d\big),\|\cdot\|_{\sup}\big)$ denote the space of $\mathbb{R}^d$-valued continuous applications defined on $[0,T]$, equipped with the uniform norm $\|x\|_{\sup} := \sup_{t\in[0,T]}|x_t|$. Let $L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega,\mathcal{F},\mathbb{P})$ denote the space of $\mathcal{C}([0,T],\mathbb{R}^d)$-valued random variable $Y=(Y_t)_{t\in[0,T]}$ having an L^p-norm $\|Y\|_p := \big[\mathbb{E}\,\|Y\|^p_{\sup}\big]^{1/p} = \big[\mathbb{E}\sup_{t\in[0,T]}|Y_t|^p\big]^{1/p} < +\infty$. For a fixed constant $C>0$, we define another norm $\|\cdot\|_{p,C,T}$ on $L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega,\mathcal{F},\mathbb{P})$ by

$$\|Y\|_{p,C,T} = \sup_{t\in[0,T]} e^{-Ct}\left\|\sup_{0\le s\le t}|Y_s|\right\|_p. \tag{5.1.1}$$

It is obvious that $\|\cdot\|_{p,C,T}$ and $\|\cdot\|_p$ are equivalent since

$$\forall\, Y\in L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega,\mathcal{F},\mathbb{P}),\quad e^{-CT}\|Y\|_p \le \|Y\|_{p,C,T} \le \|Y\|_p. \tag{5.1.2}$$

We define

$$\mathcal{H}_{p,C,T} := \big\{Y\in L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\in[0,T]},\mathbb{P}) \text{ s.t. } Y \text{ is } \mathcal{F}_t-\text{adapted.}\big\} \tag{5.1.3}$$

Lemma 5.1.1. *The space $\mathcal{H}_{p,C,T}$ equipped with $\|\cdot\|_{p,C,T}$ is a complete space.*

Proof. The space $\big(L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega,\mathcal{F},\mathbb{P}),\|\cdot\|_p\big)$ is a complete space. Moreover, it follows from (5.1.2) that $\|\cdot\|_p$ and $\|\cdot\|_{p,C,T}$ are equivalent. Thus for any Cauchy sequence

$$X^{(n)}\in\mathcal{H}_{p,C,T}\subset L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega,\mathcal{F},\mathbb{P}),$$

there exists $X^{(\infty)}\in L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega,\mathcal{F},\mathbb{P})$ such that

$$\left\|X^{(n)}-X^{(\infty)}\right\|_p \xrightarrow{n\to+\infty} 0,$$

which directly implies

$$\left\|X^{(n)}-X^{(\infty)}\right\|_{p,C,T} \le \left\|X^{(n)}-X^{(\infty)}\right\|_p \xrightarrow{n\to+\infty} 0$$

and $\left\|X^{(\infty)}\right\|_{p,C,T} \le \left\|X^{(\infty)}\right\|_p \le \liminf_n \left\|X^{(n)}\right\|_p < +\infty$ owing to Fatou's Lemma.

The fact that $\left\|X^{(n)}-X^{(\infty)}\right\|_p \xrightarrow{n\to+\infty} 0$ implies also that there exists a subsequence

$X^{\varphi(n)}$ such that

$$\left\|X^{\varphi(n)}(\omega) - X^{(\infty)}(\omega)\right\|_{\sup} \to 0 \quad a.s..$$

Thus there exists $\Omega_0 \subset \Omega$, $\Omega_0 \in \mathcal{F}$ with $\mathbb{P}(\Omega_0) = 1$ such that for every $\omega \in \Omega_0$,

$$\left\|X^{\varphi(n)}(\omega) - X^{(\infty)}(\omega)\right\|_{\sup} \to 0$$

and for every $\omega \in \Omega \backslash \Omega_0$, we can arbitrarily change the definition of $X^{(\infty)}(\omega)$. For example, for every $\omega \in \Omega \backslash \Omega_0$, set $X^{(\infty)}(\omega) = 0$. Thus, for any $t \in [0,T]$,

$$X_t^{(\infty)}(\omega) = \begin{cases} \lim X_t^{\varphi(n)}(\omega), & \omega \in \Omega_0 \\ 0, & \omega \in \Omega \backslash \Omega_0 \end{cases}.$$

This implies that $X^{(\infty)}$ is $(\mathcal{F}_t)-$adapted. Consequently, $X^{(\infty)} \in \mathcal{H}_{p,C,T}$ and the space $\left(\mathcal{H}_{p,C,T}, \|\cdot\|_{p,C,T}\right)$ is a Banach space. □

For any random variable $Y \in L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega, \mathcal{F}, \mathbb{P})$, its probability distribution P_Y naturally lies in

$$\begin{aligned}&\mathcal{P}_p\big(\mathcal{C}([0,T],\mathbb{R}^d)\big)\\ &\quad := \left\{\mu \text{ probability distribution on } \mathcal{C}([0,T],\mathbb{R}^d) \text{ s.t. } \int_{\mathcal{C}([0,T],\mathbb{R}^d)} \|\alpha\|^p_{\sup}\, \mu(d\alpha) < +\infty\right\}.\end{aligned}$$

We also define an *L^p-Wasserstein distance* $\mathbb{W}_p$ on $\mathcal{P}_p\big(\mathcal{C}([0,T],\mathbb{R}^d)\big)$ by

$$\begin{aligned}&\forall \mu, \nu \in \mathcal{P}_p\big(\mathcal{C}([0,T],\mathbb{R}^d)\big),\\ &\qquad \mathbb{W}_p(\mu,\nu) := \Big[\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{C}([0,T],\mathbb{R}^d)\times\mathcal{C}([0,T],\mathbb{R}^d)} \|x-y\|^p_{\sup}\, \pi(dx,dy)\Big]^{\frac{1}{p}}, \qquad (5.1.4)\end{aligned}$$

where $\Pi(\mu,\nu)$ denote the set of probability measures on $\mathcal{C}([0,T],\mathbb{R}^d) \times \mathcal{C}([0,T],\mathbb{R}^d)$ with respective marginals μ and ν. The space $\mathcal{P}_p\big(\mathcal{C}([0,T],\mathbb{R}^d)\big)$ equipped with $\mathbb{W}_p$ is complete and separable since $\left(\mathcal{C}([0,T],\mathbb{R}^d), \|\cdot\|_{\sup}\right)$ is a Polish space (see Bolley (2008)).

Let us consider now

$$\begin{aligned}&\mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big)\\ &\quad := \left\{(\mu_t)_{t\in[0,T]} \text{ s.t. } t \mapsto \mu_t \text{ is a continuous application from } [0,T] \text{ to } \left(\mathcal{P}_p(\mathbb{R}^d), \mathcal{W}_p\right)\right\}\end{aligned}$$

equipped with the distance

$$d_{\mathcal{C}}\big((\mu_t)_{t\in[0,T]}, (\nu_t)_{t\in[0,T]}\big) := \sup_{t\in[0,T]} \mathcal{W}_p(\mu_t, \nu_t). \qquad (5.1.5)$$

As $\big(\mathcal{P}_p(\mathbb{R}^d), \mathcal{W}_p\big)$ is a complete space (see Bolley (2008)), $\mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big)$ equipped with the uniform distance $d_{\mathcal{C}}$ is also a complete space.

For any $t \in [0,T]$, we define $\pi_t : \mathcal{C}([0,T], \mathbb{R}^d) \to \mathbb{R}^d$ by $\alpha \mapsto \pi_t(\alpha) = \alpha_t$.

Lemma 5.1.2. *The application* $\iota : \mathcal{P}_p\big(\mathcal{C}([0,T], \mathbb{R}^d)\big) \to \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big)$ *defined by*

$$\mu \mapsto \iota(\mu) = (\mu \circ \pi_t^{-1})_{t\in[0,T]} = (\mu_t)_{t\in[0,T]}$$

is well-defined.

Proof. For any $\mu \in \mathcal{P}_p\big(\mathcal{C}([0,T], \mathbb{R}^d)\big)$, there exists $X : (\Omega, \mathcal{F}, \mathbb{R}) \to \mathcal{C}([0,T], \mathbb{R}^d)$ such that $P_X = \mu$ and $\mathbb{E}\,\|X\|_{\sup}^p < +\infty$ so that $\sup_{t\in[0,T]} \mathbb{E}\,|X_t|^p < +\infty$. Hence, for any $t \in [0,T]$, we have $\mu_t \in \mathcal{P}_p(\mathbb{R}^d)$.

For a fixed $t \in [0,T]$, choose $(t_n)_{n\in\mathbb{N}^*} \in [0,T]^{\mathbb{N}^*}$ such that $t_n \to t$. Then, for any $\omega \in \Omega$, $X_{t_n}(\omega) \to X_t(\omega)$ since for any $\omega \in \Omega$, $X(\omega)$ has a continuous path. Moreover,

$$\sup_n \|X_{t_n}\|_p \vee \|X_t\|_p \le \left\| \sup_{0\le s\le T} |X_s| \right\|_p < +\infty,$$

Hence, $\|X_{t_n} - X_t\|_p \to 0$ owing to the dominated convergence theorem, which implies that $\mathcal{W}_p(\mu_{t_n}, \mu_t) \to 0$ as $n \to +\infty$, that is, $t \mapsto \mu_t$ is a continuous application. Hence,

$$\iota(\mu) = (\mu_t)_{t\in[0,T]} \in \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big).$$

□

If we have a probability distribution $\mu \in \mathcal{P}_p\big(\mathcal{C}([0,T], \mathbb{R}^d)\big)$, with a slight abuse of notation, we denote directly $(\mu_t)_{t\in[0,T]} \coloneqq \iota(\mu) \in \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big)$. The relation between $d_{\mathcal{C}}$ and $\mathbb{W}_p$ has been introduced by D. Lacker in Lacker (2018). He defines an application $\mathbb{W}_{p,t}$ on $\mathcal{P}_p\big(\mathcal{C}([0,T], \mathbb{R}^d)\big) \times \mathcal{P}_p\big(\mathcal{C}([0,T], \mathbb{R}^d)\big)$, called "truncated Wasserstein distance", by

$$\mathbb{W}_{p,t}(\mu,\nu) \coloneqq \Big[\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{C}([0,T],\mathbb{R}^d)\times\mathcal{C}([0,T],\mathbb{R}^d)} \sup_{s\in[0,t]} |x_s - y_s|^p \, \pi(dx,dy) \Big]^{\frac{1}{p}} \tag{5.1.6}$$

and indicates the relation between $\sup_{s\in[0,t]} \mathcal{W}_p(\mu_s,\nu_s)$ and $\mathbb{W}_{p,t}(\mu,\nu)$ as follows.

Lemma 5.1.3. *For any* $\mu, \nu \in \mathcal{P}_p\big(\mathcal{C}([0,T], \mathbb{R}^d)\big)$, *we have*

$$\forall t \in [0,T], \quad \sup_{s\in[0,t]} \mathcal{W}_p(\mu_s,\nu_s) \le \mathbb{W}_{p,t}(\mu,\nu),$$

where $\mu_s = \mu \circ \pi_s^{-1}$. *In particular, for any* $\mu, \nu \in \mathcal{P}_p\big(\mathcal{C}([0,T],\mathbb{R}^d)\big)$,

$$d_{\mathcal{C}}(\iota(\mu),\iota(\nu)) \leq \mathbb{W}_p(\mu,\nu)$$

and the application ι *is continuous.*

Proof. We consider the canonical space $\Omega = \mathcal{C}([0,T],\mathbb{R}^d) \times \mathcal{C}([0,T],\mathbb{R}^d)$ equipped with the σ-algebra $\mathcal{F}$ generated by the distance

$$d\big((\omega^1,\omega^2),(\alpha^1,\alpha^2)\big) \coloneqq \big\|\omega^1 - \alpha^1\big\|_{\sup} \vee \big\|\omega^2 - \alpha^2\big\|_{\sup}$$

and $\mathbb{P} \in \Pi(\mu,\nu)$ where $\Pi(\mu,\nu)$ is the set of probability measures with marginals μ and ν. For any $\omega = (\omega^1,\omega^2) \in \Omega$, we define the canonical projections $X : \Omega \to \mathcal{C}([0,T],\mathbb{R}^d)$ and $Y : \Omega \to \mathcal{C}([0,T],\mathbb{R}^d)$ by

$$\forall \omega = (\omega^1,\omega^2), \quad \forall t \in [0,T], \qquad X_t(\omega) = \omega_t^1 \text{ and } Y_t(\omega) = \omega_t^2.$$

The couple (X,Y) makes up the canonical process on Ω. Since $\mathbb{P} \in \Pi(\mu,\nu)$, then X has probability distribution μ and Y has probability distribution ν. Moreover, we have

$$\sup_{s\in[0,t]} \mathcal{W}_p^p(\mu_s,\nu_s) \leq \sup_{s\in[0,t]} \mathbb{E}\,|X_s - Y_s|^p \leq \mathbb{E} \sup_{s\in[0,t]} |X_s - Y_s|^p .$$

Then we can choose by the usual arguments $\mathbb{P} \in \Pi(\mu,\nu)$ such that

$$\mathbb{E} \sup_{s\in[0,t]} |X_s - Y_s|^p = \big(\mathbb{W}_{p,t}(\mu_s,\nu_s)\big)^p$$

to conclude the proof. □

We define a distance $d_{\mathcal{H}\times\mathcal{P}}$ on $\mathcal{H}_{p,C,T} \times \mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)$ as follows:

$$\begin{aligned}&\forall \big(X,(\mu_t)_{t\in[0,T]}\big), \big(Y,(\nu_t)_{t\in[0,T]}\big) \in \mathcal{H}_{p,C,T} \times \mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big),\\ &\quad d_{\mathcal{H}\times\mathcal{P}}\Big(\big(X,(\mu_t)_{t\in[0,T]}\big), \big(Y,(\nu_t)_{t\in[0,T]}\big)\Big) = \|X - Y\|_{p,C,T} + \sup_{t\in[0,T]} e^{-Ct}\mathcal{W}_p(\mu_t,\nu_t).\end{aligned} \tag{5.1.7}$$

We define also a distance $d_{p,C,T}$ on $\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)$ as follows:

$$\begin{aligned}&\forall (\mu_t)_{t\in[0,T]}, (\nu_t)_{t\in[0,T]} \in \mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big),\\ &\qquad d_{p,C,T}\big((\mu_t)_{t\in[0,T]},(\nu_t)_{t\in[0,T]}\big) \coloneqq \sup_{t\in[0,T]} e^{-Ct}\mathcal{W}_p(\mu_t,\nu_t).\end{aligned}$$

Lemma 5.1.4. *Both* $\big(\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big),d_{p,C,T}\big)$ *and* $\big(\mathcal{H}_{p,C,T}\times\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big),d_{\mathcal{H}\times\mathcal{P}}\big)$ *are complete metric spaces.*

Proof. The distance $d_{p,C,T}$ and $d_{\mathcal{C}}$ are equivalent since for any $(\mu_t):=(\mu_t)_{t\in[0,T]},(\nu_t):=(\nu_t)_{t\in[0,T]}\in\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)$, we have

$$d_{p,C,T}\big((\mu_t),(\nu_t)\big)\le d_{\mathcal{C}}\big((\mu_t),(\nu_t)\big)\le e^{CT}d_{p,C,T}\big((\mu_t),(\nu_t)\big).$$

Thus $\big(\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big),d_{p,C,T}\big)$ is complete. Moreover, it follows from Lemma 5.1.1 that $\big(\mathcal{H}_{p,C,T}\times\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big),d_{\mathcal{H}\times\mathcal{P}}\big)$ is also a complete metric space as the product of two complete metric spaces. □

Before proving that the McKean-Vlasov equation (5.0.1) has a unique strong solution under Assumption (I), we firstly recall two important technical tools used throughout the proof: the generalized Minkowski Inequality and the Burkölder-Davis-Gundy Inequality. We refer the proof of these two inequalities to Pagès (2018)[Section 7.8] among other references.

Lemma 5.1.5 (The Generalized Minkowski Inequality)**.** *For any (bi-measurable) process* $X=(X_t)_{t\ge0}$*, for every* $p\in[1,\infty)$ *and for every* $T\in[0,+\infty]$*,*

$$\left\|\int_0^T X_t dt\right\|_p\le\int_0^T\|X_t\|_p\,dt.$$

Lemma 5.1.6 (Burkölder-Davis-Gundy Inequality (continuous time))**.** *For every* $p\in(0,+\infty)$*, there exists two real constants* $c_p^{BDG}>0$ *and* $C_p^{BDG}>0$ *such that, for every continuous local martingale* $(X_t)_{t\in[0,T]}$ *null at 0,*

$$c_p^{BDG}\left\|\sqrt{\langle X\rangle_T}\right\|_p\le\left\|\sup_{t\in[0,T]}|X_t|\right\|_p\le C_p^{BDG}\left\|\sqrt{\langle X\rangle_T}\right\|_p.$$

In particular, if (B_t) is an $(\mathcal{F}_t)$-standard Brownian motion and $(H_t)_{t\ge0}$ is an $(\mathcal{F}_t)$-progressively measurable process having values in $\mathbb{M}_{d,q}(\mathbb{R})$ such that $\int_0^T\|H_t\|^2\,dt<+\infty$ $\mathbb{P}-a.s.$, then the d-dimensional local martingale $\int_0^{\cdot}H_s dB_s$ satisfies

$$\left\|\sup_{t\in[0,T]}\left|\int_0^t H_s dB_s\right|\right\|_p\le C_{d,p}^{BDG}\left\|\sqrt{\int_0^T\|H_t\|^2\,dt}\right\|_p. \tag{5.1.8}$$

Now we start to prove the existence and uniqueness of a strong solution of the McKean-Vlasov equation (5.0.1). Firstly, under Assumption (I), the coefficient functions b and σ have the following properties.

Lemma 5.1.7. *Under Assumption (I), we have*

(a) The functions b and σ have a linear growth in the sense that there exists a constant $C_{b,\sigma,L,T}$ depending on b, σ, L and T such that

$$\forall t \in [0,T], \forall x \in \mathbb{R}^d, \forall \mu \in \mathcal{P}_p(\mathbb{R}^d),\ |b(t,x,\mu)| \vee |||\sigma(t,x,\mu)||| \leq C_{b,\sigma,L,T}(1+|x|+\mathcal{W}_p(\mu,\delta_0)),$$

where δ_0 denotes the Dirac mass at $\{0\}$.

(b) For any $\big(X,(\mu_t)_{t\in[0,T]}\big), \big(Y,(\nu_t)_{t\in[0,T]}\big) \in \mathcal{H}_{p,C,T} \times \mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)$ and for any $t \in [0,T]$,

$$\Big\| \sup_{s\in[0,t]} \Big| \int_0^s \big[b(u,X_u,\mu_u) - b(u,Y_u,\nu_u)\big] du \Big| \Big\|_p \leq L \int_0^t \big[\, \|X_u - Y_u\|_p + \|\mathcal{W}_p(\mu_u,\nu_u)\|_p \,\big] du,$$

and

$$\Big\| \sup_{s\in[0,t]} \Big| \int_0^s \big[\sigma(u,X_u,\mu_u) - \sigma(u,Y_u,\nu_u)\big] dB_u \Big| \Big\|_p \leq C_{d,p,L} \Big\{ \int_0^t \big[\, \|X_u - Y_u\|_p^2 + \|\mathcal{W}_p(\mu_u,\nu_u)\|_p^2 \,\big] du \Big\}^{\frac{1}{2}}$$

where $C_{d,p,L}$ is a constant only depending on d, p, L.

Proof. (a) For any $x \in \mathbb{R}^d$ and for any $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, Assumption (I) implies that

$$\forall t \in [0,T],\ \ |b(t,x,\mu)| - |b(t,0,\delta_0)| \leq |b(t,x,\mu) - b(t,0,\delta_0)| \leq L\big(|x| + \mathcal{W}_p(\mu,\delta_0)\big).$$

Hence,

$$|b(t,x,\mu)| \leq |b(t,0,\delta_0)| + L\big(|x| + \mathcal{W}_p(\mu,\delta_0)\big) \leq (|b(t,0,\delta_0)| \vee L)(1 + |x| + \mathcal{W}_p(\mu,\delta_0))$$

Similarly, we have $|||\sigma(t,x,\mu)||| \leq (|||\sigma(t,0,\delta_0)||| \vee L)(1 + |x| + \mathcal{W}_p(\mu,\delta_0))$, so we can take $C_{b,\sigma,L,T} = \sup_{t\in[0,T]} |b(t,0,\delta_0)| \vee \sup_{t\in[0,T]} |||\sigma(t,0,\delta_0)||| \vee L$ to complete the proof.

(b) For any $\big(X,(\mu_t)_{t\in[0,T]}\big), \big(Y,(\nu_t)_{t\in[0,T]}\big) \in \mathcal{H}_{p,C,T} \times \mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)$, for any $t \in [0,T]$, we have

$$\begin{aligned}
&\Big\| \sup_{s\in[0,t]} \Big| \int_0^t \big[b(u,X_u,\mu_u) - b(u,Y_u,\nu_u)\big] du \Big| \Big\|_p \\
&\leq \Big\| \sup_{s\in[0,t]} \int_0^s \big|b(u,X_u,\mu_u) - b(u,Y_u,\nu_u)\big| du \Big\|_p = \Big\| \int_0^t \big|b(u,X_u,\mu_u) - b(u,Y_u,\nu_u)\big| du \Big\|_p \\
&\leq \int_0^t \|b(u,X_u,\mu_u) - b(u,Y_u,\nu_u)\|_p \, du \ \text{ (by Lemma 5.1.5)} \\
&\leq \int_0^t \big\| L\big[\, |X_u - Y_u| + \mathcal{W}_p(\mu_u,\nu_u)\big] \big\|_p \, du \leq L \int_0^t \big[\, \|X_u - Y_u\|_p + \|\mathcal{W}_p(\mu_u,\nu_u)\|_p \,\big] du,
\end{aligned}$$

and

$$
\begin{aligned}
&\Big\| \sup_{s\in[0,t]} \Big| \int_0^s \big[\sigma(u,X_u,\mu_u) - \sigma(u,Y_u,\nu_u)\big] dB_u \Big| \Big\|_p \\
&\quad \le C_{d,p}^{BDG} \Big\| \sqrt{\int_0^t |||\sigma(u,X_u,\mu_u) - \sigma(u,Y_u,\nu_u)|||^2 du} \Big\|_p \quad \text{(by Lemma 5.1.6)} \\
&\quad \le C_{d,p}^{BDG} \Big\| \int_0^t |||\sigma(u,X_u,\mu_u) - \sigma(u,Y_u,\nu_u)|||^2 du \Big\|_{\frac{p}{2}}^{\frac{1}{2}} \\
&\qquad \text{(since } \Big\|\sqrt{U}\Big\|_p = \big[\mathbb{E}U^{\frac{p}{2}}\big]^{\frac{2}{p}\times\frac{1}{2}} = \|U\|_{\frac{p}{2}}^{\frac{1}{2}}, \text{ when } U \ge 0) \\
&\quad \le C_{d,p}^{BDG} \Big[\int_0^t \Big\| |||\sigma(u,X_u,\mu_u) - \sigma(u,Y_u,\nu_u)|||^2 \Big\|_{\frac{p}{2}} du \Big]^{\frac{1}{2}} \\
&\qquad \text{(by Minkowski's inequality, since } p \in [2,+\infty)) \\
&\quad \le C_{d,p}^{BDG} \Big[\int_0^t \big\| |||\sigma(u,X_u,\mu_u) - \sigma(u,Y_u,\nu_u)||| \big\|_p^2 du \Big]^{\frac{1}{2}} \\
&\qquad \big(\text{since } \Big\| |U|^2 \Big\|_{\frac{p}{2}} = \Big[\big(\mathbb{E}\,|U|^p\big)^{\frac{1}{p}} \Big]^2 = \|U\|_p^2 \big) \\
&\quad \le C_{d,p}^{BDG} \Big[\int_0^t \big\| L\big[\,|X_u - Y_u| + \mathcal{W}_p(\mu_u,\nu_u)\big] \big\|_p^2 du \Big]^{\frac{1}{2}} \\
&\qquad \text{(by Assumption (I))} \\
&\quad \le C_{d,p}^{BDG} L \Big[\int_0^t \big[\, \|X_u - Y_u\|_p + \|\mathcal{W}_p(\mu_u,\nu_u)\|_p \,\big]^2 du \Big]^{\frac{1}{2}} \\
&\quad \le \sqrt{2}\, C_{d,p}^{BDG} L \Big[\int_0^t \big[\, \|X_u - Y_u\|_p^2 + \|\mathcal{W}_p(\mu_u,\nu_u)\|_p^2 \,\big] du \Big]^{\frac{1}{2}}.
\end{aligned}
$$

Then we can conclude the proof by letting $C_{d,p,L} = \sqrt{2}\, C_{d,p}^{BDG} L$. □

The idea of our proof follows from Feyel's approach, originally developped for the existence and uniqueness of a strong solution for SDE $dX_t = b(X_t)dt + \sigma(X_t)dB_t$ (see Bouleau (1988)[Section 7]). We define an application Φ_C: $\mathcal{H}_{p,C,T} \times \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big) \to \mathcal{H}_{p,C,T} \times \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big)$[(1)] by

$$
\begin{aligned}
&\forall (Y, (\nu_t)_{t\in[0,T]}) \in \mathcal{H}_{p,C,T} \times \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big), \\
&\Phi_C(Y, (\nu_t)_{t\in[0,T]}) = \Big(\underbrace{\big(X_0 + \int_0^t b(s,Y_s,\nu_s)ds + \int_0^t \sigma(s,Y_s,\nu_s)dB_s\big)_{t\in[0,T]}}_{=:\Phi_C^{(1)}(Y,(\nu_t)_{t\in[0,T]})}, \iota(P_{\Phi_C^{(1)}(Y,(\nu_t)_{t\in[0,T]})}) \Big).
\end{aligned}
$$

The application Φ_C has the following property.

Proposition 5.1.1. *(i) The function Φ_C is well-defined.*

(1) The C in the subscripe of Φ_C is the same constant C as in $(\mathcal{H}_{p,C,T}, \|\cdot\|_{p,C,T})$, the same below.

(*ii*) *Under Assumption (I),* Φ_C *is Lipschitz continuous in the sense that: for any* $\big(X,\iota(P_X)\big)$ *and* $\big(Y,\iota(P_Y)\big)$ *in* $\mathcal{H}_{p,C,T}\times\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)$,

$$d_{\mathcal{H}\times\mathcal{P}}\Big(\Phi_C\big(X,\iota(P_X)\big),\Phi_C\big(Y,\iota(P_Y)\big)\Big)\leq\big(\frac{K_1}{C}+\frac{K_2}{\sqrt{C}}\big)\cdot d_{\mathcal{H}\times\mathcal{P}}\Big(\big(X,\iota(P_X)\big),\big(Y,\iota(P_X)\big)\Big),$$

where K_1, K_2 *are real constants which do not depend on the constant* C.

Proof. (*i*) It follows from Lemma 5.1.2 that for every $X\in\mathcal{H}_{p,C,T}$, $\iota(P_X)\in\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)$.

Let $\nu=P_Y$. Next, we prove $\Phi_C^{(1)}\big(Y,\iota(\nu)\big)\in\mathcal{H}_{p,C,T}$. For any $t\in[0,T]$,

$$\begin{aligned}\Big\|\sup_{s\in[0,t]}\big|\Phi_C^{(1)}\big(Y,\iota(\nu)\big)_s\big|\Big\|_p&=\Big\|\sup_{s\in[0,t]}\big|X_0+\int_0^s b(u,Y_u,\nu_u)du+\int_0^s\sigma(u,Y_u,\nu_u)dB_u\big|\Big\|_p\\&\leq\Big\|X_0+\int_0^t\big|b(u,Y_u,\nu_u)\big|du+\sup_{s\in[0,t]}\Big|\int_0^s\sigma(u,Y_u,\nu_u)dB_u\Big|\Big\|_p\\&\leq\|X_0\|_p+\Big\|\int_0^t\big|b(u,Y_u,\nu_u)\big|du\Big\|_p+\Big\|\sup_{s\in[0,t]}\Big|\int_0^s\sigma(u,Y_u,\nu_u)dB_u\Big|\Big\|_p\end{aligned}\tag{5.1.9}$$

Owing to Assumption (I), we have $\|X_0\|_p<+\infty$. For the second part of (5.1.9), it follows from Lemma 5.1.7-(a) that

$$\begin{aligned}&\Big\|\int_0^t b(u,Y_u,\nu_u)du\Big\|_p\leq\int_0^t\|b(u,Y_u,\nu_u)\|_p\,du\leq\int_0^t C_{b,\sigma,L,T}\big(1+\|Y_u\|_p+\|\mathcal{W}_p(\nu_u,\delta_0)\|_p\big)du\\&\leq 2C_{b,\sigma,L,T}\int_0^t\big(1+\|Y_u\|_p\big)du\leq 2C_{b,\sigma,L,T}\int_0^t\big(1+e^{CT}\|Y\|_{p,C,T}\big)du<+\infty.\end{aligned}$$

Moreover,

$$\begin{aligned}&\Big\|\sup_{s\in[0,t]}\Big|\int_0^s\sigma(u,Y_u,\nu_u)dB_u\Big|\Big\|_p\\&\leq C_{d,p}^{BDG}\Big\|\sqrt{\int_0^t|||\sigma(u,Y_u,\nu_u)|||^2du}\Big\|_p\quad\text{(by Lemma 5.1.6)}\\&\leq C_{d,p}^{BDG}\Big\|\int_0^t|||\sigma(u,Y_u,\nu_u)|||^2du\Big\|_{\frac{p}{2}}^{\frac12}\quad\Big(\text{since }\Big\|\sqrt{X}\Big\|_p=\big[\mathbb{E}X^{\frac{p}{2}}\big]^{\frac{2}{p}\times\frac12}=\|X\|_{\frac{p}{2}}^{\frac12}\Big)\\&\leq C_{d,p}^{BDG}\Big[\int_0^t\Big\||||\sigma(u,Y_u,\nu_u)|||^2\Big\|_{\frac{p}{2}}du\Big]^{\frac12}\text{ (by Minkowski's inequality, since }p\in[2,+\infty))\\&\leq C_{d,p}^{BDG}\Big[\int_0^t\big\||||\sigma(u,Y_u,\nu_u)|||\big\|_p^2du\Big]^{\frac12}\quad\Big(\text{since }\Big\||X|^2\Big\|_{\frac{p}{2}}=\Big(\mathbb{E}\,|X|^{2\times\frac{p}{2}}\Big)^{\frac{2}{p}}=\|X\|_p^2\Big)\end{aligned}$$

$$
\begin{aligned}
&\le C_{d,p}^{BDG}\Big\{\int_0^t \big\|C_{b,\sigma,L,T}\big[1+|Y_u|+\mathcal{W}_p(\nu_u,\delta_0)\big]\big\|_p^2\,du\Big\}^{\frac{1}{2}} \quad \text{(by Lemma 5.1.7-}(a)\text{)}\\
&\le C_{d,p}^{BDG}\cdot C_{b,\sigma,L,T}\Big\{\int_0^t \big[1+\|Y_u\|_p+\mathcal{W}_p(\nu_u,\delta_0)\big]^2 du\Big\}^{\frac{1}{2}}\\
&\le C_{d,p}^{BDG}\cdot C_{b,\sigma,L,T}\Big\{\int_0^t \big[1+2\,\|Y_u\|_p\big]^2 du\Big\}^{\frac{1}{2}} \quad \text{(since } \mathcal{W}_p(\nu_u,\delta_0)\le\|Y_u\|_p\text{)}\\
&\le C_{d,p}^{BDG}\cdot C_{b,\sigma,L,T}\Big\{2T+\int_0^t 4\,\|Y_u\|_p\,du\Big\}^{\frac{1}{2}} < +\infty \text{ (since } (a+b)^2\le 2(a^2+b^2)\text{),}
\end{aligned}
$$

where the last inequality of the above formula is due to

$$
\int_0^t 4\,\|Y_u\|_p\,du \le \int_0^t 4e^{CT}\,\|Y\|_{p,C,T}\,du \le 4T\cdot e^{CT}\,\|Y\|_{p,C,T} < +\infty.
$$

Hence for every $t\in[0,T]$, $\Big\|\sup_{s\in[0,t]}\big|\Phi_C^{(1)}\big(Y,\iota(\nu)\big)_s\big|\Big\|_p < +\infty$, which directly implies

$$
\Big\|\Phi_C^{(1)}\big(Y,\iota(\nu)\big)\Big\|_{p,C,T} = \sup_{t\in[0,T]} e^{-Ct}\Big\|\sup_{s\in[0,t]}\big|\Phi_C^{(1)}\big(Y,\iota(\nu)\big)_s\big|\Big\|_p < +\infty.
$$

Thus $\Phi_C^{(1)}\big(Y,\iota(\nu)\big)\in\mathcal{H}_{p,C,T}$.

(ii) We will first prove that for any $X,Y\in\mathcal{H}_{p,C,T}$, $d_{p,C,T}\big(\iota(P_X),\iota(P_Y)\big)\le\|X-Y\|_{p,C,T}$. In fact

$$
\begin{aligned}
d_{p,C,T}\big(\iota(P_X),\iota(P_Y)\big) &= \sup_{t\in[0,T]} e^{-Ct}\mathcal{W}_p(P_X\circ\pi_t^{-1},P_Y\circ\pi_t^{-1}) \le \sup_{t\in[0,T]} e^{-Ct}\,\|X_t-Y_t\|_p\\
&\le \sup_{t\in[0,T]} e^{-Ct}\Big\|\sup_{s\in[0,t]}|X_s-Y_s|\Big\|_p \le \|X-Y\|_{p,C,T}\,.
\end{aligned}
$$

Next, we will prove that $\Phi_C^{(1)}$ is Lipschitz continuous. For any $X,Y\in\mathcal{H}_{p,C,T}$, set $\mu=P_X$ and $\nu=P_Y$. Then

$$
\begin{aligned}
&\Big\|\Phi_C^{(1)}\big(X,\iota(\mu)\big)-\Phi_C^{(1)}\big(Y,\iota(\nu)\big)\Big\|_{p,C,T}\\
&= \Big\|\int_0^{\cdot}\big(b(u,X_u,\mu_u)-b(u,Y_u,\nu_u)\big)du+\int_0^{\cdot}\big(\sigma(u,X_u,\mu_u)-\sigma(u,Y_u,\nu_u)\big)dB_u\Big\|_{p,C,T}\\
&\le \Big\|\int_0^{\cdot}\big(b(u,X_u,\mu_u)-b(u,Y_u,\nu_u)\big)du\Big\|_{p,C,T}+\Big\|\int_0^{\cdot}\big(\sigma(u,X_u,\mu_u)-\sigma(u,Y_u,\nu_u)\big)dB_u\Big\|_{p,C,T}\\
&= \sup_{t\in[0,T]} e^{-Ct}\Big\|\sup_{s\in[0,t]}\Big|\int_0^s\big[b(u,X_u,\mu_u)-b(u,Y_u,\nu_u)\big]du\Big|\Big\|_p
\end{aligned}
$$

$$+ \sup_{t\in[0,T]} e^{-Ct} \left\| \sup_{s\in[0,t]} \left| \int_0^s \big[\sigma(u, X_u, \mu_u) - \sigma(u, Y_u, \nu_u)\big] dBu \right| \right\|_p$$

Owing to Lemma 5.1.7, we have

$$\begin{aligned}
&\sup_{t\in[0,T]} e^{-Ct} \left\| \sup_{s\in[0,t]} \left| \int_0^s \big[b(u, X_u, \mu_u) - b(u, Y_u, \nu_u)\big] du \right| \right\|_p \\
&\quad \le L \sup_{t\in[0,T]} e^{-Ct} \int_0^t \big(\|X_u - Y_u\|_p + \mathcal{W}_p(\mu_u, \nu_u)\big) du \\
&\quad \le L \sup_{t\in[0,T]} e^{-Ct} \int_0^t (2 \|X_u - Y_u\|_p) du \quad (\text{since } \mathcal{W}_p(\mu_u, \nu_u) \le \|X_u - Y_u\|_p) \\
&\quad \le 2L \sup_{t\in[0,T]} e^{-Ct} \int_0^t e^{Cu} \big(e^{-Cu} \|X_u - Y_u\|_p \big) du \\
&\quad \le 2L \sup_{t\in[0,T]} e^{-Ct} \int_0^t e^{Cu} du \cdot \|X - Y\|_{p,C,T} \quad (\text{since } e^{-Cu} \|X_u - Y_u\|_p \le \|X - Y\|_{p,C,T}) \\
&\quad = 2L \sup_{t\in[0,T]} e^{-Ct} \frac{e^{Ct} - 1}{C} \cdot \|X - Y\|_{p,C,T} \\
&\quad \le \frac{2L}{C} \|X - Y\|_{p,C,T},
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{t\in[0,T]} e^{-Ct} \left\| \sup_{s\in[0,t]} \left| \int_0^s \big[\sigma(u, X_u, \mu_u) - \sigma(u, Y_u, \nu_u)\big] dBu \right| \right\|_p \\
&\quad \le \sup_{t\in[0,T]} e^{-Ct} C_{d,p,L} \Big\{ \int_0^t \big[\|X_u - Y_u\|_p^2 + \mathcal{W}_p^2(\mu_u, \nu_u) du \big] \Big\}^{\frac{1}{2}} \quad (\text{by Lemma 5.1.7}) \\
&\quad \le \sup_{t\in[0,T]} e^{-Ct} C_{d,p,L} \Big\{ \int_0^t 2 \|X_u - Y_u\|_p^2 du \Big\}^{\frac{1}{2}} \quad (\text{since } \mathcal{W}_p(\mu_u, \nu_u) \le \|X_u - Y_u\|_p) \\
&\quad \le \sqrt{2}\, C_{d,p,L} \sup_{t\in[0,T]} e^{-Ct} \Big\{ \int_0^t e^{2Cu} \Big(e^{-Cu} \|X_u - Y_u\|_p \Big)^2 du \Big\}^{\frac{1}{2}} \\
&\quad \le \sqrt{2}\, C_{d,p,L} \|X - Y\|_{p,C,T} \sup_{t\in[0,T]} e^{-Ct} \Big\{ \int_0^t e^{2Cu} du \Big\}^{\frac{1}{2}} \\
&\qquad (\text{since } e^{-Cu} \|X_u - Y_u\|_p \le \|X - Y\|_{p,C,T}) \\
&\quad \le \sqrt{2}\, C_{d,p,L} \|X - Y\|_{p,C,T} \cdot \sup_{t\in[0,T]} e^{-Ct} \Big[\frac{e^{2Ct} - 1}{2C} \Big]^{\frac{1}{2}} \\
&\quad \le \frac{C_{d,p,L}}{\sqrt{C}} \cdot \|X - Y\|_{p,C,T},
\end{aligned}$$

since $\sup_{t\in[0,T]} e^{-Ct}\left[\frac{e^{2Ct}-1}{2C}\right]^{\frac{1}{2}} \le \sup_{t\in[0,T]}\left[\frac{1-e^{-2Ct}}{2C}\right]^{\frac{1}{2}} = \frac{1}{\sqrt{2C}}$. Consequently,

$$\begin{aligned}
&\left\|\Phi_C^{(1)}\big(X,\iota(\mu)\big)-\Phi_C^{(1)}\big(Y,\iota(\nu)\big)\right\|_{p,C,T}\\
&\quad\le \left\|\int_0^{\cdot} b(u,X_u,\mu_u)du-\int_0^{\cdot} b(u,Y_u,\nu_u)du\right\|_{p,C,T}\\
&\qquad\quad+\left\|\int_0^{\cdot} \sigma(u,X_u,\mu_u)du-\int_0^{\cdot} \sigma(u,Y_u,\nu_u)dB_u\right\|_{p,C,T}\\
&\quad\le \Big(\frac{2L}{C}+\frac{C_{d,p,L}}{\sqrt{C}}\Big)\left\|X-Y\right\|_{p,C,T}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&d_{\mathcal{H}\times\mathcal{P}}\Big(\Phi_C\big(X,\iota(\mu)\big),\Phi_C\big(Y,\iota(\nu)\big)\Big)\\
&\quad=\left\|\Phi_C^{(1)}\big(X,\iota(\mu)\big)-\Phi_C^{(1)}\big(Y,\iota(\nu)\big)\right\|_{p,C,T}+d_{p,C,T}(P_{\Phi_C^{(1)}(X,\iota(\mu))},P_{\Phi_C^{(1)}(Y,\iota(\nu))})\\
&\quad\le 2\left\|\Phi_C^{(1)}\big(X,\iota(\mu)\big)-\Phi_C^{(1)}\big(Y,\iota(\nu)\big)\right\|_{p,C,T}\le 2\Big(\frac{2L}{C}+\frac{C_{d,p,L}}{\sqrt{C}}\Big)\left\|X-Y\right\|_{p,C,T}.\\
&\quad\le 2\Big(\frac{2L}{C}+\frac{C_{d,p,L}}{\sqrt{C}}\Big)\cdot d_{\mathcal{H}\times\mathcal{P}}\Big((X,\mu),(Y,\nu)\Big).
\end{aligned}$$

Hence we can conclude the proof by letting $K_1=4L$ and $K_2=2C_{d,p,L}$. □

Proposition 5.1.1 directly implies the existence and uniqueness of a strong solution of the McKean-Vlasov equation (5.0.1) as shown below.

Theorem 5.1.1. *Under Assumption (I), the McKean-Vlasov equation defined in* (5.0.1) *has a unique strong solution.*

Proof. Proposition 5.1.1 implies that Φ_C is a Lipschitz continuous function. Thus, $F_C \coloneqq \Phi_C\big(\mathcal{H}_{p,C,T}\times\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)\big)$ is a closed set in $\mathcal{H}_{p,C,T}\times\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)$. Moreover, For a large enough constant C, we have $\big(\frac{K_1}{C}+\frac{K_2}{\sqrt{C}}\big)<1$, then Φ_C is a contraction mapping. Therefore, Φ_C has a unique fixed point $\big(H,\iota(P_H)\big)\in F_C\subset\mathcal{H}_{p,C,T}\times\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)$ and this process H is the unique strong solution of the McKean-Vlasov equation (5.0.1). □

5.2 Convergence rate of the theoretical Euler scheme

We add the following assumption in this section.

Assumption (II): *For every $s,t\in[0,T]$ with $s<t$, there exist positive constants $\tilde{L},\gamma$*

such that

$$\forall x \in \mathbb{R}^d, \forall \mu \in \mathcal{P}(\mathbb{R}^d),$$
$$|b(t,x,\mu) - b(s,x,\mu)| \vee |||\sigma(t,x,\mu) - \sigma(s,x,\mu)||| \leq \tilde{L}\big(1 + |x| + \mathcal{W}_p(\mu, \delta_0)\big)(t-s)^{\gamma}.$$

Let $(X_t)_{t\in[0,T]}$ be the unique solution of (5.0.1) and let $\mu_t = P_{X_t}, t \in [0,T]$ be its marginal distribution at time $t \in [0,T]$. Moreover, let $(\bar{X}_{t_m})_{m=0,\dots,M}$ be the Euler scheme defined by (5.0.2) and let $\bar{\mu}_{t_m} = P_{\bar{X}_{t_m}}, m = 0, ..., M$. The main result of this section is the following proposition.

Proposition 5.2.1 (Convergence rate of the theoretical Euler Scheme)**.** *Under Assumption (I) and (II), one has*

$$\sup_{0\leq m\leq M} \mathcal{W}_p(\bar{\mu}^M_{t_m}, \mu_{t_m}) \leq \left\| \sup_{0\leq m\leq M} \left| X_{t_m} - \bar{X}^M_{t_m} \right| \right\|_p \leq \tilde{C} h^{\frac{1}{2}\wedge\gamma}, \tag{5.2.1}$$

where $\tilde{C}$ is a constant depending on $L, \tilde{L}, p, d, \|X_0\|_p, T, \gamma$.

Remark 5.2.1. If the McKean-Vlasov equation (5.0.1) is homogeneous, i.e. the coefficient functions b and σ do not depend on t, Assumption (II) is directly satisfied with γ as large as we want. In this case, the convergence rate of theoretical Euler scheme is

$$\sup_{0\leq m\leq M} \mathcal{W}_p(\bar{\mu}^M_{t_m}, \mu_{t_m}) \leq \left\| \sup_{0\leq m\leq M} \left| X_{t_m} - \bar{X}^M_{t_m} \right| \right\|_p \leq \tilde{C} h^{\frac{1}{2}}. \tag{5.2.2}$$

In order to prove Proposition 5.2.1, we introduce the *continuous time Euler scheme* $(\bar{X}_t)_{t\in[0,T]}$ which reads as follows: set $\bar{X}_0 = X_0$ and for every $t \in [t_m, t_{m+1})$, define

$$\bar{X}_t := \bar{X}_{t_m} + b(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m})(t - t_m) + \sigma(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m})(B_t - B_{t_m}). \tag{5.2.3}$$

The above definition implies that $\bar{X} := (\bar{X}_t)_{t\in[0,T]}$ is a $\mathcal{C}([0,T], \mathbb{R}^d)$-valued stochastic process. Let $\bar{\mu}$ denote the probability distribution of $\bar{X}$ and for every $t \in [0,T]$, let $\bar{\mu}_t$ denote the marginal distribution of $\bar{X}_t$. Then $(\bar{X}_t)_{t\in[0,T]}$ is the solution of

$$\begin{cases} d\bar{X}_t = b(\underline{t}, \bar{X}_{\underline{t}}, \bar{\mu}_{\underline{t}})dt + \sigma(\underline{t}, \bar{X}_{\underline{t}}, \bar{\mu}_{\underline{t}})dB_t, \\ \bar{X}_0 = X_0, \end{cases} \tag{5.2.4}$$

where for every $t \in [t_m, t_{m+1})$, $\underline{t} := t_m$.

Now we recall a variant version of Gronwall's Lemma and we refer to Pagès (2018)[Lemma 7.3] for a proof (among many others).

Lemma 5.2.1 ("À la Gronwall" Lemma)**.** *Let $f : [0,T] \to \mathbb{R}_+$ be a Borel, locally bounded,*

non-negative and non-decreasing function and let $\psi : [0,T] \to \mathbb{R}_+$ *be a non-negative non-decreasing function satisfying*

$$\forall t \in [0,T],\ f(t) \leq A\int_0^t f(s)ds + B\left(\int_0^t f^2(s)ds\right)^{\frac{1}{2}} + \psi(t),$$

where A, B *are two positive real constants. Then, for any* $t \in [0,T]$,

$$f(t) \leq 2e^{(2A+B^2)t}\psi(t).$$

The proof of Proposition 5.2.1 relies on the following lemma.

Lemma 5.2.2. *Under Assumption (I), let* X *be the unique strong solution of* (5.0.1) *and let* $(\bar{X}_t)_{t\in[0,T]}$ *be the process defined in* (5.2.3)*. Then*

(a) There exists a constant $C_{p,d,b,\sigma}$ *depending on* p, d, b, σ *such that for every* $t \in [0,T]$,

$$\forall M \geq 1,\quad \left\|\sup_{u\in[0,t]}|X_u|\right\|_p \vee \left\|\sup_{u\in[0,t]}\left|\bar{X}_u^M\right|\right\|_p \leq C_{p,d,b,\sigma}e^{C_{p,d,b,\sigma}t}(1+\|X_0\|_p).$$

(b) There exists a constant κ *depending on* $L, b, \sigma, \|X_0\|, p, d, T$ *such that for any* $s, t \in [0,T], s < t$,

$$\forall M \geq 1,\quad \left\|\bar{X}_t^M - \bar{X}_s^M\right\|_p \vee \|X_t - X_s\|_p \leq \kappa\sqrt{t-s}.$$

Proof. (a) If X is the unique strong solution of (5.0.1), then its probability distribution μ is the unique weak solution. We define two new coefficient functions depending on $\iota(\mu) = (\mu_t)_{t\in[0,T]}$ by

$$\tilde{b}(t,x) := b(t,x,\mu_t) \text{ and } \tilde{\sigma}(t,x) := \sigma(t,x,\mu_t).$$

Now we discuss the continuity in t of $\tilde{b}$ and $\tilde{\sigma}$. In fact,

$$\begin{aligned}\left|\tilde{b}(t,x) - \tilde{b}(s,x)\right| &\leq |b(t,x,\mu_t) - b(s,x,\mu_s)| \\ &\leq |b(t,x,\mu_t) - b(s,x,\mu_t)| + |b(s,x,\mu_t) - b(s,x,\mu_s)| \\ &\leq |b(t,x,\mu_t) - b(s,x,\mu_t)| + \mathcal{W}_p(\mu_t,\mu_s), \end{aligned} \tag{5.2.5}$$

and we have a similar inequality for $\tilde{\sigma}$. Moreover, we know from Assumption (I) that b and σ are continuous in t and from Lemma 5.1.2 that $\iota(\mu) = (\mu_t)_{t\in[0,T]} \in \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big)$. Hence, $\tilde{b}$ and $\tilde{\sigma}$ are continuous in t. Moreover, it is obvious that $\tilde{b}$ and $\tilde{\sigma}$ are still Lipschitz in x. Consequently, X is also the unique strong solution of the following stochastic

differential equation

$$dX_t = \tilde{b}(t, X_t)dt + \tilde{\sigma}(t, X_t)dB_t$$

with X_0 same as in (5.0.1).

Hence, the inequality

$$\left\| \sup_{u\in[0,t]} |X_u| \right\|_p \leq C_{p,d,b,\sigma} e^{C_{p,d,b,\sigma} t}(1 + \|X_0\|_p)$$

can be obtained by the usual method for the regular stochastic differential equation for which we refer to Pagès (2018)[Proposition 7.2 and (7.12)] among many other references.

Next, we prove the inequality for $\left\|\sup_{u\in[0,t]} |\bar{X}_u^M|\right\|_p$.

We go back the discrete Euler scheme

$$\bar{X}_{t_{m+1}}^M = \bar{X}_{t_m}^M + h \cdot b(t_m, \bar{X}_{t_m}^M, \bar{\mu}_{t_m}^M) + \sqrt{h}\sigma(t_m, \bar{X}_{t_m}^M, \bar{\mu}_{t_m}^M)Z_{m+1}.$$

We write $\bar{X}_{t_m}$ instead of $\bar{X}_{t_m}^M$ in the following. By Minkovski's inequality, we have

$$\left\|\bar{X}_{t_{m+1}}\right\|_p = \left\|\bar{X}_{t_m}\right\|_p + h\left\|b(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m})\right\|_p + \sqrt{h}\left\| |||\sigma(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m})||| \, |Z_{m+1}| \right\|_p.$$

As Z_{m+1} is independent of the $\sigma-$algebra generated by $\bar{X}_{t_0}, ..., \bar{X}_{t_m}$, one can imply the linear growth result in Lemma 5.1.7 and obtain

$$\left\|\bar{X}_{t_{m+1}}\right\|_p = \left\|\bar{X}_{t_m}\right\|_p + C_{b,\sigma,L,T}(h + c_p h^{1/2})\big(1 + \left\|\bar{X}_{t_m}\right\|_p + \mathcal{W}_p(\delta_0, \bar{X}_{t_m})\big),$$

where $C_{b,\sigma,L,T}$ and c_p are two real constants. As $\mathcal{W}_p(\delta_0, \bar{X}_{t_m}) \leq \left\|\bar{X}_{t_m}\right\|_p$, there exists a constant C such that

$$\left\|\bar{X}_{t_{m+1}}\right\|_p \leq C\left\|\bar{X}_{t_m}\right\|_p,$$

which in turn implies by induction that

$$\max_{m=0,...,M} \left\|\bar{X}_{t_m}\right\|_p < +\infty$$

since $\left\|\bar{X}_0\right\|_p = \|X_0\|_p < +\infty$.

For every $t \in [t_m, t_{m+1}]$, it follows from the definition (5.2.3) that

$$\left\|\bar{X}_t^M\right\|_p \leq \left\|\bar{X}_{t_m}\right\|_p + (t - t_m)\left\|b(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m})\right\|_p + \left\| |||\sigma(t_m, \bar{X}_{t_m}, \bar{\mu}_{t_m})||| \, |B_t - B_{t_m}| \right\|_p.$$

We write $\bar{X}_t$ instead of $\bar{X}_t^M$ in the following when there is no ambiguity.

As $B_t - B_{t_m}$ is independent to $\sigma(\mathcal{F}_s, s \le t_m)$, it follows that

$$\begin{aligned}\left\|\bar{X}_t\right\|_p &\le \left\|\bar{X}_{t_m}\right\|_p + C_{b,\sigma,L,T}\big(1 + \left\|\bar{X}_{t_m}\right\|_p + \mathcal{W}_p(\delta_0, \bar{X}_{t_m})\big)\big(h + c_p(t-t_m)^p\big)\\ &\le C_1 \left\|\bar{X}_{t_m}\right\|_p + C_2,\end{aligned}$$

where C_1 and C_2 are two constants. Finally, for every $M \ge 1$,

$$\sup_{t\in[0,T]} \left\|\bar{X}_t^M\right\|_p < +\infty. \tag{5.2.6}$$

Consequently,

$$\begin{aligned}&\left\|\sup_{u\in[0,t]}\left|\bar{X}_u^M\right|\right\|_p\\ &\le \|X_0\|_p + \left\|\int_0^t \left|b(s,\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\right| ds\right\|_p + \left\|\sup_{u\in[0,t]}\left|\int_0^u \sigma(s,\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})dB_s\right|\right\|_p\\ &\quad\text{(Minkowski's Inequality)}\\ &\le \|X_0\|_p + \int_0^t \left\|b(s,\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\right\|_p ds + C_{d,p}^{BDG}\left\|\sqrt{\int_0^t \left|\left|\left|\sigma(s,\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\right|\right|\right|^2 ds}\right\|_p\\ &\quad\text{(by Lemma 5.1.5 and 5.1.6)}\\ &\le \|X_0\|_p + \int_0^t C_{b,\sigma,L,T}\left\|1 + \left|\bar{X}_{\underline{s}}\right| + \mathcal{W}_p(\bar{\mu}_{\underline{s}},\delta_0)\right\|_s ds\\ &\quad + C_{d,p,L}^{BDG}\left\|\sqrt{\int_0^t \left|1 + \left|\bar{X}_{\underline{s}}\right| + \mathcal{W}_p(\bar{\mu}_{\underline{s}},\delta_0)\right|^2 ds}\right\|_p \quad\text{(by Lemma 5.1.7 } - (a))\\ &\le \|X_0\|_p + \int_0^t C_{b,\sigma,L,T}(1 + 2\left\|\bar{X}_{\underline{s}}\right\|_p)ds + C_{d,p,L}^{BDG}\left\|\sqrt{\int_0^t 4\big(1 + \left|\bar{X}_{\underline{s}}\right|^2 + \mathcal{W}_p^2(\bar{\mu}_{\underline{s}},\delta_0)\big)ds}\right\|_p\\ &\le \|X_0\|_p + \int_0^t C_{b,\sigma,L,T}(1 + 2\left\|\bar{X}_{\underline{s}}\right\|_p)ds\\ &\quad + C_{d,p,L}^{BDG}\left\|\sqrt{4\big[t + \int_0^t \left|\bar{X}_{\underline{s}}\right|^2 ds + \int_0^t \mathcal{W}_p^2(\bar{\mu}_{\underline{s}},\delta_0)ds\big]}\right\|_p\\ &\le \|X_0\|_p + \int_0^t C_{b,\sigma,L,T}(1 + 2\left\|\bar{X}_{\underline{s}}\right\|_p)ds + C_{d,p,L}^{BDG'}\left\|\sqrt{t} + \sqrt{\int_0^t \left|\bar{X}_{\underline{s}}\right|^2 ds} + \sqrt{\int_0^t \mathcal{W}_p^2(\bar{\mu}_{\underline{s}},\delta_0)ds}\right\|_p\\ &\le \|X_0\|_p + \int_0^t C_{b,\sigma,L,T}(1 + 2\left\|\bar{X}_{\underline{s}}\right\|_p)ds\\ &\quad + C_{d,p,L}^{BDG'}\Big[\sqrt{t} + \left\|\sqrt{\int_0^t \left|\bar{X}_{\underline{s}}\right|^2 ds}\right\|_p + \sqrt{\int_0^t \mathcal{W}_p^2(\bar{\mu}_{\underline{s}},\delta_0)ds}\Big]\\ &\le \|X_0\|_p + \int_0^t C_{b,\sigma,L,T}(1 + 2\left\|\bar{X}_{\underline{s}}\right\|_p)ds\end{aligned}$$

$$
\begin{aligned}
&+ C_{d,p,L}^{BDG'}\Big[\sqrt{t} + \Big\|\int_0^t |\bar{X}_{\underline{s}}|^2\, ds\Big\|_{\frac{p}{2}}^{\frac{1}{2}} + \Big(\int_0^t \mathcal{W}_p^2(\bar{\mu}_{\underline{s}}, \delta_0) ds\Big)^{\frac{1}{2}}\Big] \\
&\le \|X_0\|_p + \int_0^t C_{b,\sigma,L,T}(1 + 2\,\|\bar{X}_{\underline{s}}\|_p) ds \\
&\quad + C_{d,p,L}^{BDG'}\Big[\sqrt{t} + \Big[\int_0^t \Big\||\bar{X}_{\underline{s}}|^2\Big\|_{\frac{p}{2}} ds\Big]^{\frac{1}{2}} + \Big[\int_0^t \mathcal{W}_p^2(\bar{\mu}_{\underline{s}}, \delta_0) ds\Big]^{\frac{1}{2}}\Big] \\
&\text{(by Lemma 5.1.5 since } \frac{p}{2} \ge 1\text{).}
\end{aligned} \tag{5.2.7}
$$

It follows from $\big\||\bar{X}_{\underline{s}}|^2\big\|_{\frac{p}{2}} = \big[\mathbb{E}\,|\bar{X}_{\underline{s}}|^{2\cdot\frac{p}{2}}\big]^{\frac{2}{p}} = \|\bar{X}_{\underline{s}}\|_p^2$ and

$$
\Big[\int_0^t \mathcal{W}_p^2(\bar{\mu}_{\underline{s}}, \delta_0) ds\Big]^{\frac{1}{2}} \le \Big[\int_0^t \|\mathcal{W}_p(\bar{\mu}_{\underline{s}}, \delta_0)\|_p^2\, ds\Big]^{\frac{1}{2}} \le \Big[\int_0^t \|\bar{X}_{\underline{s}}\|_p^2\, ds\Big]^{\frac{1}{2}}
$$

that

$$
\Big\|\sup_{u\in[0,t]} |\bar{X}_u^M|\Big\|_p \le \|X_0\|_p + \int_0^t C_{b,\sigma,L,T}(1 + 2\,\|\bar{X}_{\underline{s}}\|_p) ds + C_{d,p,L}^{BDG'}\Big(\sqrt{t} + \Big[\int_0^t \|\bar{X}_{\underline{s}}\|_p^2\, ds\Big]^{\frac{1}{2}}\Big). \tag{5.2.8}
$$

Hence, for any $t \in [0,T]$, (5.2.8) implies that, for every $M \ge 1$,

$$
\Big\|\sup_{u\in[0,t]} |\bar{X}_u^M|\Big\|_p < +\infty
$$

owing to (5.2.6).

In order to establish the uniformity in M, we come back to (5.2.8). As $\|\bar{X}_{\underline{s}}\|_p \le \big\|\sup_{u\in[0,s]} |\bar{X}_u|\big\|_p$, it follows that

$$
\begin{aligned}
\Big\|\sup_{u\in[0,t]} |\bar{X}_u^M|\Big\|_p &\le \|X_0\|_p + C_{b,\sigma,L,T}(t + C_{d,p,L}^{BDG'}\sqrt{t}) \\
&\quad + C_{b,\sigma,L,T}\Big\{\int_0^t \Big\|\sup_{u\in[0,s]} |\bar{X}_u|\Big\|_p ds + C_{d,p,L}^{BDG'}\Big[\int_0^t \Big\|\sup_{u\in[0,s]} |\bar{X}_u|\Big\|_p^2 ds\Big]^{\frac{1}{2}}\Big\}.
\end{aligned}
$$

Hence,

$$
\Big\|\sup_{u\in[0,t]} |\bar{X}_u^M|\Big\|_p \le 2e^{(2C_{b,\sigma,L,T} + C_{d,p,L}^{BDG'2})t}(\|X_0\|_p + C_{b,\sigma,L,T}(t + C_{d,p,L}^{BDG}\sqrt{t})),
$$

by appling Lemma 5.2.1. Thus one can take

$$C_{p,d,b,\sigma} = (2C_{b,\sigma,L,T} + C_{d,p,L}^{BDG'^2}) \vee 2C_{b,\sigma,L,T}(T + C_{d,p,L}^{BDG}\sqrt{T}) \vee 2$$

to conclude the proof.

(b) It follows from $|X_t - X_s| = \left|\int_s^t b(u, X_u, \mu_u)du + \int_s^t \sigma(u, X_u, \mu_u)dB_u\right|$ that,

$$\begin{aligned}
\|X_t - X_s\|_p &\leq \left\|\int_s^t b(u, X_u, \mu_u)du\right\|_p + \left\|\int_s^t \sigma(u, X_u, \mu_u)dB_u\right\|_p \\
&\leq \int_s^t \|b(u, X_u, \mu_u)\|_p \, du + C_{d,p}^{BDG} \left\|\int_s^t |||\sigma(u, X_u, \mu_u)|||^2 du\right\|_{\frac{p}{2}}^{\frac{1}{2}} \\
&\qquad \text{(by Lemma 5.1.5 and Lemma 5.1.6)} \\
&\leq \int_s^t C_{b,\sigma,L,T}\Big[1 + \|X_u\|_p + \|\mathcal{W}_p(\mu_p, \delta_0)\|_p\Big]du \\
&\qquad + C_{d,p}^{BDG} \left\|\int_s^t C_{b,\sigma,L,T}\Big[1 + \|X_u\|_p + \|\mathcal{W}_p(\mu_p, \delta_0)\|_p\Big]^2 du\right\|_{\frac{p}{2}}^{\frac{1}{2}} \quad \text{(by Lemma 5.1.7} - (a)) \\
&\leq \int_s^t C_{b,\sigma,L,T}\Big[1 + 2\|X_u\|_p\Big]du + 4C_{d,p}^{BDG} \cdot C_{b,\sigma,L,T} \left\|\int_s^t \Big[1 + \|X_u\|_p^2 + \mathcal{W}_p^2(\mu_p, \delta_0)\Big]du\right\|_{\frac{p}{2}}^{\frac{1}{2}} \\
&\leq \int_s^t C_{b,\sigma,L,T}\Big[1 + 2\|X_u\|_p\Big]du \\
&\qquad + 4C_{d,p}^{BDG} \cdot C_{b,\sigma,L,T}\Big[(t-s) + \left\|\int_s^t |X_u|^2 du\right\|_{\frac{p}{2}} + \left\|\int_s^t \mathcal{W}_p^2(\mu_u, \delta_0)du\right\|_{\frac{p}{2}}\Big]^{\frac{1}{2}} \\
&\leq \int_s^t C_{b,\sigma,L,T}\Big[1 + 2\|X_u\|_p\Big]du \\
&\qquad + 4C_{d,p}^{BDG} \cdot C_{b,\sigma,L,T}\Big[\sqrt{t-s} + \Big[\int_s^t \left\||X_u|^2\right\|_{\frac{p}{2}} du\Big]^{\frac{1}{2}} + \Big[\int_s^t \left\|\mathcal{W}_p^2(\mu_u, \delta_0)\right\|_{\frac{p}{2}} du\Big]^{\frac{1}{2}} \\
&\leq \int_s^t C_{b,\sigma,L,T}\Big[1 + 2\left\|\sup_{u\in[0,T]} |X_u|\right\|_p\Big]du \\
&\qquad + 4C_{d,p}^{BDG} \cdot C_{b,\sigma,L,T}\Big\{\sqrt{t-s} + \sqrt{\int_s^t \|X_u\|_p^2 \, du} + \sqrt{\int_s^t \|\mathcal{W}_p(\mu_u, \delta_0)\|_p^2 \, du}\Big\} \\
&\leq C_{b,\sigma,L,T}\Big[1 + 2\left\|\sup_{u\in[0,T]} |X_u|\right\|_p\Big](t-s) \\
&\qquad + 4C_{d,p}^{BDG} \cdot C_{b,\sigma,L,T}\Big\{\sqrt{t-s} + 2\sqrt{t-s}\left\|\sup_{u\in[0,T]} |X_u|\right\|_p^2\Big\} \\
&\leq \Big\{C_{b,\sigma,L,T}\Big[1 + 2\left\|\sup_{u\in[0,T]} |X_u|\right\|_p\Big]\sqrt{T} + 4C_{d,p}^{BDG} \cdot C_{b,\sigma,L,T}[1 + 2\left\|\sup_{u\in[0,T]} |X_u|\right\|_p^2]\Big\}\sqrt{t-s}.
\end{aligned}$$

Owing to the result in (a), $\left\|\sup_{u\in[0,T]}|X_u|\right\|_p \le C_{p,d,b,\sigma}e^{C_{p,d,b,\sigma}t}(1+\|X_0\|_p)$, then one can conclude by setting

$$\begin{aligned}\kappa = C_{L,b,\sigma,\|X_0\|,p,d,T} :=& C_{b,\sigma,L,T}\Big[1+2C_{p,d,b,\sigma}e^{C_{p,d,b,\sigma}t}(1+\|X_0\|_p)\Big]\sqrt{T}\\ &+4C_{d,p}^{BDG}\cdot C_{b,\sigma,L,T}[1+2C_{p,d,b,\sigma}^2e^{2C_{p,d,b,\sigma}t}(1+\|X_0\|_p)^2].\end{aligned}$$

□

Proof of Proposition 5.2.1. We write $\bar{X}_t$ and $\bar{\mu}_t$ instead of $\bar{X}_t^M$ and $\bar{\mu}_t^M$ to simplify the notation in this proof. For every $s\in[0,T]$, set

$$\varepsilon_s := X_s-\bar{X}_s = \int_0^s \big(b(u,X_u,\mu_u)-b(\underline{u},\bar{X}_{\underline{u}},\bar{\mu}_{\underline{u}})\big)du+\int_0^s \big(\sigma(u,X_u,\mu_u)-\sigma(\underline{u},\bar{X}_{\underline{u}},\bar{\mu}_{\underline{u}})\big)dB_u,$$

and let

$$f(t) := \left\|\sup_{s\in[0,t]}|\varepsilon_s|\right\|_p = \left\|\sup_{s\in[0,t]}|X_s-\bar{X}_s|\right\|_p.$$

It follows from Lemma 5.2.2-(a) that $\bar{X}=(\bar{X}_t)_{t\in[0,T]}\in L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega,\mathcal{F},\mathbb{P})$. Consequently, $\bar{\mu}\in\mathcal{P}_p\big(\mathcal{C}([0,T],\mathbb{R}^d)\big)$ and $\iota(\mu)=(\mu_t)_{t\in[0,T]}\in\mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)$ by applying Lemma 5.1.2. Hence,

$$\begin{aligned}
f(t) &= \left\|\sup_{s\in[0,t]}|X_s-\bar{X}_s|\right\|_p\\
&\le \left\|\int_0^t\big|b(s,X_s,\mu_s)-b(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\big|\,ds+\sup_{s\in[0,t]}\left|\int_0^s\big(\sigma(u,X_u,\mu_u)-\sigma(\underline{u},\bar{X}_{\underline{u}},\bar{\mu}_{\underline{u}})\big)dB_u\right|\right\|_p\\
&\le \int_0^t\big\|b(s,X_s,\mu_s)-b(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\big\|_p\,ds+C_{d,p}^{BDG}\left\|\sqrt{\int_0^t\big\|\|\sigma(s,X_s,\mu_s)-\sigma(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\|\big\|^2ds}\right\|_p\\
&= \int_0^t\big\|b(s,X_s,\mu_s)-b(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\big\|_p\,ds+C_{d,p}^{BDG}\left\|\int_0^t\big\|\|\sigma(s,X_s,\mu_s)-\sigma(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\|\big\|^2ds\right\|_{\frac{p}{2}}^{\frac12}\\
&\le \int_0^t\big\|b(s,X_s,\mu_s)-b(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\big\|_p\,ds+C_{d,p}^{BDG}\Big[\int_0^t\Big\|\big\|\|\sigma(s,X_s,\mu_s)-\sigma(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\|\big\|^2\Big\|_{\frac{p}{2}}ds\Big]^{\frac12}\\
&= \int_0^t\big\|b(s,X_s,\mu_s)-b(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\big\|_p\,ds+C_{d,p}^{BDG}\Big[\int_0^t\big\|\|\|\sigma(s,X_s,\mu_s)-\sigma(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\|\|_p^2ds\Big]^{\frac12}\\
&\le \int_0^t\|b(s,X_s,\mu_s)-b(\underline{s},X_s,\mu_s)\|_p\,ds+\int_0^t\big\|b(\underline{s},X_s,\mu_s)-b(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\big\|_p\,ds\\
&+C_{d,p}^{BDG}\Big[\int_0^t\big\|\|\|\sigma(s,X_s,\mu_s)-\sigma(\underline{s},X_s,\mu_s)\|\|+\|\|\sigma(\underline{s},X_s,\mu_s)-\sigma(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\|\|\big\|_p^2ds\Big]^{\frac12},
\end{aligned}\tag{5.2.9}$$

where the last term of (5.2.9) can be upper-bounded by

$$
\begin{aligned}
&C_{d,p}^{BDG}\Big[\int_0^t \big\|\|\|\sigma(s,X_s,\mu_s)-\sigma(\underline{s},X_s,\mu_s)\|\|+\big\|\|\sigma(\underline{s},X_s,\mu_s)-\sigma(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\|\|\big\|_p^2 ds\Big]^{\frac12}\\
&\le C_{d,p}^{BDG}\Big[\int_0^t \big[\big\|\|\|\sigma(s,X_s,\mu_s)-\sigma(\underline{s},X_s,\mu_s)\|\|\big\|_p+\big\|\|\|\sigma(\underline{s},X_s,\mu_s)-\sigma(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\|\|\big\|_p\big]^2 ds\Big]^{\frac12}\\
&\le \sqrt{2}C_{d,p}^{BDG}\big[\int_0^t \big\|\|\|\sigma(s,X_s,\mu_s)-\sigma(\underline{s},X_s,\mu_s)\|\|\big\|_p^2 ds\big]^{\frac12}\\
&\qquad+\sqrt{2}C_{d,p}^{BDG}\big[\int_0^t \big\|\|\|\sigma(\underline{s},X_s,\mu_s)-\sigma(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\|\|\big\|_p^2 ds\big]^{\frac12}. \qquad (5.2.10)
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\int_0^t \|b(s,X_s,\mu_s)-b(\underline{s},X_s,\mu_s)\|_p\,ds+\sqrt{2}C_{d,p}^{BDG}\big[\int_0^t \big\|\|\|\sigma(s,X_s,\mu_s)-\sigma(\underline{s},X_s,\mu_s)\|\|\big\|_p^2 ds\big]^{\frac12}\\
&\le \int_0^t \big\|(s-\underline{s})^{\gamma}\tilde{L}\big(1+|X_s|+\mathcal{W}_p(\mu_s,\delta_0)\big)\big\|_p\,ds\\
&\qquad+\sqrt{2}C_{d,p}^{BDG}\big[\int_0^t \big\|(s-\underline{s})^{\gamma}\tilde{L}\big(1+|X_s|+\mathcal{W}_p(\mu_s,\delta_0)\big)\big\|_p^2 ds\big]^{\frac12}\quad\text{(by Assumption (II))}\\
&\le h^{\gamma}T\tilde{L}(1+2\big\|\sup_{s\in[0,T]}|X_s|\,\big\|_p)+\sqrt{2}h^{\gamma}\tilde{L}C_{d,p}^{BDG}\big[T(2+4\big\|\sup_{s\in[0,T]}|X_s|\,\big\|_p^2)\big]^{\frac12}\\
&\le h^{\gamma}T\tilde{L}(1+2\big\|\sup_{s\in[0,T]}|X_s|\,\big\|_p)+\sqrt{2}h^{\gamma}\tilde{L}C_{d,p}^{BDG}\big[\sqrt{2T}+2\sqrt{T}\big\|\sup_{s\in[0,T]}|X_s|\,\big\|_p\big] \qquad (5.2.11)
\end{aligned}
$$

and

$$
\begin{aligned}
&\int_0^t \big\|b(\underline{s},X_s,\mu_s)-b(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\big\|_p\,ds+\sqrt{2}C_{d,p}^{BDG}\big[\int_0^t \big\|\|\|\sigma(\underline{s},X_s,\mu_s)-\sigma(\underline{s},\bar{X}_{\underline{s}},\bar{\mu}_{\underline{s}})\|\|\big\|_p^2 ds\big]^{\frac12}\\
&\le \int_0^t \big\|L\big(\,|X_s-\bar{X}_{\underline{s}}\,|+\mathcal{W}_p(\mu_s,\bar{\mu}_{\underline{s}})\big)\big\|_p\,ds+\sqrt{2}C_{d,p}^{BDG}\Big[\int_0^t \big\|L\big(\,|X_s-\bar{X}_{\underline{s}}\,|+\mathcal{W}_p(\mu_s,\bar{\mu}_{\underline{s}})\big)\big\|_p^2\,ds\Big]^{\frac12}\\
&\le \int_0^t 2L\,\big\|X_s-\bar{X}_{\underline{s}}\big\|_p\,ds+\sqrt{2}C_{d,p}^{BDG}\Big[\int_0^t 4L^2\,\big\|X_s-\bar{X}_{\underline{s}}\,\big\|_p^2\,ds\Big]^{\frac12}\\
&\le \int_0^t 2L\Big[\underbrace{\big\|X_s-X_{\underline{s}}\big\|_p}_{\le\kappa\sqrt{s-\underline{s}}\le\kappa\sqrt{h}}+\big\|X_{\underline{s}}-\bar{X}_{\underline{s}}\big\|_p\Big]ds\\
&\qquad+\sqrt{2}C_{d,p}^{BDG}\Big[\int_0^t 4L^2\big[\underbrace{\big\|X_s-X_{\underline{s}}\big\|_p}_{\le\kappa\sqrt{s-\underline{s}}\le\kappa\sqrt{h}}+\big\|X_{\underline{s}}-\bar{X}_{\underline{s}}\big\|_p\big]^2 ds\Big]^{\frac12}\text{(by applying Lemma 5.2.2-}(b)\text{)}\\
&\le \int_0^t 2L\Big[\kappa\sqrt{h}+\big\|X_{\underline{s}}-\bar{X}_{\underline{s}}\big\|_p\Big]ds\\
&\qquad+\sqrt{2}C_{d,p}^{BDG}\Big[\int_0^t 4L^2\big[\kappa\sqrt{h}+\big\|X_{\underline{s}}-\bar{X}_{\underline{s}}\big\|_p\big]^2 ds\Big]^{\frac12}\\
&\le 2Lt\kappa\sqrt{h}+4C_{d,p}^{BDG}L\sqrt{t}\kappa\sqrt{h}+2L\int_0^t f(s)ds+\sqrt{2}C_{d,p}^{BDG}4L\big[\int_0^t f(s)^2 ds\big]^{\frac12}. \qquad (5.2.12)
\end{aligned}
$$

Let $\tilde{\kappa}(T,\|X_0\|_p)=C_{p,d,b,\sigma}e^{C_{p,d,b,\sigma}t}(1+\|X_0\|_p)$, which is the right hand side of results in

Lemma 5.2.2-(a). A combination of (5.2.9), (5.2.10), (5.2.11) and (5.2.12) leads to

$$\begin{aligned}
f(t) &= \Big\| \sup_{s\in[0,t]} |X_s - \bar{X}_s| \Big\|_p \\
&\leq h^\gamma T\tilde{L}(1 + 2\big\| \sup_{s\in[0,T]} |X_s| \big\|_p) + \sqrt{2}h^\gamma \tilde{L} C_{d,p}^{BDG}\big[\sqrt{2T} + 2\sqrt{T}\big\| \sup_{s\in[0,T]} |X_s| \big\|_p\big] \\
&\quad + 2Lt\kappa\sqrt{h} + \sqrt{2}C_{d,p}^{BDG} 2\sqrt{2}L\sqrt{t}\kappa\sqrt{h} + 2L\int_0^t f(s)ds + \sqrt{2}C_{d,p}^{BDG} 4L\big[\int_0^t f(s)^2 ds\big]^{\frac{1}{2}}. \\
&\leq h^{\frac{1}{2}\wedge\gamma}\psi(T) + 2L\int_0^t f(s)ds + \sqrt{2}C_{d,p}^{BDG} 4L\big[\int_0^t f(s)^2 ds\big]^{\frac{1}{2}},
\end{aligned}$$

where

$$\begin{aligned}
\psi(T) =& T^{\gamma-\gamma\wedge\frac{1}{2}}\big[T\tilde{L}\big(1 + 2\tilde{\kappa}(T, \|X_0\|_p)\big) + \sqrt{2}\tilde{L}C_{d,p}^{BDG}\big(\sqrt{2T} + 2\sqrt{T}\tilde{\kappa}(T, \|X_0\|_p)\big)\big] \\
&+ T^{\frac{1}{2}-\gamma\wedge\frac{1}{2}}\big[2LT\kappa + 4C_{d,p}^{BDG}L\sqrt{T}\kappa\big].
\end{aligned}$$

Then it follows from lemma 5.2.1 that $f(t) \leq 2e^{(4L+16C_{d,p}^{BDG^2}L^2)T} \cdot \psi(T)h^{\gamma\wedge\frac{1}{2}}$. Then we can conclude the proof by letting $\tilde{C} = 2e^{(4L+16C_{d,p}^{BDG^2}L^2)T} \cdot \psi(T)$. □

The proof of Proposition 5.2.1 directly derives the following result.

Corollary 5.2.1. *Let $\bar{X} := (\bar{X}_t)_{t\in[0,T]}$ denote the process defined by the continuous time Euler scheme (5.2.3) with step $h = \frac{T}{M}$ and let $X := (X_t)_{t\in[0,T]}$ denote the unique solution of the McKean-Vlasov equation (5.0.1). Then under Assumption (I) and (II), one has*

$$\mathbb{W}_p(\bar{X}, X) \leq \Big\| \sup_{t\in[0,T]} |X_t - \bar{X}_t| \Big\|_p \leq \tilde{C}h^{\frac{1}{2}\wedge\gamma}, \tag{5.2.13}$$

where $\tilde{C}$ is the same as in Proposition 5.2.1.

Chapter 6

Functional Convex Order for the McKean-Vlasov Equation

The aim of this section is to establish functional convex order results for d-dimensional scaled McKean-Vlasov equation, which extends results in Pagès (2016) obtained for one dimensional martingale diffusion, solution of stochastic differential equations of the form $dX_t = \sigma(t, X_t)dB_t$. The convex order result is also an direct application of the convergence of the theoretical Euler scheme proved in Chapter 5, even this scheme is not directly computable.

Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of all probability distributions on $\mathbb{R}^d$. Let σ, θ be two functions defined on $[0,T] \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)$ and valued in $\mathbb{M}_{d\times q}$. We define two McKean-Vlasov processes $(X_t)_{t\in[0,T]}$ and $(Y_t)_{t\in[0,T]}$ by

$$dX_t = (\alpha X_t + \beta)dt + \sigma(t, X_t, \mu_t)dB_t, \quad X_0 \in L^p(\mathbb{P}), \tag{6.0.1}$$

$$dY_t = (\alpha\, Y_t + \beta)dt + \theta(t,\, Y_t,\, \nu_t)\, dB_t, \quad Y_0 \in L^p(\mathbb{P}), \tag{6.0.2}$$

where $p \geq 1$ α, $\beta \in \mathbb{R}$ and for any $t \in [0,T]$, μ_t and ν_t respectively denote the probability distribution of X_t and Y_t. The main goal of this section is to prove if σ and θ are ordered for some matrix order, then the process $(X_t)_{t\in[0,T]}$ and $(Y_t)_{t\in[0,T]}$ defined in (6.0.1), (6.0.2) are accordingly ordered for the functional convex order. To be more precise, let us first recall the definition of convex order for two $\mathbb{R}^d$-valued random variables U and V and generalize this definition to two probability distributions μ, ν on $\big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big)$.

Definition 6.0.1. *(i) Let $U, V : (\Omega, \mathcal{F}, \mathbb{P}) \to \big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big)$ be two random variables. We call U is dominated by V for the convex order - denoted by $U \preceq_{cv} V$ - if for any convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbb{E}\,\varphi(U) \leq \mathbb{E}\,\varphi(V),$$

as soon as these two expectations make sense in $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$.

(ii) Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. *We call the distribution* μ *is dominated by* ν *for the convex order - denoted by* $\mu \preceq_{cv} \nu$ *- if for every convex function* $\varphi : \mathbb{R}^d \to \mathbb{R}$,

$$\int_{\mathbb{R}^d} \varphi(\xi)\mu(d\xi) \leq \int_{\mathbb{R}^d} \varphi(\xi)\nu(d\xi),$$

as soon as these two integrals make sense in $\overline{\mathbb{R}}$.

If we denote by $P_X = \mathbb{P} \circ X^{-1}$ the probability distribution of a random variable X, it is obvious that if $X \preceq_{cv} Y$, then $P_X \preceq_{cv} P_Y$ and *vice versa*.

We define a *partial order* between matrices in $\mathbb{M}_{d\times q}$ as follows:

$$\forall A, B \in \mathbb{M}_{d\times q}, \text{ we write } A \preceq B, \text{ if } BB^* - AA^* \text{ is a positive semi-definite matrix.} \tag{6.0.3}$$

Moreover, we assume that X_0, Y_0, σ and θ in (6.0.1) and (6.0.2) satisfy the following conditions:

Assumption (III): *(i) For every fixed* $t \in \mathbb{R}_+, \mu \in \mathcal{P}(\mathbb{R}^d)$, $\sigma(t, \cdot, \mu)$ *is convex in* x *in the sense that*

$$\forall\, x, y \in \mathbb{R}^d, \forall \lambda \in [0,1], \quad \sigma(t, \lambda x + (1-\lambda)y, \mu) \preceq \lambda\sigma(t, x, \mu) + (1-\lambda)\sigma(t, y, \mu). \tag{6.0.4}$$

(ii) For every fixed $t \in \mathbb{R}_+, x \in \mathbb{R}^d$, $\sigma(t, x, \cdot)$ *is non-decreasing in* μ *with respect to the convex order, that is,*

$$\forall\, \mu, \nu \in \mathcal{P}(\mathbb{R}^d) \quad \mu \preceq_{cv} \nu, \qquad \Longrightarrow \qquad \sigma(t, x, \mu) \preceq \sigma(t, x, \nu). \tag{6.0.5}$$

(iii) For every $(t, x, \mu) \in \mathbb{R}_+ \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)$,

$$\sigma(t, x, \mu) \preceq \theta(t, x, \mu). \tag{6.0.6}$$

(iv) $X_0 \preceq_{cv} Y_0$.

The main theorem of this section is the following

Theorem 6.0.1. *Let* $X := (X_t)_{t\in[0,T]}$, $Y := (Y_t)_{t\in[0,T]}$ *respectively denote the solution of McKean-Vlasov equations (6.0.1) and (6.0.2). Assume that the equations (6.0.1) and (6.0.2) satisfy Assumption (I), (II) and (III). Then for any convex function* F :

$\mathcal{C}([0,T],\mathbb{R}^d)\to\mathbb{R}$ *with* $(r,\|\cdot\|_{\sup})$*-polynomial growth,* $1\le r\le p$ *in the sense that*

$$\forall\alpha\in\mathcal{C}([0,T],\mathbb{R}^d),\ \textit{there exists } C>0,\quad s.t.\quad |F(\alpha)|\le C\big(1+\|\alpha\|_{\sup}^r\big),\tag{6.0.7}$$

one has

$$\mathbb{E}\,F(X)\le\mathbb{E}\,F(Y).\tag{6.0.8}$$

Let $M\in\mathbb{N}^*$ and let $h=\frac{T}{M}$. For $m=0,...,M$, we write $t_m^M:=h\cdot m=\frac{T}{M}\cdot m$[(1)]. The Euler schemes of $(X_t)_{t\in[0,T]}$ and $(Y_t)_{t\in[0,T]}$ are

$$\begin{cases}\bar X^M_{t_{m+1}}=\bar X^M_{t_m}+h\cdot(\alpha\bar X^M_{t_m}+\beta)+\sqrt{h}\cdot\sigma(t_m^M,\bar X^M_{t_m},\bar\mu^M_{t_m})Z_{m+1}, & \bar X_0=X_0,\\ \bar Y^M_{t_{m+1}}=\bar Y^M_{t_m}+h\cdot(\alpha\bar Y^M_{t_m}+\beta)+\sqrt{h}\cdot\theta(t_m^M,\bar Y^M_{t_m},\bar\nu^M_{t_m})Z_{m+1}, & \bar Y_0=Y_0,\end{cases}\tag{6.0.9}$$

where $Z_m, m=1,...,M,\overset{i.i.d}{\sim}\mathcal{N}(0,\mathbf{I}_q)$ and $\bar\mu^M_{t_m},\bar\nu^M_{t_m}$ respectively denote the probability distribution of $\bar X^M_{t_m}$ and $\bar Y^M_{t_m}$, $m=0,...,M$.

We first show the functional convex order for the Euler scheme $\bar X^M_{t_m}$ and $\bar Y^M_{t_m}$ in Section 6.1 by proving

$$\mathbb{E}\,F(\bar X^M_{t_0},...,\bar X^M_{t_M})\le\mathbb{E}\,F(\bar Y^M_{t_0},...,\bar Y^M_{t_M})\tag{6.0.10}$$

for any convex function $F:(\mathbb{R}^d)^{M+1}\to\mathbb{R}$ with r-polynomial growth, $1\le r\le p$. Next, based on the convergence of the theoretical Euler scheme established in Section 5.2, we derive the functional convex order result (6.0.8) from (6.0.10) by letting $M\to+\infty$.

6.1 Convex order for the Euler scheme

In order to simplify the notations, we rewrite the Euler scheme defined by (6.0.9) by letting $\bar X_m:=\bar X^M_{t_m}$, $\bar Y_m:=\bar Y^M_{t_m}$, $\bar\mu_m:=\bar\mu^M_{t_m}$ and $\bar\nu_m:=\bar\nu^M_{t_m}$ as follows,

$$\bar X_{m+1}=\bar\alpha\bar X_m+\bar\beta+\sigma_m(\bar X_m,\bar\mu_m)Z_{m+1},\qquad \bar X_0=X_0,\tag{6.1.1}$$

$$\bar Y_{m+1}=\bar\alpha\bar Y_m+\bar\beta+\theta_m(\bar Y_m,\bar\nu_m)Z_{m+1},\qquad \bar Y_0=Y_0,\tag{6.1.2}$$

where $\bar\alpha=\alpha h+1$, $\bar\beta=\beta h$, and for every $m=0,...,M$,

$$\sigma_m(x,\mu):=\sqrt{h}\cdot\sigma(t_m,x,\mu),\quad \theta_m(x,\mu):=\sqrt{h}\cdot\theta(t_m,x,\mu).$$

Then it follows from Assumption (III) that X_0, Y_0, $\sigma_m,\theta_m,m=0,...,M$, satisfy the following conditions.

(1) When there is no ambiguity, we write t_m instead of t_m^M.

Assumption (III'): (i) *Convex in* x *:*

$$\forall x, y \in \mathbb{R}^d, \forall \lambda \in [0,1], \qquad \sigma_m\big(\lambda x + (1-\lambda)y, \mu\big) \preceq \lambda \sigma_m(x,\mu) + (1-\lambda)\sigma_m(y,\mu). \tag{6.1.3}$$

(ii) *Non-decreasing in* μ *with respect to the convex order:*

$$\forall \mu, \nu \in \mathcal{P}(\mathbb{R}^d), \mu \preceq_{cv} \nu, \qquad \sigma_m(x,\mu) \preceq \sigma_m(x,\nu). \tag{6.1.4}$$

(iii) *Order of* σ_m *and* θ_m*:*

$$\forall (x,\mu) \in \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d), \qquad \sigma_m(x,\mu) \preceq \theta_m(x,\mu). \tag{6.1.5}$$

(iv) $\bar{X}_0 \preceq_{cv} \bar{Y}_0$.

The main result of this section is the following proposition.

Proposition 6.1.1. *Under Assumption (III), for any convex function* $F : (\mathbb{R}^d)^{M+1} \to \mathbb{R}$ *with* r*-polynomial growth,* $1 \le r \le p$*, in the sense that*

$$\forall x = (x_0, ..., x_M) \in (\mathbb{R}^d)^{M+1}, \exists\, C > 0, \quad \textit{such that} \quad |F(x)| \le C\big(1 + \sup_{0 \le i \le M} |x_i|^r\big), \tag{6.1.6}$$

we have

$$\mathbb{E}\, F(\bar{X}_0, ..., \bar{X}_M) \le \mathbb{E}\, F(\bar{Y}_0, ..., \bar{Y}_M).$$

The proof of Proposition 6.1.1 relies on the following two lemmas.

Lemma 6.1.1 (see Jourdain and Pagès (2019) and Fadili (2019))**.** *Let* $Z \sim \mathcal{N}(0, \mathbf{I}_q)$*. If* $u_1, u_2 \in \mathbb{M}_{d\times q}$ *with* $u_1 \preceq u_2$*, then* $u_1 Z \preceq_{cv} u_2 Z$*.*

Proof. We define $M_1 \coloneqq u_1 Z$ and $M_2 \coloneqq M_1 + \sqrt{u_2 u_2^* - u_1 u_1^*} \cdot \widetilde{Z}$, where $\sqrt{A}$ denotes the square root of a positive semi-definite matrix A and $\widetilde{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$, $\widetilde{Z}$ is independent to Z. Hence the probability distribution of M_2 is $\mathcal{N}(0, u_2 u_2^*)$, which is the same distribution as $u_2 Z$.

For any convex function φ such that $\mathbb{E}\,\varphi(M_1)$ and $\mathbb{E}\,\varphi(M_2)$ make sense, we have

$$\begin{aligned}
\mathbb{E}\left[\varphi(M_2)\right] &= \mathbb{E}\left[\varphi\big(M_1 + \sqrt{u_2 u_2^* - u_1 u_1^*} \cdot \widetilde{Z}\big)\right] \\
&= \mathbb{E}\left[\ \mathbb{E}\left[\varphi\big(M_1 + \sqrt{u_2 u_2^* - u_1 u_1^*} \cdot \widetilde{Z}\big) \mid Z\right]\ \right] \\
&\ge \mathbb{E}\left[\varphi\big(\mathbb{E}\left[M_1 + \sqrt{u_2 u_2^* - u_1 u_1^*} \cdot \widetilde{Z} \mid Z\right]\big)\ \right] \\
&= \mathbb{E}\left[\varphi\big(M_1 + \mathbb{E}\big[\sqrt{u_2 u_2^* - u_1 u_1^*} \cdot \widetilde{Z}\big]\big)\right] = \mathbb{E}\,\varphi(M_1).
\end{aligned} \tag{6.1.7}$$

Hence, $u_1 Z \preceq_{cv} u_2 Z$ owing to the equivalence of convex order of the random variable and its probability distribution (see Definition 6.0.1). □

Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a Borel convex function. We define an operator $Q : \mathcal{C}(\mathbb{R}^d, \mathbb{R}) \to \mathcal{C}(\mathbb{R}^d \times \mathbb{M}_{d\times q}, \mathbb{R})$ associated to a random variable $Z \sim \mathcal{N}(0, \mathbf{I}_q)$ by

$$(x, u) \in \mathbb{R}^d \times \mathbb{M}_{d\times q} \mapsto (Q\varphi)(x, u) := \mathbb{E}\,\varphi(\bar{\alpha}x + \bar{\beta} + uZ) \tag{6.1.8}$$

The following lemma is a generalisation to dimension d of Pagès (2016)[Lemma 2.1].

Lemma 6.1.2 (Revisited Jensen's Lemma)**.** *Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then,*

(i) The function $Q\varphi$ defined by (6.1.8) is convex.

(ii) For any fixed $x \in \mathbb{R}^d$, the function $Q\varphi(x, \cdot)$ reaches its minimum at $\mathbf{0}_{d\times q}$, where $\mathbf{0}_{d\times q}$ is the zero-matrix of size $d \times q$.

(iii) The function $Q\varphi(x, \cdot)$ is non-decreasing in u with respect to the partial order of $d \times q$ matrix (6.0.3).

Proof. *(i)* For every $(x_1, u_1), (x_2, u_2) \in \mathbb{R}^d \times \mathbb{M}_{d\times q}$ and $\lambda \in [0, 1]$,

$$\begin{aligned}
&Q\varphi\big(\lambda(x_1, u_1) + (1-\lambda)(x_2, u_2)\big) \\
&\quad = \mathbb{E}\left[\varphi\Big(\bar{\alpha}\big(\lambda x_1 + (1-\lambda)x_2\big) + \bar{\beta} + \big(\lambda u_1 + (1-\lambda)u_2\big)Z\Big)\right] \\
&\quad = \mathbb{E}\left[\varphi\Big(\lambda\big(\bar{\alpha}x_1 + \bar{\beta}\big) + (1-\lambda)\big(\bar{\alpha}x_2 + \bar{\beta}\big) + \lambda u_1 Z + (1-\lambda)u_2 Z\Big)\right] \\
&\quad = \lambda\mathbb{E}\left[\varphi\big(\bar{\alpha}x_1 + \bar{\beta} + u_1 Z\big)\right] + (1-\lambda)\mathbb{E}\left[\varphi\big(\bar{\alpha}x_2 + \bar{\beta} + u_2 Z\big)\right] \\
&\qquad \text{(by the convexity of } \varphi \text{ and linearity of the expectation)} \\
&\quad = \lambda Q\varphi(x_1, u_1) + (1-\lambda)Q\varphi(x_2, u_2).
\end{aligned}$$

Hence, $Q\varphi$ is a convex function.

(ii) If we fix an $x \in \mathbb{R}^d$, then for any $u \in \mathbb{M}_{d\times q}$,

$$\begin{aligned}
Q\varphi(x, u) = \mathbb{E}\left[\varphi(\bar{\alpha}x + \bar{\beta} + uZ)\right] &\geq \varphi\big(\mathbb{E}\left[\bar{\alpha}x + \bar{\beta} + uZ\right]\big) \\
&= \varphi\big(\bar{\alpha}x + \bar{\beta} + \mathbf{0}_{d\times 1}\big) = Q\varphi(x, \mathbf{0}_{d\times q}).
\end{aligned} \tag{6.1.9}$$

(iii) For a fixed $x \in \mathbb{R}^d$, it is obvious that $\varphi(\bar{\alpha}x + \bar{\beta} + \cdot)$ is also a convex function. Thus, Lemma 6.1.1 directly implies that if $u_1 \preceq u_2$, then $\mathbb{E}\,\varphi(\bar{\alpha}x + \bar{\beta} + u_1 Z) \leq \mathbb{E}\,\varphi(\bar{\alpha}x + \bar{\beta} + u_2 Z)$, which is equivalent to $Q\varphi(x, u_1) \leq Q\varphi(x, u_2)$. □

Before proving Proposition 6.1.1, we first show in the next section by a forward induction that the Euler scheme defined in (6.1.1) and (6.1.2) propagates the marginal convex order step by step, i.e. $\bar{X}_m \preceq_{cv} \bar{Y}_m,\ m = 0, ..., M$.

6.1.1 Marginal convex order

Let $Z_m, m = 1, ..., M$ be i.i.d random variable with distributions $\mathcal{N}(0, \mathbf{I}_q)$ in the definition of Euler scheme (6.1.1) and (6.1.2). For every $m = 1, ..., M$, we define an operator $Q_m : \mathcal{C}(\mathbb{R}^d, \mathbb{R}) \to \mathcal{C}\big(\mathbb{R}^d \times \mathbb{M}_{d\times q}, \mathbb{R}\big)$ associated with Z_m by

$$(x, u) \in \mathbb{R}^d \times \mathbb{M}_{d\times q} \mapsto (Q_m\,\varphi)(x, u) := \mathbb{E}\big[\varphi(\bar{\alpha}x + \bar{\beta} + uZ_m)\big]. \tag{6.1.10}$$

For every $m = 0, ..., M$, let $\mathcal{F}_m$ denote the σ-algebra generated by $X_0, Z_1, ..., Z_m$.

Proposition 6.1.2. *Let $(\bar{X}_m)_{m=0,...,M}$, $(\bar{Y}_m)_{m=0,...,M}$ be random variables defined by (6.1.1) and (6.1.2). If for every $m = 0, ..., M$, σ_m and θ_m satisfy Assumption (III'), then*

$$\bar{X}_m \preceq_{cv} \bar{Y}_m, m = 0, ..., M.$$

The proof of Proposition 6.1.2 relies on the following lemma.

Lemma 6.1.3. *If $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a convex function, then for a fixed $\mu \in \mathcal{P}(\mathbb{R}^d)$, the function $x \mapsto \mathbb{E}\big[\varphi\big(\bar{\alpha}x + \bar{\beta} + \sigma_m(x, \mu)Z_m\big)\big]$ is also convex, $m = 0, ..., M$.*

Proof of Lemma 6.1.3. Let $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$. For every $m = 0, ..., M$, we have

$$\begin{aligned}
&\mathbb{E}\left[\varphi\Big(\bar{\alpha}\big(\lambda x + (1-\lambda)y\big) + \bar{\beta} + \sigma_m\big(\lambda x + (1-\lambda)y, \mu\big)Z_m\Big)\right] \\
&\quad\le \mathbb{E}\left[\varphi\Big(\lambda\big(\bar{\alpha}x + \bar{\beta}\big) + (1-\lambda)\big(\bar{\alpha}y + \bar{\beta}\big) + \lambda\sigma_m(x, \mu)Z_m + (1-\lambda)\sigma_m(y, \mu)Z_m\Big)\right] \\
&\qquad\text{(by Assumption (6.1.3) and Lemma 6.1.2)} \\
&\quad\le \lambda\,\mathbb{E}\left[\varphi\big(\bar{\alpha}x + \bar{\beta} + \sigma_m(x, \mu)Z_m\big)\right] + (1-\lambda)\,\mathbb{E}\left[\varphi\big(\bar{\alpha}y + \bar{\beta} + \sigma_m(y, \mu)Z_m\big)\right] \\
&\qquad\text{(by the convexity of } \varphi).
\end{aligned}$$

□

Proof of Proposition 6.1.2. Assumption (III') directly implies $X_0 \preceq_{cv} Y_0$.

Assume that $X_m \preceq_{cv} Y_m$, then for any convex function φ such that $\mathbb{E}\,\varphi(\bar{X}_{m+1})$ and $\mathbb{E}\,\varphi(\bar{Y}_{m+1})$ make sense,

$$\mathbb{E}\left[\varphi(\bar{X}_{m+1})\right] = \mathbb{E}\left[\varphi\big(\bar{\alpha}\bar{X}_m + \bar{\beta} + \sigma_m(\bar{X}_m, \bar{\mu}_m)Z_{m+1}\big)\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\mathbb{E}\left[\varphi(\bar{\alpha}\bar{X}_m + \bar{\beta} + \sigma_m(\bar{X}_m, \bar{\mu}_m)Z_{m+1}) \mid \mathcal{F}_m\right]\right] \\
&= \int_{\mathbb{R}^d} \bar{\mu}_m(dx)\mathbb{E}\left[\varphi(\bar{\alpha}x + \bar{\beta} + \sigma_m(x, \bar{\mu}_m)Z_{m+1})\right] \\
&\leq \int_{\mathbb{R}^d} \bar{\mu}_m(dx)\mathbb{E}\left[\varphi(\bar{\alpha}x + \bar{\beta} + \sigma_m(x, \bar{\nu}_m)Z_{m+1})\right] \\
&\qquad \text{(by Lemma 6.1.2 and Assumption (6.1.4), since } \bar{\mu}_m \preceq_{cv} \bar{\nu}_m \text{)} \\
&\leq \int_{\mathbb{R}^d} \bar{\nu}_m(dx)\mathbb{E}\left[\varphi(\bar{\alpha}x + \bar{\beta} + \sigma_m(x, \bar{\nu}_m)Z_{m+1})\right] \\
&\qquad \text{(by Lemma 6.1.3, since } \bar{\mu}_m \preceq_{cv} \bar{\nu}_m\text{)} \\
&\leq \int_{\mathbb{R}^d} \bar{\nu}_m(dx)\mathbb{E}\left[\varphi(\bar{\alpha}x + \bar{\beta} + \theta_m(x, \bar{\nu}_m)Z_{m+1})\right] \\
&\qquad \big(\text{by Lemma 6.1.2 and Assumption (6.1.5)}\big) \\
&= \mathbb{E}\left[\varphi(\bar{Y}_{m+1})\right].
\end{aligned}
$$

Thus one concludes by a forward induction. □

6.1.2 Global functional convex order

The main goal of this section is to prove Proposition 6.1.1. For any $m_1, m_2 \in \mathbb{N}^*$ with $m_1 \leq m_2$, we denote by $x_{m_1:m_2} := (x_{m_1}, x_{m_1+1}, ..., x_{m_2}) \in (\mathbb{R}^d)^{m_2-m_1+1}$. Similarly, we denote by $\mu_{m_1:\, m_2} := (\mu_{m_1}, ..., \mu_{m_2}) \in \big(\mathcal{P}(\mathbb{R}^d)\big)^{m_2-m_1+1}$. We recursively define a function sequence $\Phi_m : (\mathbb{R}^d)^{m+1} \times \big(\mathcal{P}(\mathbb{R}^d)\big)^{M-m+1} \to \mathbb{R}, m = 0, ..., M$ as follows.

Set

$$
\Phi_M(x_{0:M}, \mu_M) := F(x_0, ..., x_M), \tag{6.1.11}
$$

where the function F is the same as in Proposition 6.1.1. For $m = 0, ..., M-1$, set

$$
\begin{aligned}
\Phi_m(x_{0:m}, \mu_{m:M}) &:= \big(Q_{m+1}\Phi_{m+1}(x_{0:m}, \cdot\,, \mu_{m+1:M})\big)\big(x_m, \sigma_m(x_m, \mu_m)\big) \\
&= \mathbb{E}\left[\Phi_{m+1}\big(x_{0:m}, \bar{\alpha}x_m + \bar{\beta} + \sigma_m(x_m, \mu_m)Z_{m+1}, \mu_{m+1:M}\big)\right].
\end{aligned} \tag{6.1.12}
$$

The functions $\Phi_m, m = 0, ..., M$ have the following properties.

Lemma 6.1.4. *For every $m = 0, ..., M$,*

(i) for a fixed $\mu_{m:M} \in \big(\mathcal{P}(\mathbb{R}^d)\big)^{M-m+1}$, the function $\Phi_m(\,\cdot\,, \mu_{m:M})$ is convex in $x_{0:m}$,

(ii) for a fixed $x_{0:m} \in (\mathbb{R}^d)^{m+1}$, the function $\Phi_m(x_{0:m}, \,\cdot\,)$ is non-decreasing in $\mu_{m:M}$ with respect to the convex order in the sense that for any $\mu_{m:M}, \nu_{m:M} \in \big(\mathcal{P}(\mathbb{R}^d)\big)^{M-m+1}$ such that $\mu_i \preceq_{cv} \nu_i, i = m, ..., M$,

$$
\Phi_m(x_{0:m}, \mu_{m:M}) \leq \Phi_m(x_{0:m}, \nu_{m:M}). \tag{6.1.13}
$$

Proof. (i) The function Φ_M is convex in $x_{0:M}$ owing to the hypotheses on F. Now assume that Φ_{m+1} is convex in $x_{0:m+1}$. For any $x_{0:m}, y_{0:m} \in (\mathbb{R}^d)^{m+1}$ and $\lambda \in [0,1]$, it follows that

$$
\begin{aligned}
&\Phi_m\big(\lambda x_{0:m} + (1-\lambda)y_{0:m}, \mu_{m:M}\big) \\
&\quad = \mathbb{E}\,\Phi_{m+1}\Big(\lambda x_{0:m} + (1-\lambda)y_{0:m}, \bar{\alpha}\big(\lambda x_m + (1-\lambda)y_m\big) + \bar{\beta} \\
&\qquad\qquad + \sigma_m(\lambda x_m + (1-\lambda)y_m, \mu_m) Z_{m+1}, \mu_{m+1:M}\Big) \\
&\quad \le \mathbb{E}\,\Phi_{m+1}\Big(\lambda x_{0:m} + (1-\lambda)y_{0:m}, \lambda(\bar{\alpha}x_m + \bar{\beta}) + (1-\lambda)\cdot(\bar{\alpha}y_m + \bar{\beta}) \\
&\qquad\qquad + \big[\lambda\sigma_m(x_m, \mu_m) + (1-\lambda)\sigma_m(y_m, \mu_m)\big] Z_{m+1}, \mu_{m+1:M}\Big) \\
&\qquad \text{(by the Assumption (6.1.3) and Lemma 6.1.2} \\
&\qquad\quad \text{since } \Phi_{m+1}(x_{0:m}, \cdot, \mu_{m+1:M}) \text{ is a convex function)} \\
&\quad \le \lambda\mathbb{E}\,\Phi_{m+1}\big(x_{0:m}, \bar{\alpha}x_m + \bar{\beta} + \sigma(x_m, \mu_m) Z_{m+1}, \mu_{m+1:M}\big) \\
&\qquad + (1-\lambda)\mathbb{E}\,\Phi_{m+1}\big(y_{0:m}, \bar{\alpha}y_m + \bar{\beta} + \sigma(y_m, \mu_m) Z_{m+1}, \mu_{m+1:M}\big) \\
&\qquad \text{(since } \Phi_{m+1}(x_{0:m}, \cdot, \mu_{m+1:M}) \text{ is a convex function)} \\
&\quad = \lambda\Phi_m(x_{0:m}, \mu_{m:M}) + (1-\lambda)\Phi_m(y_{0:m}, \mu_{m:M}).
\end{aligned}
$$

Thus one concludes by a backward induction.

(ii) Firstly, it is obvious that for any $\mu_M, \nu_M \in \mathcal{P}(\mathbb{R}^d)$ such that $\mu_M \preceq_{cv} \nu_M$, we have

$$\Phi_M(x_{0:M}, \mu_M) = F(x_{0:M}) = \Phi_M(x_{0:M}, \nu_M).$$

Assume that $\Phi_{m+1}(x_{0:m+1}, \cdot)$ increases with respect to the convex order of $\mu_{m+1:M}$. For any $\mu_{m:M}, \nu_{m:M} \in \big(\mathcal{P}(\mathbb{R}^d)\big)^{M-m+1}$ such that $\mu_i \preceq_{cv} \nu_i, i = m, ..., M$, we have

$$
\begin{aligned}
&\Phi_m(x_{0:m}, \mu_{m:M}) = \mathbb{E}\Big[\Phi_{m+1}\big(x_{0:m}, \bar{\alpha}x_m + \bar{\beta} + \sigma_m(x_m, \mu_m) Z_{m+1}, \mu_{m+1:M}\big)\Big] \\
&\le \mathbb{E}\Big[\Phi_{m+1}\big(x_{0:m}, \bar{\alpha}x_m + \bar{\beta} + \sigma_m(x_m, \nu_m) Z_{m+1}, \mu_{m+1:M}\big)\Big] \\
&\text{(by Assumption (6.1.4) and Lemma 6.1.2 since } \Phi_{m+1}(x_{0:m}, \cdot, \mu_{m+1:M}) \text{ is a convex function)} \\
&\le \mathbb{E}\Big[\Phi_{m+1}\big(x_{0:m}, \bar{\alpha}x_m + \bar{\beta} + \sigma_m(x_m, \nu_m) Z_{m+1}, \nu_{m+1:M}\big)\Big] \quad \text{(by the assumption on } \Phi_{m+1}) \\
&= \Phi_m(x_{0:m}, \nu_{m:M}).
\end{aligned}
$$

We can conclude by a backward induction. □

As F has an r-polynomial growth, then the integrability of

$$F(\bar{X}_0, ..., \bar{X}_M) \text{ and } F(\bar{Y}_0, ..., \bar{Y}_M)$$

is guaranteed by Lemma 5.2.2 since $\|\bar{X}_0\|_r = \|\bar{Y}_0\|_r < +\infty$ as $X_0, Y_0 \in L^p(\mathbb{P})$, $p \geq r$. We define for every $m = 0, ..., M$,

$$\mathcal{X}_m := \mathbb{E}\left[F(\bar{X}_0, ..., \bar{X}_M) \mid \mathcal{F}_m\right].$$

Recall that $\bar{\mu}_m = P_{\bar{X}_m}, m = 0, ..., M$.

Lemma 6.1.5. *For every $m = 0, ..., M$, $\Phi_m(\bar{X}_{0:m}, \bar{\mu}_{m:M}) = \mathcal{X}_m$.*

Proof. It is obvious that

$$\Phi_M(\bar{X}_{0:M}, \bar{\mu}_M) = F(\bar{X}_0, ..., \bar{X}_M) =: \mathcal{X}_M.$$

Assume that $\Phi_{m+1}(\bar{X}_{0:m+1}, \bar{\mu}_{m+1:M}) = \mathcal{X}_{m+1}$. Then

$$\begin{aligned}
\mathcal{X}_m &= \mathbb{E}\left[\mathcal{X}_{m+1} \mid \mathcal{F}_m\right] = \mathbb{E}\left[\Phi_{m+1}(\bar{X}_{0:m+1}, \bar{\mu}_{m+1:M}) \mid \mathcal{F}_m\right] \\
&= \mathbb{E}\left[\Phi_{m+1}(\bar{X}_{0:m}, \bar{\alpha}\bar{X}_m + \bar{\beta} + \sigma_m(\bar{X}_m, \bar{\mu}_m)Z_{m+1}, \bar{\mu}_{m+1:M}) \mid \mathcal{F}_m\right] \\
&= \left(Q_{m+1}\Phi_{m+1}(\bar{X}_{0:m}, \cdot, \bar{\mu}_{m+1:M})\right)\left(\bar{X}_m, \sigma_m(\bar{X}_m, \bar{\mu}_m)\right) \\
&= \Phi_M(\bar{X}_{0:m}, \bar{\mu}_{m:M}).
\end{aligned}$$

We conclude by a backward induction. □

Similarly, we define $\Psi_m : (\mathbb{R}^d)^{m+1} \times \left(\mathcal{P}(\mathbb{R}^d)\right)^{M-m+1} \to \mathbb{R}$, $m = 0, ..., M$ by

$$\begin{aligned}
\Psi_M(x_{0:M}, \mu_M) &:= F(x_{0:M}) \\
\Psi_m(x_{0:m}, \mu_{m:M}) &:= \left(Q_{m+1}\Psi_{m+1}(x_{0:m}, \cdot, \mu_{m+1:M})\right)\left(x_m, \theta_m(x_m, \mu_m)\right) \\
&= \mathbb{E}\left[\Psi_{m+1}\left(x_{0:m}, \bar{\alpha}x_m + \bar{\beta} + \theta_m(x_m, \mu_m)Z_{m+1}, \mu_{m+1:M}\right)\right]. \quad (6.1.14)
\end{aligned}$$

Recall that $\bar{\nu}_m := P_{\bar{Y}_m}$. It follows from the same reasoning as in Lemma 6.1.5 that

$$\Psi_m(\bar{Y}_0, ..., \bar{Y}_m, \bar{\nu}_m, ..., \bar{\nu}_M) = \mathbb{E}\left[F(\bar{Y}_0, ..., \bar{Y}_m) \mid \mathcal{F}_m\right].$$

Proof of Proposition 6.1.1. We first prove, this time by a backward induction that for every $m = 0, ..., M$, $\Phi_m \leq \Psi_m$.

It follows from the definition of Φ_M and Ψ_M that $\Phi_M = \Psi_M$. Assume $\Phi_{m+1} \leq \Psi_{m+1}$. Then for any $x_{0:m} \in (\mathbb{R}^d)^{m+1}$ and $\mu_{m:M} \in \left(\mathcal{P}(\mathbb{R}^d)\right)^{M-m+1}$, we have

$$\begin{aligned}
&\Phi_m(x_{0:m}, \mu_{m:M}) \\
&= \mathbb{E}\left[\Phi_{m+1}\left(x_{0:m}, \bar{\alpha}x_m + \bar{\beta} + \sigma_m(x_m, \mu_m)Z_{m+1}, \mu_{m+1:M}\right)\right] \\
&\leq \mathbb{E}\left[\Phi_{m+1}\left(x_{0:m}, \bar{\alpha}x_m + \bar{\beta} + \theta_m(x_m, \mu_m)Z_{m+1}, \mu_{m+1:M}\right)\right]
\end{aligned}$$

(by Assumption (6.1.5) and Lemma 6.1.2,
since Lemma 6.1.4 shows that Φ_{m+1} is convex in $x_{0:m+1}$)
$$\leq \mathbb{E}\left[\Psi_{m+1}\big(x_{0:m}, \bar{\alpha}x_m + \bar{\beta} + \theta_m(x_m, \mu_m)Z_{m+1}, \mu_{m+1:M}\big)\right] = \Psi(x_{0:m}, \mu_{m:M}).$$

Thus, the backward induction is completed and

$$\forall m = 0, ..., M, \qquad \Phi_m \leq \Psi_m. \tag{6.1.15}$$

Consequently,

$$\begin{aligned}
\mathbb{E}\left[F(\bar{X}_0, ..., \bar{X}_M)\right] &= \mathbb{E}\,\Phi_0(\bar{X}_0, \bar{\mu}_{0:M}) \\
&= \mathbb{E}\,\Phi_0(\bar{Y}_0, \bar{\mu}_{0:M}) \qquad \text{(by Lemma 6.1.4-}(i)\text{ since } \bar{X}_0 \preceq_{cv} \bar{Y}_0) \\
&\leq \mathbb{E}\,\Phi_0(\bar{Y}_0, \bar{\nu}_{0:M}) \qquad \text{(by Lemma 6.1.4-}(ii)\text{ and Proposition 6.1.2)} \\
&\leq \mathbb{E}\,\Psi_0(\bar{Y}_0, \bar{\nu}_{0:M}) \qquad \text{(by (6.1.15))} \\
&= \mathbb{E}\left[F(\bar{Y}_0, ..., \bar{Y}_M)\right], \qquad (6.1.16)
\end{aligned}$$

owing to the martingale property. □

6.2 Functional convex order for the McKean-Vlasov process

This section is devoted to prove Theorem 6.0.1. Recall that $t_m^M = m \cdot \frac{T}{M}, m = 0, ..., M$. We define two interpolators as follows.

Definition 6.2.1. (*i*) *For every integer $M \geq 1$, we define the piecewise affine interpolator $i_M : x_{0:M} \in (\mathbb{R}^d)^{M+1} \mapsto i_M(x_{0:M}) \in \mathcal{C}([0,T], \mathbb{R}^d)$ by*

$$\forall\, m = 0, ..., M-1,\ \forall\, t \in [t_m^M, t_{m+1}^M],$$
$$i_M(x_{0:M})(t) = \frac{M}{T}\left[(t_{m+1}^M - t)x_m + (t - t_m^M)x_{m+1}\right].$$

(*ii*) *For every $M \geq 1$, we define the functional interpolator $I_M : \mathcal{C}\big([0,T], \mathbb{R}^d\big) \to \mathcal{C}\big([0,T], \mathbb{R}^d\big)$ by*

$$\forall \alpha \in \mathcal{C}([0,T], \mathbb{R}^d), \qquad I_M(\alpha) = i_M\big(\alpha(t_0^M), ..., \alpha(t_M^M)\big).$$

It is obvious that

$$\forall x_{0:M} \in (\mathbb{R}^d)^{M+1}, \qquad \|i_M(x_{0:M})\|_{\sup} \leq \max_{0 \leq m \leq M} |x_m| \tag{6.2.1}$$

since the norm $|\cdot|$ is convex. Consequently,

$$\forall \alpha \in \mathcal{C}([0,T],\mathbb{R}^d), \qquad \|I_M(\alpha)\|_{\sup} \leq \|\alpha\|_{\sup}. \tag{6.2.2}$$

Moreover, for any $\alpha \in \mathcal{C}([0,T],\mathbb{R}^d)$, we have

$$\|I_M(\alpha) - \alpha\|_{\sup} \leq w(\alpha, \frac{T}{M}), \tag{6.2.3}$$

where w denotes the uniform continuity modulus of α. The proof of Theorem 6.0.1 relies on the following lemma.

Lemma 6.2.1 (Lemma 2.2 in Pagès (2016))**.** *Let $X^M, M \geq 1$ be a sequence of continuous processes weakly converging towards X as $M \to +\infty$ for the $\|\cdot\|_{sup}$-norm. Then the sequence of interpolating processes $\widetilde{X}^M = I_M(X^M), M \geq 1$ is weakly converging toward X for the $\|\cdot\|_{\sup}$-norm topology.*

Proof of Theorem 6.0.1. Let $M \in \mathbb{N}^*, h = \frac{T}{M}, t_m^M = m \cdot h = m \cdot \frac{T}{M}$. Let $(\bar{X}_{t_m}^M)_{m=0,\dots,M}$ and $(\bar{Y}_{t_m}^M)_{m=0,\dots,M}$ denote the Euler scheme defined in (6.0.9). Let $\bar{X}^M := (\bar{X}_t^M)_{t\in[0,T]}$, $\bar{Y}^M := (\bar{Y}_t^M)_{t\in[0,T]}$ (defined as follows) be the continuous Euler scheme of $(X_t)_{t\in[0,T]}$, $(Y_t)_{t\in[0,T]}$,

$$\begin{aligned}
&\forall m = 0, ..., M-1, \ \ \forall t \in [t_m, t_{m+1}),\\
&\qquad \bar{X}_t^M = \bar{X}_{t_m}^M + (\alpha \bar{X}_{t_m}^M + \beta)(t - t_m) + \sigma(t_m^M, \bar{X}_{t_m}^M, \bar{\mu}_{t_m}^M)(B_t - B_{t_m}), &(6.2.4)\\
&\qquad \bar{Y}_t^M = \bar{Y}_{t_m}^M + (\alpha \bar{Y}_{t_m}^M + \beta)(t - t_m) + \theta(t_m^M, \bar{Y}_{t_m}^M, \bar{\nu}_{t_m}^M)(B_t - B_{t_m}). &(6.2.5)
\end{aligned}$$

By Lemma 5.2.2, there exists a constant $\tilde{C}$ such that

$$\begin{aligned}
&\left\| \sup_{t\in[0,T]} \left|\bar{X}_t^M\right| \right\|_r \vee \left\| \sup_{t\in[0,T]} |X_t| \right\|_r \leq \tilde{C}(1 + \|X_0\|_r) = \tilde{C}(1 + \|X_0\|_p) < +\infty,\\
&\left\| \sup_{t\in[0,T]} \left|\bar{Y}_t^M\right| \right\|_r \vee \left\| \sup_{t\in[0,T]} |Y_t| \right\|_r \leq \tilde{C}(1 + \|Y_0\|_r) = \tilde{C}(1 + \|Y_0\|_p) < +\infty
\end{aligned} \tag{6.2.6}$$

as $1 \leq r \leq p$ and $X_0, Y_0 \in L^p(\mathbb{P})$. Hence, $F(X)$ and $F(Y)$ are in $L^1(\mathbb{P})$ since F has a r-polynomial growth.

We define a function $F_M : (\mathbb{R}^d)^{M+1} \to \mathbb{R}$ by

$$x_{0:M} \in (\mathbb{R}^d)^{M+1} \mapsto F_M(x_{0:M}) := F\big(i_M(x_{0:M})\big). \tag{6.2.7}$$

The function F_M is obviously convex since i_M is a linear application. Moreover, F_M has also an r-polynomial growth by (6.2.1).

Furthermore, we have $I_M(\bar{X}^M) = i_M\big((\bar{X}^M_{t_0}, ..., \bar{X}^M_{t_M})\big)$ by the definition of continuous Euler scheme and interpolators i_M and I_M, so that

$$F_M(\bar{X}^M_{t_0}, ..., \bar{X}^M_{t_M}) = F\Big(i_M\big((\bar{X}^M_{t_0}, ..., \bar{X}^M_{t_M})\big)\Big) = F\big(I_M(\bar{X}^M)\big).$$

It follows from Proposition 6.1.1 that

$$\begin{aligned}\mathbb{E}\, F\big(I_M(\bar{X}^M)\big) &= \mathbb{E}\, F\big(i_M(\bar{X}^M_0, ..., \bar{X}^M_M)\big) = \mathbb{E}\, F_M\big(\bar{X}^M_0, ..., \bar{X}^M_M\big)\\ &\le \mathbb{E}\, F_M\big(\bar{Y}^M_0, ..., \bar{Y}^M_M\big) = \mathbb{E}\, F\big(i_M(\bar{Y}^M_0, ..., \bar{Y}^M_M)\big) = \mathbb{E}\, F\big(I_M(\bar{Y}^M)\big). \qquad (6.2.8)\end{aligned}$$

The function F is $\|\cdot\|_{\sup}$ −continuous since it is convex with $\|\cdot\|_{\sup}$ −polynomial growth (see Lemma 2.1.1 in Lucchetti (2006)). Moreover the process $\bar{X}^M$ weakly converges to X as $M \to +\infty$ by Corollary 5.2.1. Then $I_M(\bar{X}^M)$ weakly converges to X by applying Lemma 6.2.1. Hence the inequality (6.2.8) implies that

$$\mathbb{E}\, F(X) \le \mathbb{E}\, F(Y),$$

by letting $M \to +\infty$ and by applying the Lebesgue dominated convergence theorem owing to (6.2.6) since F has a r-polynomial growth. □

Remark 6.2.1. The functional convex order result, in a general setting, can be used to establish a robust option price bound (see e.g. Alfonsi et al. (2019)). However, in the McKean-Vlasov setting, the functional convex order result Theorem 6.0.1, is established by using the theoretical Euler scheme (C) which is not directly computable so that there are still some work to do to produce simulatable approximations which are consistent for the convex order. In the next chapter, we propose the computable particle method for (6.0.1) and (6.0.2), which reads,

$$\begin{cases}\forall n \in \{1, ..., N\},\\ \bar{X}^{n,N}_{t_{m+1}} = \bar{X}^{n,N}_{t_m} + h(\alpha \bar{X}^{n,N}_{t_m} + \beta) + \sqrt{h}\sigma(\bar{X}^{n,N}_{t_m}, \bar{\mu}^N_{t_m}) Z^n_{m+1}, \text{ with } \bar{\mu}^N_{t_m} \coloneqq \frac{1}{N}\sum_{n=1}^N \delta_{\bar{X}^{n,N}_{t_m}},\\ \bar{Y}^{n,N}_{t_{m+1}} = \bar{Y}^{n,N}_{t_m} + h(\alpha \bar{Y}^{n,N}_{t_m} + \beta) + \sqrt{h}\theta(\bar{Y}^{n,N}_{t_m}, \bar{\nu}^N_{t_m}) Z^n_{m+1}, \text{ with } \bar{\nu}^N_{t_m} \coloneqq \frac{1}{N}\sum_{n=1}^N \delta_{\bar{Y}^{n,N}_{t_m}},\end{cases}$$

where $t_m = t^M_m \coloneqq m \cdot \frac{T}{M}$, $M \in \mathbb{N}^*$, $\bar{X}^{n,N}_0$ are i.i.d copies of X_0 and $\bar{Y}^{n,N}_0$ are i.i.d copies of Y_0.

Unfortunately, this scheme based on the particle method does not propagate nor preserve the convex order like in Proposition 6.1.2 since we cannot obtain for a convex function φ that,

$$\frac{1}{N}\sum_{n=1}^N \varphi\big(X^{n,N}_{t_m}(\omega)\big) \le \frac{1}{N}\sum_{n=1}^N \varphi\big(Y^{n,N}_{t_m}(\omega)\big), \quad a.s.$$

under the condition that $X_{t_m}^{n,N} \preceq_{cv} Y_{t_m}^{n,N}$, $n = 1, ..., N$, even if the random variables $X_{t_m}^{n,N}$, $n = 1, ..., N$ and $Y_{t_m}^{n,N}$, $n = 1, ..., N$ were both i.i.d. (see the same paper Alfonsi et al. (2019)).

6.3 Extension of the functional convex order result

In this section, we will extend the result of Theorem 6.0.1 to functionals of both the path of process and its marginal distributions. If we consider a function

$$G : \big(\alpha, (\gamma_t)_{t\in[0,T]}\big) \in \mathcal{C}\big([0,T], \mathbb{R}^d\big) \times \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big) \mapsto G\big(\alpha, (\gamma_t)_{t\in[0,T]}\big) \in \mathbb{R}$$

satisfying the following conditions:

(*i*) G is convex in α,

(*ii*) G has an r-polynomial growth, $1 \le r \le p$, in the sense that

$$\begin{aligned}&\forall \big(\alpha, (\gamma_t)_{t\in[0,T]}\big) \in \mathcal{C}\big([0,T], \mathbb{R}^d\big) \times \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big), \text{ there exists } C \in \mathbb{R}_+ \text{ s.t.}\\ &\qquad G\big(\alpha, (\gamma_t)_{t\in[0,T]}\big) \le C\big[1 + \|\alpha\|_{\sup}^r + \sup_{t\in[0,T]} \mathcal{W}_p^r(\gamma_t, \delta_0)\big],\end{aligned} \tag{6.3.1}$$

(*iii*) G is continuous in $(\gamma_t)_{t\in[0,T]}$ with respect to the distance $d_{\mathcal{C}}$ defined in (5.1.5) and non-decreasing in $(\gamma_t)_{t\in[0,T]}$ with respect to the convex order in the sense that

$$\begin{aligned}&\forall (\gamma_t)_{t\in[0,T]}, (\tilde{\gamma}_t)_{t\in[0,T]} \in \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big) \text{ such that } \forall t \in [0,T],\ \gamma_t \preceq_{cv} \tilde{\gamma}_t,\\ &\qquad \forall \alpha \in \mathcal{C}\big([0,T], \mathbb{R}^d\big), \qquad G\big(\alpha, (\gamma_t)_{t\in[0,T]}\big) \le G\big(\alpha, (\tilde{\gamma}_t)_{t\in[0,T]}\big),\end{aligned} \tag{6.3.2}$$

the result in Theorem 6.0.1 can be extended as follows.

Theorem 6.3.1. *Let $X := (X_t)_{t\in[0,T]}$, $Y := (Y_t)_{t\in[0,T]}$ respectively denote the solution of the McKean-Vlasov equation (6.0.1) and (6.0.2). For every $t \in [0,T]$, let μ_t, ν_t respectively denote the probability distribution of X_t and Y_t. If the equations (6.0.1) and (6.0.2) satisfy conditions in Assumption (I), (II) and (III), then for any function G satisfying the above conditions (i), (ii) and (iii), one has*

$$\mathbb{E}\, G\big(X, (\mu_t)_{t\in[0,T]}\big) \le \mathbb{E}\, G\big(Y, (\nu_t)_{t\in[0,T]}\big). \tag{6.3.3}$$

The proof of Theorem 6.3.1 is very similar to the proof of Theorem 6.0.1. Firstly, in order to prove the functional convex order result for the Euler schemes (6.1.1) and (6.1.2)

$$\mathbb{E}\, \tilde{G}(\bar{X}_0, ..., \bar{X}_m, \bar{\mu}_0, ..., \bar{\mu}_M) \le \mathbb{E}\, \tilde{G}(\bar{Y}_0, ..., \bar{Y}_m, \bar{\nu}_0, ..., \bar{\nu}_M) \tag{6.3.4}$$

with

$$\tilde{G} : (x_{0:M}, \gamma_{0:M}) \in (\mathbb{R}^d)^{M+1} \times \big(\mathcal{P}_p(\mathbb{R}^d)\big)^{M+1} \mapsto \tilde{G}(x_{0:M}, \gamma_{0:M}) \in \mathbb{R}$$

convex in $x_{0:M}$, non-decreasing in $\gamma_{0:M}$ with respect to the convex order and having an r-polynomial growth, we just need to replace the definition of Φ_m in (6.1.11) and (6.1.12) by $\Phi'_m : (\mathbb{R}^d)^{m+1} \times \big(\mathcal{P}_p(\mathbb{R}^d)\big)^{M+1}, m = 0, ..., M$, which are defined by

$$\forall (x_{0:m}, \gamma_{0:M}) \in (\mathbb{R}^d)^{m+1} \times \big(\mathcal{P}_p(\mathbb{R}^d)\big)^{M+1},$$
$$\Phi'_M(x_{0:M}, \gamma_{0:M}) = \tilde{G}(x_{0:M}, \gamma_{0:M})$$

and

$$\Phi'_m = \big(Q_{m+1}\Phi'_{m+1}(x_{0:m}, \cdot, \gamma_{0:M})\big)\big(x_m, \sigma_m(x_m, \gamma_m)\big).$$

Now we discuss the key step from the functional convex order of Euler scheme (6.3.4) to the functional convex order of process and its marginal probability distribution (6.3.3).

Let $\lambda \in [0, 1]$. For any two random variables X_1, X_2 with respective probability distributions $\gamma_1, \gamma_2 \in \mathcal{P}_p(\mathbb{R}^d)$, we define a linear combination of γ_1, γ_2, denoted by $\lambda\gamma_1 + (1 - \lambda)\gamma_2$, by

$$\forall A \in \mathcal{B}(\mathbb{R}^d), \quad \big(\lambda\gamma_1 + (1 - \lambda)\gamma_2\big)(A) \coloneqq \lambda\gamma_1(A) + (1 - \lambda)\gamma_2(A). \tag{6.3.5}$$

It is obvious from the above definition (6.3.5) that $\lambda\gamma_1 + (1 - \lambda)\gamma_2 \in \mathcal{P}_p(\mathbb{R}^d)$ and $\lambda\gamma_1 + (1 - \lambda)\gamma_2$ is in fact the distribution of

$$\mathbb{1}_{\{U \le \lambda\}} X_1 + \mathbb{1}_{\{U > \lambda\}} X_2,$$

where U is a random variable with probability distribution $\mathcal{U}([0, 1])$ and independent to (X_1, X_2). Moreover, for a fixed $(\gamma_1, \gamma_2) \in \big(\mathcal{P}_p(\mathbb{R}^d)\big)^2$, the application $\lambda \in [0, 1] \mapsto \lambda\gamma_1 + (1 - \lambda)\gamma_2 \in \mathcal{P}_p(\mathbb{R}^d)$ is continuous with respect to $\mathcal{W}_p$.

From (6.3.5) we can extend the definition of interpolator i_M (*respectively* I_M) to the probability distribution space $\big(\mathcal{P}_p(\mathbb{R}^d)\big)^{M+1}$ (*resp.* $\mathcal{C}\big([0, T], \mathcal{P}_p(\mathbb{R}^d)\big)$) as follows

$$\begin{aligned}
&\forall\, m = 0, ..., M - 1,\ \forall\, t \in [t_m^M, t_{m+1}^M],\\
&\forall \gamma_{0:M} \in \big(\mathcal{P}_p(\mathbb{R}^d)\big)^{M+1},\\
&\qquad i_M(\gamma_{0:M})(t) = \frac{M}{T}\big[(t_{m+1}^M - t)\gamma_m + (t - t_m^M)\gamma_{m+1}\big],\\
&\forall (\gamma_t)_{t\in[0,T]} \in \mathcal{C}\big([0, T], \mathcal{P}_p(\mathbb{R}^d)\big),\\
&\qquad I_M\big((\gamma_t)_{t\in[0,T]}\big) = i_M\big(\gamma_{t_0^M}, ..., \gamma_{t_M^M}\big).
\end{aligned} \tag{6.3.6}$$

Let $\bar{\mu}^M$ and $\bar{\nu}^M$ respectively denote the probability distribution of $\bar{X}^M = (\bar{X}_t^M)_{t\in[0,T]}$

and $\bar{Y}^M = (\bar{Y}_t^M)_{t\in[0,T]}$, which is defined by (6.2.4) and (6.2.5). For every $t \in [0,T]$, let $\tilde{\mu}_t^M := I_M\big((\bar{\mu}_t^M)_{t\in[0,T]}\big)_t$.

We know from Lemma 5.1.3 and Corollary 5.2.1 that for any $p \geq 2$

$$\sup_{t\in[0,T]} \mathcal{W}_p(\mu_t, \bar{\mu}_t^M) \leq \mathbb{W}_p(\mu, \bar{\mu}^M) \to 0 \ \text{ as } \ M \to +\infty. \tag{6.3.7}$$

Now we prove that $\sup_{t\in[0,T]} \mathcal{W}_p(\bar{\mu}_t^M, \tilde{\mu}_t^M) \to 0$ as $M \to +\infty$. For every $t \in [t_m^M, t_{m+1}^M]$, let

$$\tilde{X}_t^M := \mathbb{1}_{\left\{U_m \leq \frac{M\left(t_{m+1}^M - t\right)}{T}\right\}} \bar{X}_{t_m}^M + \mathbb{1}_{\left\{U_m > \frac{M\left(t_{m+1}^M - t\right)}{T}\right\}} \bar{X}_{t_{m+1}}^M,$$

where $(U_0, ..., U_M)$ is independent to the Brownian Motion $(B_t)_{t\in[0,T]}$ in (6.0.1), (6.0.2) and $(Z_0, ..., Z_M)$ in (6.0.9). Thus, for every $t \in [t_m^M, t_{m+1}^M]$, $\tilde{X}_t^M$ has the probability distribution $\tilde{\mu}_t^M$. It follows that

$$\begin{aligned}
&\forall m \in \{0, ..., M\}, \forall t \in [t_m^M, t_{m+1}^M], \\
&\qquad \mathcal{W}_p^p(\bar{\mu}_t^M, \tilde{\mu}_t^M) \leq \mathbb{E}\left|\bar{X}_t^M - \tilde{X}_t^M\right|^p \\
&\qquad = \mathbb{E}\left|\bar{X}_t^M - \mathbb{1}_{\left\{U_m \leq \frac{M\left(t_{m+1}^M - t\right)}{T}\right\}} \bar{X}_{t_m}^M - \mathbb{1}_{\left\{U_m > \frac{M\left(t_{m+1}^M - t\right)}{T}\right\}} \bar{X}_{t_{m+1}}^M\right|^p \\
&\qquad \leq C_p\Big(\mathbb{E}\left|\bar{X}_t^M - \bar{X}_{t_m}^M\right|^p + \mathbb{E}\left|\bar{X}_t^M - \bar{X}_{t_{m+1}}^M\right|^p\Big)
\end{aligned}$$

and it follows from Lemma 5.2.2-(b) that

$$\begin{aligned}
&\forall s, t \in [t_m^M, t_{m+1}^M], \ \ s < t, \\
&\qquad \mathbb{E}\left|\bar{X}_t^M - \bar{X}_s^M\right|^p \leq (\kappa\sqrt{t-s}\,)^p \leq \kappa^p \big(\frac{T}{M}\big)^{\frac{p}{2}} \to 0, \ \text{as } M \to +\infty.
\end{aligned}$$

Thus we have $\sup_{t\in[0,T]} \mathcal{W}_p^p(\bar{\mu}_t^M, \tilde{\mu}_t^M) \to 0$ as $M \to +\infty$. Hence,

$$\sup_{t\in[0,T]} \mathcal{W}_p^p(\mu_t, \tilde{\mu}_t^M) \leq \sup_{t\in[0,T]} \mathcal{W}_p^p(\bar{\mu}_t^M, \mu_t) + \sup_{t\in[0,T]} \mathcal{W}_p^p(\bar{\mu}_t^M, \tilde{\mu}_t^M) \to 0 \text{ as } M \to +\infty.$$

Consequently,

$$\begin{aligned}
&\mathbb{E}\, G\big(I_M(\bar{X}^M), (\tilde{\mu}_t)_{t\in[0,T]}\big) = \mathbb{E}\, G\big(I_M(\bar{X}^M), I_M\big((\bar{\mu}_t^M)_{t\in[0,T]}\big)\big) \\
&\quad = \mathbb{E}\, G\big(i_M(\bar{X}_{t_0}^M, ..., \bar{X}_{t_M}^M), i_M(\bar{\mu}_{t_0}^M, ..., \bar{\mu}_{t_M}^M)\big) = \mathbb{E}\, G_M\big(\bar{X}_{t_0}^M, ..., \bar{X}_{t_M}^M, \bar{\mu}_0^M, ..., \bar{\mu}_{t_M}^M\big) \\
&\quad \leq \mathbb{E}\, G_M\big(\bar{Y}_{t_0}^M, ..., \bar{Y}_{t_M}^M, \bar{\nu}_{t_0}^M, ..., \bar{\nu}_{t_M}^M\big) = \mathbb{E}\, G\big(i_M(\bar{Y}_{t_0}^M, ..., \bar{Y}_{t_M}^M), i_M(\bar{\nu}_{t_0}^M, ..., \bar{\nu}_{t_M}^M)\big) \\
&\quad = \mathbb{E}\, G\big(I_M\big(\bar{Y}^M, (\bar{\nu}_t^M)_{t\in[0,T]}\big)\big),
\end{aligned} \tag{6.3.8}$$

where for any $(x_{0:M}, \gamma_{0:M}) \in (\mathbb{R}^d)^{M+1} \times \big(\mathcal{P}_p(\mathbb{R}^d)\big)^{M+1}$,

$$G_M(x_{0:M}, \gamma_{0:M}) \coloneqq G\big(i_M(x_{0:M}), i_M(\gamma_{0:M})\big).$$

Thus one can obtain (6.3.3) by the assumption (iii) on G and by applying the Lebesgue dominated convergence theorem.

Chapter 7

Particle Method, Quantization Based and Hybrid Scheme, Examples of Simulation

In this chapter, we consider the *homogeneous* McKean-Vlasov equation

$$\begin{cases} dX_t = b(X_t, \mu_t)dt + \sigma(X_t, \mu_t)dB_t \\ \forall t \in [0,T], \quad \mu_t = P_{X_t} \end{cases} \tag{7.0.1}$$

and establish the error analysis of the particle method (Section 7.1) and several different quantization based schemes (Section 7.2-7.5) under Assumption (I). At the end of this chapter, we compare the performances of these schemes on two examples in dimension 1 and 3. In Section 7.2-7.5, we assume that the conditions in Assumption (I) is satisfied with $p = 2$ and $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$ induced by the inner product $\langle \cdot \mid \cdot \rangle$.

7.1 Convergence rate of the particle method ($D \to C$)

Recall that the particle method is the following time discretized system,

$$(D): \quad \begin{cases} \forall n \in \{1, ..., N\}, \\ \bar{X}^{n,N}_{t_{m+1}} = \bar{X}^{n,N}_{t_m} + hb(\bar{X}^{n,N}_{t_m}, \bar{\mu}^N_{t_m}) + \sqrt{h}\sigma(\bar{X}^{n,N}_{t_m}, \bar{\mu}^N_{t_m})Z^n_{m+1}, \\ \bar{\mu}^N_{t_m} := \frac{1}{N}\sum_{n=1}^N \delta_{\bar{X}^{n,N}_{t_m}} \end{cases}$$

where $t_m = t^M_m := m \cdot \frac{T}{M}$, $M \in \mathbb{N}^*$, $X^{n,N}_0 \overset{i.i.d}{\sim} X_0$.

In this section, we study the convergence of $\mathcal{W}_p(\bar{\mu}^N_{t_m}, \bar{\mu}_{t_m})$ as $N \to +\infty$, where $\bar{\mu}_{t_m}$ is the probability distribution of $\bar{X}_{t_m}$ defined in the theoretical Euler scheme (C) and $\bar{\mu}^N_{t_m}$ is defined in the above Euler scheme of the N-particle system (D). The main result of this section is the following proposition.

Proposition 7.1.1. *Assume that Assumption (I) is in force. Then,*

(a) *Let $\bar{\mu}$ be the probability distribution of $\bar{X} = (\bar{X}_t)_{t\in[0,T]}$ defined in (5.2.3) and let ν^N denote the empirical measure of $\bar{\mu}$ generated by i.i.d copies of process $\bar{X} = (\bar{X}_t)_{t\in[0,T]}$. Then*

$$\left\| \sup_{1\le m\le M} \mathcal{W}_p(\bar{\mu}^N_{t_m}, \bar{\mu}_{t_m}) \right\|_p \le C_{d,p,L,T} \left\| \mathbb{W}_p(\bar{\mu}, \nu^N) \right\|_p .$$

(b) *If moreover, $\|X_0\|_{p+\varepsilon} < +\infty$ for some $\varepsilon > 0$, then*

$$\left\| \sup_{1\le m\le M} \mathcal{W}_p(\bar{\mu}^N_{t_m}, \bar{\mu}_m) \right\|_p \le \widetilde{C} \times \begin{cases} N^{-\frac{1}{2p}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p > d/2 \text{ and } \varepsilon \neq p \\ N^{-\frac{1}{2p}} \big[\log(1+N)\big]^{\frac{1}{p}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p = d/2 \text{ and } \varepsilon \neq p \\ N^{-\frac{1}{d}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p \in (0, d/2) \\ & \text{and } p+\varepsilon \neq \frac{d}{(d-p)} \end{cases},$$

where $\widetilde{C}$ is a constant depending on $p, \varepsilon, d, b, \sigma, L, T$.

We define the continuous time Euler scheme of (D), as what we did in Section 5.2 for the theoretical Euler scheme. For any $n \in \{1, ..., N\}$ and for any $t \in [t_m, t_{m+1})$, set

$$\bar{X}^{n,N}_t = \bar{X}^{n,N}_{t_m} + b\big(\bar{X}^{n,N}_{t_m}, \bar{\mu}^N_{t_m}\big)(t - t_m) + \sigma\big(\bar{X}^{n,N}_{t_m}, \bar{\mu}^N_{t_m}\big)(B^n_t - B^n_{t_m}) \tag{7.1.1}$$

where $B^n := (B^n_t)_{t\in[0,T]}, n = 1, ..., N$ are independent standard Brownian motions defined on $(\Omega, \mathcal{F}, \mathbb{P})$. For any $t \in [t_m, t_{m+1})$, define $\underline{t} = t_m$. Then, for every $n \in \{1, ..., N\}$, $\bar{X}^{n,N}_t$ is the solution of

$$d\bar{X}^{n,N}_t = b(\bar{X}^{n,N}_{\underline{t}}, \bar{\mu}^N_{\underline{t}})dt + \sigma(\bar{X}^{n,N}_{\underline{t}}, \bar{\mu}^N_{\underline{t}})dB^n_t, \tag{7.1.2}$$

where $\bar{\mu}^N_{\underline{t}} = \frac{1}{N}\sum_{n=1}^N \delta_{\bar{X}^{n,N}_{\underline{t}}}$.

Now we construct an i.i.d sample of size N of the process $\bar{X} = (\bar{X}_t)_{t\in[0,T]}$ defined in (5.2.3). It follows from Lemma 5.2.2-(a) that $\bar{X} \in L^p_{\mathcal{C}([0,T],\mathbb{R}^d)}(\Omega, \mathcal{F}, \mathbb{P})$, hence its probability distribution $\bar{\mu} \in \mathcal{P}_p\big(\mathcal{C}([0,T], \mathbb{R}^d)\big)$ and $\iota(\bar{\mu}) = (\bar{\mu}_t)_{t\in[0,T]} \in \mathcal{C}\big([0,T], \mathcal{P}_p(\mathbb{R}^d)\big)$ (see Lemma 5.1.2). Based on the same Brownian motions $B^n, n = 1, ..., N$ in (7.1.1), we define N Itô processes $Y^n, n = 1, ..., N$, by

$$\begin{cases} dY^n_t = b(Y^n_{\underline{t}}, \bar{\mu}_{\underline{t}})dt + \sigma(Y^n_{\underline{t}}, \bar{\mu}_{\underline{t}})dB^n_t \\ Y^n_0 = \bar{X}^{n,N}_0 \end{cases}.$$

Then $Y^n, n = 1, ..., N$, are i.i.d copies of $\bar{X}$ and

$$\nu^{N,\omega} := \frac{1}{N}\sum_{n=1}^{N} \delta_{Y^n(\omega)}, \tag{7.1.3}$$

is the empirical measure of $\bar{\mu}$. When there is no ambiguity, we will write ν^N instead of $\nu^{N,\omega}$.

The random measure $\nu^{N,\omega}$ is valued in $\mathcal{P}_p\big(\mathcal{C}([0,T],\mathbb{R}^d)\big)$. In fact, for every $\omega \in \Omega$, $Y^n(\omega)$ lies in $\mathcal{C}([0,T],\mathbb{R}^d)$ so that $\|Y^n(\omega)\|_{\sup} < +\infty$. Hence, for every $\omega \in \Omega$,

$$\int_{\mathcal{C}([0,T],\mathbb{R}^d)} \|\xi\|^p_{\sup}\, \nu^{N,\omega}(d\xi) = \frac{1}{N}\sum_{n=1}^{N} \|Y^n(\omega)\|^p_{\sup} < +\infty.$$

Notice that

$$\nu^{N,\omega}_t := \nu^{N,\omega} \circ \pi_t^{-1} = \frac{1}{N}\sum_{n=1}^{N} \delta_{Y^n_t(\omega)}$$

and it follows from Lemma 5.1.2 that $\iota(\nu^{N,\omega}) = (\nu^{N,\omega}_t)_{t\in[0,T]} \in \mathcal{C}\big([0,T],\mathcal{P}_p(\mathbb{R}^d)\big)$.

Let us recall the following theorem from Fournier and Guillin (2015), which yields a non-asymptotic upper bound of the convergence rate in the Wasserstein distance of the empirical measures.

Theorem 7.1.1. *(Fournier and Guillin (2015)[see Theorem 1]) Let $p > 0$ and let $\mu \in \mathcal{P}_q(\mathbb{R}^d)$ for some $q > p$. Let $U^1(\omega), ..., U^n(\omega), ...$ be i.i.d random variables with distribution μ. Let μ^ω_n denote the empirical measure of μ defined by*

$$\mu^\omega_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{U^i(\omega)}.$$

Then, there exists a real constant C only depending on p, d, q such that, for all $n \geq 1$,

$$\mathbb{E}\Big(\mathcal{W}^p_p(\mu^\omega_n, \mu)\Big) \leq C M^{p/q}_q(\mu) \times \begin{cases} n^{-1/2} + n^{-(q-p)/q} & \text{if } p > d/2 \text{ and } q \neq 2p \\ n^{-1/2}\log(1+n) + n^{-(q-p)/q} & \text{if } p = d/2 \text{ and } q \neq 2p \\ n^{-p/d} + n^{-(q-p)/q} & \text{if } p \in (0, d/2) \text{ and } q \neq d/(d-p) \end{cases},$$

where $M_q(\mu) := \int_{\mathbb{R}^d} |\xi|^q \, \mu(d\xi)$.

In particular, Theorem 7.1.1 implies that if $p \geq 2$,

$$\|\mathcal{W}_p(\mu^\omega_n, \mu)\|_p \leq C M^{1/q}_q(\mu) \times \begin{cases} n^{-1/2p} + n^{-(q-p)/qp} & \text{if } p > d/2 \text{ and } q \neq 2p \\ n^{-1/2p}\big(\log(1+n)\big)^{1/p} + n^{-(q-p)/qp} & \text{if } p = d/2 \text{ and } q \neq 2p \\ n^{-1/d} + n^{-(q-p)/qp} & \text{if } p \in (0, d/2) \text{ and } q \neq d/(d-p) \end{cases}. \tag{7.1.4}$$

Proof of Proposition 7.1.1. (a) For any $n \in \{1, ..., N\}$, we have

$$\left|Y_t^n - \bar{X}_t^{n,N}\right| = \left|\int_0^t \left[b(Y_{\underline{u}}^n, \bar{\mu}_{\underline{u}}) - b(\bar{X}_{\underline{u}}^{n,N}, \bar{\mu}_{\underline{u}}^N)\right] du + \int_0^t \left[\sigma(Y_{\underline{u}}^n, \bar{\mu}_{\underline{u}}) - \sigma(\bar{X}_{\underline{u}}^{n,N}, \bar{\mu}_{\underline{u}}^N)\right] dB_u\right|.$$

Hence,

$$\begin{aligned}
&\left\| \sup_{s\in[0,t]} \left|Y_s^n - \bar{X}_s^{n,N}\right| \right\|_p \\
&= \left\| \sup_{s\in[0,t]} \left| \int_0^s \left[b(Y_{\underline{u}}^n, \bar{\mu}_{\underline{u}}) - b(\bar{X}_{\underline{u}}^{n,N}, \bar{\mu}_{\underline{u}}^N)\right] du + \int_0^s \left[\sigma(Y_{\underline{u}}^n, \bar{\mu}_{\underline{u}}) - \sigma(\bar{X}_{\underline{u}}^{n,N}, \bar{\mu}_{\underline{u}}^N)\right] dB_u \right| \right\|_p \\
&\le \left\| \sup_{s\in[0,t]} \left| \int_0^s \left[b(Y_{\underline{u}}^n, \bar{\mu}_{\underline{u}}) - b(\bar{X}_{\underline{u}}^{n,N}, \bar{\mu}_{\underline{u}}^N)\right] du \right| + \sup_{s\in[0,t]} \left| \int_0^s \left[\sigma(Y_{\underline{u}}^n, \bar{\mu}_{\underline{u}}) - \sigma(\bar{X}_{\underline{u}}^{n,N}, \bar{\mu}_{\underline{u}}^N)\right] dB_u \right| \right\|_p \\
&\le \left\| \sup_{s\in[0,t]} \left| \int_0^s \left[b(Y_{\underline{u}}^n, \bar{\mu}_{\underline{u}}) - b(\bar{X}_{\underline{u}}^{n,N}, \bar{\mu}_{\underline{u}}^N)\right] du \right| \right\|_p \\
&\qquad + \left\| \sup_{s\in[0,t]} \left| \int_0^s \left[\sigma(Y_{\underline{u}}^n, \bar{\mu}_{\underline{u}}) - \sigma(\bar{X}_{\underline{u}}^{n,N}, \bar{\mu}_{\underline{u}}^N)\right] dB_u \right| \right\|_p \qquad \text{(by the Minkowski inequality)} \\
&\le L \int_0^t \left[\left\|Y_{\underline{u}}^n - \bar{X}_{\underline{u}}^{n,N}\right\|_p + \left\|\mathcal{W}_p(\bar{\mu}_{\underline{u}}, \bar{\mu}_{\underline{u}}^N)\right\|_p \right] du \\
&\qquad + C_{d,p,L} \Big[\int_0^t \left[\left\|Y_{\underline{u}}^n - \bar{X}_{\underline{u}}^{n,N}\right\|_p^2 + \left\|\mathcal{W}_p(\bar{\mu}_{\underline{u}}, \bar{\mu}_{\underline{u}}^N)\right\|_p^2 \right] du \Big]^{\frac{1}{2}} \quad \big(\text{by Lemma 5.1.7-(b)}\big) \\
&\le L \int_0^t \left\| \sup_{v\in[0,u]} \left|Y_v^n - \bar{X}_v^{n,N}\right| \right\|_p du + C_{d,p,L} \Big[\int_0^t \left\| \sup_{v\in[0,u]} \left|Y_v^n - \bar{X}_v^{n,N}\right| \right\|_p^2 du \Big]^{\frac{1}{2}} + \psi(t),
\end{aligned}$$

where

$$\psi(t) = L \int_0^t \left\|\mathcal{W}_p(\bar{\mu}_{\underline{u}}, \bar{\mu}_{\underline{u}}^N)\right\|_p du + C_{d,p,L} \Big[\int_0^t \left\|\mathcal{W}_p(\bar{\mu}_{\underline{u}}, \bar{\mu}_{\underline{u}}^N)\right\|_p^2 du \Big]^{\frac{1}{2}}, \tag{7.1.5}$$

owing to $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a \ge 0, b \ge 0$. Then by Lemma 5.2.1, we have

$$\left\| \sup_{s\in[0,t]} \left|Y_s^n - \bar{X}_s^{n,N}\right| \right\|_p \le 2\, e^{(2L + C_{d,p,L}^2)\, t} \psi(t).$$

Moreover, the empirical measure $\frac{1}{N}\sum_{n=1}^N \delta_{(\bar{X}^{n,N}, Y^n)}$ is a coupling of the random measures $\bar{\mu}^N$ and ν^N. Thus

$$\begin{aligned}
\mathbb{E}\, \mathbb{W}_{p,t}^p(\bar{\mu}^N, \nu^N) &= \mathbb{E}\Big[\inf_{\pi\in\Pi(\bar{\mu}^N, \nu^N)} \int_{\mathcal{C}([0,T],\mathbb{R}^d)\times\mathcal{C}([0,T],\mathbb{R}^d)} \sup_{s\in[0,t]} |x_s - y_s|^p \, \pi(dx, dy) \Big] \\
&\le \mathbb{E}\Big[\int_{\mathcal{C}([0,T],\mathbb{R}^d)\times\mathcal{C}([0,T],\mathbb{R}^d)} \sup_{s\in[0,t]} |x_s - y_s|^p \, \frac{1}{N} \sum_{n=1}^N \delta_{(\bar{X}^{n,N}, Y^n)}(dx, dy) \Big]
\end{aligned}$$

$$= \mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^{N}\sup_{s\in[0,t]}\big|\bar{X}_s^{n,N}-Y_s^n\big|^p\Big] = \frac{1}{N}\sum_{n=1}^{N}\Big\|\sup_{s\in[0,t]}\big|\bar{X}_s^{n,N}-Y_s^n\big|\Big\|_p^p$$
$$\leq \Big[2\,e^{(2L+C_{d,p,L}^2)\,t}\psi(t)\Big]^p \leq \Big[2\,e^{(2L+C_{d,p,L}^2)\,T}\psi(t)\Big]^p.$$

By Lemma 5.1.3, we have $\sup_{s\in[0,t]}\mathcal{W}_p^p(\bar{\mu}_s^N,\nu_s^N)\leq\mathbb{W}_{p,t}^p(\bar{\mu}^N,\nu^N)$, so that

$$\Big\|\sup_{s\in[0,t]}\mathcal{W}_p(\bar{\mu}_s^N,\nu_s^N)\Big\|_p \leq C_{d,p,L,T}\psi(t), \tag{7.1.6}$$

with $C_{d,p,L,T}=2\,e^{(2L+C_{d,p,L}^2)\,T}$. It follows that,

$$\begin{aligned}
\Big\|\sup_{s\in[0,t]}\mathcal{W}_p(\bar{\mu}_s^N,\bar{\mu}_s)\Big\|_p &\leq \Big\|\sup_{s\in[0,t]}\big[\mathcal{W}_p(\bar{\mu}_s^N,\nu_s^N)+\mathcal{W}_p(\nu_s^N,\bar{\mu}_s)\big]\Big\|_p\\
&\leq \Big\|\sup_{s\in[0,t]}\mathcal{W}_p(\bar{\mu}_s^N,\nu_s^N)\Big\|_p + \Big\|\sup_{s\in[0,t]}\mathcal{W}_p(\nu_s^N,\bar{\mu}_s)\Big\|_p\\
&\leq C_{d,p,L,T}\psi(t) + \Big\|\sup_{s\in[0,t]}\mathcal{W}_p(\nu_s^N,\bar{\mu}_s)\Big\|_p \quad \text{(by applying (7.1.6))}\\
&\leq \Big\|\sup_{s\in[0,t]}\mathcal{W}_p(\nu_s^N,\bar{\mu}_s)\Big\|_p + C_{d,p,L,T}\cdot L\int_0^t\Big\|\mathcal{W}_p(\bar{\mu}_{\underline{u}},\bar{\mu}_{\underline{n}}^N)\Big\|_p\,du\\
&\qquad + C_{d,p,L,T}\cdot C_{d,p,L}\Big[\int_0^t\Big\|\mathcal{W}_p(\bar{\mu}_{\underline{u}},\bar{\mu}_{\underline{n}}^N)\Big\|_p^2\,du\Big]^{\frac{1}{2}}\\
&\qquad \text{(by the defintion of } \psi(t) \text{ in (7.1.5))}\\
&\leq \Big\|\sup_{s\in[0,t]}\mathcal{W}_p(\nu_s^N,\bar{\mu}_s)\Big\|_p + C_{d,p,L,T}\cdot L\int_0^t\Big\|\sup_{v\in[0,u]}\mathcal{W}_p(\bar{\mu}_v,\bar{\mu}_v^N)\Big\|_p\,du\\
&\qquad + C_{d,p,L,T}\cdot C_{d,p,L}\Big[\int_0^t\Big\|\sup_{v\in[0,u]}\mathcal{W}_p(\bar{\mu}_v,\bar{\mu}_v^N)\Big\|_p^2\,du\Big]^{\frac{1}{2}}.
\end{aligned}$$

Then, by Lemma 5.2.1, we obtain

$$\Big\|\sup_{s\in[0,t]}\mathcal{W}_p(\bar{\mu}_s^N,\bar{\mu}_s)\Big\|_p \leq 2e^{(2A+B^2)T}\Big\|\sup_{s\in[0,t]}\mathcal{W}_p(\bar{\mu}_s,\nu_s^N)\Big\|_p, \tag{7.1.7}$$

where $A=C_{d,p,L,T}L$ and $B=C_{d,p,L,T}\cdot C_{d,p,L}$. Finally,

$$\begin{aligned}
\Big\|\sup_{1\leq m\leq M}\mathcal{W}_p(\bar{\mu}_{t_m}^N,\bar{\mu}_m)\Big\|_p &\leq 2e^{(2A+B^2)T}\Big\|\sup_{s\in[0,T]}\mathcal{W}_p(\bar{\mu}_s,\nu_s^N)\Big\|_p\\
&\leq 2e^{(2A+B^2)T}\Big\|\mathbb{W}_p(\bar{\mu},\nu^N)\Big\|_p \longrightarrow 0 \quad \text{as} \quad N\to+\infty
\end{aligned}$$

by applying Lemma 5.1.3.

(b) If $\|X_0\|_{p+\varepsilon} < +\infty$ for some $\varepsilon > 0$, then Lemma 5.2.2 implies

$$\big\|\bar{X}\big\|_{p+\varepsilon} = \Big\| \sup_{u\in[0,T]} \big|\bar{X}_u\big| \Big\|_{p+\varepsilon} \le C_{p,d,b,\sigma} e^{C_{p,d,b,\sigma}} \big(1 + \|X_0\|_{p+\varepsilon}\big) < +\infty.$$

Thus $\bar{\mu} \in \mathcal{P}_{p+\varepsilon}\big(\mathcal{C}([0,T],\mathbb{R}^d)\big)$, which implies that $\bar{\mu}_s \in \mathcal{P}_{p+\varepsilon}(\mathbb{R}^d)$ for any $s \in [0,T]$ by Lemma 5.1.2.

For any $s \in [0,T]$, ν_s^N is the empirical measure of $\bar{\mu}_s$. It follows from Theorem 7.1.1 that for any $s \in [0,T]$,

$$\Big\|\mathcal{W}_p(\nu_s^N, \bar{\mu}_s)\Big\|_p \le C M_{p+\varepsilon}^{1/p+\varepsilon}(\bar{\mu}_s) \times \begin{cases} N^{-1/2p} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p > d/2 \text{ and } \varepsilon \neq p \\ N^{-1/2p}\big(\log(1+N)\big)^{1/p} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p = d/2 \text{ and } \varepsilon \neq p \\ N^{-1/d} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p \in (0, d/2) \text{ and } p + \varepsilon \neq \frac{d}{(d-p)} \end{cases}. \tag{7.1.8}$$

Moreover, Lemma 5.2.2 implies that

$$\begin{aligned} \sup_{s\in[0,T]} M_{p+\varepsilon}(\bar{\mu}_s) = \sup_{s\in[0,T]} \mathbb{E}\big[\,|X_s|^{p+\varepsilon}\,\big] &\le \Big\| \sup_{s\in[0,T]} |X_s| \Big\|_{p+\varepsilon}^{p+\varepsilon} \\ &\le \Big[C_{p,d,b,\sigma} e^{C_{p,d,b,\sigma}T}(1 + \|X_0\|_{p+\varepsilon})\Big]^{p+\varepsilon} < +\infty. \end{aligned} \tag{7.1.9}$$

Thus it follows from (7.1.7) that

$$\begin{aligned} &\Big\| \sup_{1\le m\le M} \mathcal{W}_p(\bar{\mu}_{t_m}^N, \bar{\mu}_m) \Big\|_p \\ &\le \widetilde{C} \times \begin{cases} N^{-\frac{1}{2p}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p > d/2 \text{ and } \varepsilon \neq p \\ N^{-\frac{1}{2p}}\big[\log(1+N)\big]^{\frac{1}{p}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p = d/2 \text{ and } \varepsilon \neq p \\ N^{-\frac{1}{d}} + N^{-\frac{\varepsilon}{p(p+\varepsilon)}} & \text{if } p \in (0, d/2) \text{ and } p + \varepsilon \neq \frac{d}{(d-p)} \end{cases}, \end{aligned}$$

where $\widetilde{C}$ is a constant depending on $p, \varepsilon, d, b, \sigma, L, T$ and $\|X_0\|_{p+\varepsilon}$. □

7.2 L^2-error analysis of the theoretical quantization (E→C)

From now on, let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^d$ induced by the inner product $\langle \cdot \mid \cdot \rangle$. Recall that the theoretical quantization procedure reads

$$(E): \quad \begin{cases} \widetilde{X}_0 = X_0, \ \ \widehat{X}_0 = \mathrm{Proj}_{x^{(0)}}(\widetilde{X}_0) \\ \widetilde{X}_{t_{m+1}} = \widehat{X}_{t_m} + hb(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) + \sqrt{h}\sigma(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})Z_{m+1}, \ \ m = 0, ..., M-1, \\ \text{where } h = \frac{T}{M} \ \text{ and } \ \widehat{\mu}_{t_m} = P_{\widehat{X}_{t_m}} \\ \widehat{X}_{t_{m+1}} = \mathrm{Proj}_{x^{(m+1)}}(\widetilde{X}_{t_{m+1}}), \ \ m = 0, ..., M-1, \end{cases}$$

where for $m = 0, ..., M$, $x^{(m)} = (x_1^{(m)}, ..., x_{K_m}^{(m)})$ is the K_m-quantizer of $\widetilde{X}_{t_m}$.

For $m = 1, ..., M$, let $\Xi_m := e_{K, \widetilde{X}_{t_m}}\left(x^{(m)}\right)$ denote the quadratic quantization error of $\widetilde{X}_{t_m}$ induced by $x^{(m)}$. The next proposition establishes the L^2-error of the quantization method at every time t_m, $m = 1, ..., M$.

Proposition 7.2.1. *Assume that Assumption (I) is satisfied with $p = 2$.*

(a) For any $m \in \{1, ..., M\}$,

$$\left\| \bar{X}_{t_m} - \widehat{X}_{t_m} \right\|_2 \leq \sum_{j=1}^{m} \left[1 + 2Lh(1 + Lh + Lq)\right]^{m-j} \cdot e_{K, \widetilde{X}_{t_j}}\left(x^{(j)}\right). \tag{7.2.1}$$

(b) If for every $m = 0, 1, ..., M$, $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$ is an optimal quantizer of $\bar{X}_{t_m}$ and if moreover, $\|X_0\|_{2+\varepsilon} < +\infty$ for some $\varepsilon > 0$, then

$$\left\| \bar{X}_{t_m} - \widehat{X}_{t_m} \right\|_2 = \mathcal{O}(K^{-1/d}). \tag{7.2.2}$$

Remark 7.2.1. From Proposition 7.2.1 we know that in order to obtain a simulation with the minimum error by the quantization method, we need to reduce at each step m the quantization error Ξ_m. Thus we can apply the Lloyd algorithm (4.0.19) at each step m, as mentioned in Algorithm 2 and Algorithm 4.

Proof of Proposition 7.2.1. *(a)* Let $\bar{b}_m = b(\bar{X}_{t_m}, \bar{\mu}_{t_m})$, $\bar{\sigma}_m = \sigma(\bar{X}_{t_m}, \bar{\mu}_{t_m})$, $\widehat{b}_m = b(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})$ and $\widehat{\sigma}_m = \sigma(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})$. The definition of $\widetilde{X}_{t_m}$ in (E) and $\bar{X}_{t_m}$ in (C) directly imply that

$$\left| \bar{X}_{t_{m+1}} - \widetilde{X}_{t_{m+1}} \right| = \left| (\bar{X}_{t_m} - \widehat{X}_{t_m}) + \left[\bar{b}_m - \widehat{b}_m\right]h + \left[\bar{\sigma}_m - \widehat{\sigma}_m\right]\sqrt{h}Z_{m+1} \right|.$$

Hence,

$$\mathbb{E} \left| \bar{X}_{t_{m+1}} - \widetilde{X}_{t_{m+1}} \right|^2$$

$$
\begin{aligned}
&= \mathbb{E}\left|\big(\bar{X}_{t_m} - \widehat{X}_{t_m}\big) + h\big[\bar{b}_m - \widehat{b}_m\big]\right|^2 + h\mathbb{E}\left|\big[\bar{\sigma}_m - \widehat{\sigma}_m\big]Z_{m+1}\right|^2 \\
&\quad + 2\sqrt{h}\,\mathbb{E}\left\langle \big(\bar{X}_{t_m} - \widehat{X}_{t_m}\big) + h\big[\bar{b}_m - \widehat{b}_m\big] \mid \big[\bar{\sigma}_m - \widehat{\sigma}_m\big]Z_{m+1}\right\rangle \\
&= \mathbb{E}\left|\bar{X}_{t_m} - \widehat{X}_{t_m}\right|^2 + h^2\mathbb{E}\left|\bar{b}_m - \widehat{b}_m\right|^2 + 2h\,\mathbb{E}\left\langle \bar{X}_{t_m} - \widehat{X}_{t_m} \mid \bar{b}_m - \widehat{b}_m\right\rangle \\
&\quad + h\mathbb{E}\left|\big[\bar{\sigma}_m - \widehat{\sigma}_m\big]Z_{m+1}\right|^2 + 2\sqrt{h}\,\mathbb{E}\left\langle \big(\bar{X}_{t_m} - \widehat{X}_{t_m}\big) + h\big[\bar{b}_m - \widehat{b}_m\big] \mid \big[\bar{\sigma}_m - \widehat{\sigma}_m\big]Z_{m+1}\right\rangle.
\end{aligned}
\tag{7.2.3}
$$

For any $m \in \{1, ..., M\}$, define $\mathcal{F}_m$ the σ-algebra generated by $X_0, Z_1, ..., Z_m$. Then,

$$
\begin{aligned}
&\mathbb{E}\left\langle \big(\bar{X}_{t_m} - \widehat{X}_{t_m}\big) + h\big[\bar{b}_m - \widehat{b}_m\big] \mid \big[\bar{\sigma}_m - \widehat{\sigma}_m\big]Z_{m+1}\right\rangle \\
&\quad = \mathbb{E}\left[\big[\big(\bar{X}_{t_m} - \widehat{X}_{t_m}\big) + h\big(\bar{b}_m - \widehat{b}_m\big)\big]^\top \big(\bar{\sigma}_m - \widehat{\sigma}_m\big)Z_{m+1}\right] \\
&\quad = \mathbb{E}\left\{\mathbb{E}\left[\big[\big(\bar{X}_{t_m} - \widehat{X}_{t_m}\big) + h\big(\bar{b}_m - \widehat{b}_m\big)\big]^\top \big(\bar{\sigma}_m - \widehat{\sigma}_m\big)Z_{m+1} \big| \mathcal{F}_m\right]\right\} \\
&\quad = \mathbb{E}\left\{\left[\big[\big(\bar{X}_{t_m} - \widehat{X}_{t_m}\big) + h\big(\bar{b}_m - \widehat{b}_m\big)\big]^\top \big(\bar{\sigma}_m - \widehat{\sigma}_m\big)\right]\mathbb{E}\left[Z_{m+1}\right]\right\} = 0.
\end{aligned}
$$

Moreover, Assumption (I) implies that

$$
\mathbb{E}\left|\bar{b}_m - \widehat{b}_m\right|^2 \le 2L^2\Big[\mathbb{E}\left|\bar{X}_{t_m} - \widehat{X}_{t_m}\right|^2 + \mathbb{E}\mathcal{W}_2^2(\bar{\mu}_m, \widehat{\mu}_m)\Big] \le 4L^2\,\mathbb{E}\left|\bar{X}_{t_m} - \widehat{X}_{t_m}\right|^2
$$

so that

$$
\mathbb{E}\left\langle \bar{X}_{t_m} - \widehat{X}_{t_m} \mid \bar{b}_m - \widehat{b}_m\right\rangle \le \left\|\bar{X}_{t_m} - \widehat{X}_{t_m}\right\|_2 \left\|\bar{b}_m - \widehat{b}_m\right\|_2 \le 2L\,\mathbb{E}\left|\bar{X}_{t_m} - \widehat{X}_{t_m}\right|^2
$$

and

$$
\begin{aligned}
\mathbb{E}\big|(\bar{\sigma}_m - \widehat{\sigma}_m)Z_{m+1}\big|^2 &\le \mathbb{E}\big[|||\bar{\sigma}_m - \widehat{\sigma}_m|||^2 Z_{m+1}^2\big] \le \mathbb{E}\Big[\mathbb{E}\big[|||\bar{\sigma}_m - \widehat{\sigma}_m|||^2 Z_{m+1}^2 | \mathcal{F}_m\big]\Big] \\
&= \mathbb{E}\Big[|||\bar{\sigma}_m - \widehat{\sigma}_m|||^2 \mathbb{E}\big[Z_{m+1}^2\big]\Big] \le 4L^2 q\,\mathbb{E}\left|\bar{X}_{t_m} - \widehat{X}_{t_m}\right|^2.
\end{aligned}
$$

Consequently,

$$
\mathbb{E}\left|\bar{X}_{t_{m+1}} - \widetilde{X}_{t_{m+1}}\right|^2 \le \big[1 + 4Lh(1 + Lh + Lq)\big] \cdot \mathbb{E}\left|\bar{X}_{t_m} - \widehat{X}_{t_m}\right|^2
$$

so that

$$
\begin{aligned}
\left\|\bar{X}_{t_{m+1}} - \widetilde{X}_{t_{m+1}}\right\|_2 &\le \sqrt{1 + 4Lh(1 + Lh + Lq)}\left\|\bar{X}_{t_m} - \widehat{X}_{t_m}\right\|_2 \\
&\le \big[1 + 2Lh(1 + Lh + Lq)\big]\left\|\bar{X}_{t_m} - \widehat{X}_{t_m}\right\|_2
\end{aligned}
$$

and

$$\begin{aligned}\left\|\bar{X}_{t_{m+1}} - \widehat{X}_{t_{m+1}}\right\|_2 &\le \left\|\bar{X}_{t_{m+1}} - \widetilde{X}_{t_{m+1}}\right\|_2 + \left\|\widetilde{X}_{t_{m+1}} - \widehat{X}_{t_{m+1}}\right\|_2 \\ &\le \left[1 + 2Lh(1 + Lh + Lq)\right] \left\|\bar{X}_{t_m} - \widehat{X}_{t_m}\right\|_2 + \Xi_{m+1}.\end{aligned}$$

This directly implies

$$\left\|\bar{X}_{t_m} - \widehat{X}_{t_m}\right\|_2 \le \sum_{j=1}^{m} \left[1 + 2Lh(1 + Lh + Lq)\right]^{m-j} \Xi_j.$$

(b) It follows from Proposition 5.2.1 that, if $\|X_0\|_{2+\varepsilon} < +\infty$, then for every $m \in \{1, ..., M\}$, $\bar{\mu}_{t_m} \in \mathcal{P}_{2+\varepsilon}(\mathbb{R}^d)$. Thus, if for every $m = 1, ..., M$, $x^{(m)}$ is the optimal quantizer of $\bar{X}_{t_m}$, the convergence rate (7.2.2) is a direct consequence of Zador's theorem (see Proposition 4.0.1-(b)). □

7.3 Recursive quantization for the Vlasov equation (G→E)

7.3.1 Recursive quantization for a fixed quantizer sequence

For $m = 1, ..., M$, let $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)}) \in (\mathbb{R}^d)^K$ be the quantizer of $\bar{X}_{t_m}$ defined in (C) and let $\left(C_k(x^{(m)})\right)_{1 \le k \le K}$ denote the Voronoï partition generated by $x^{(m)}$. For any $m \in \{1, ..., M\}$ and $k \in \{1, ..., K\}$, let $p_k^{(m)} = \mathbb{P}(\widetilde{X}_{t_m} \in C_k(x^{(m)})) = \mathbb{P}(\widehat{X}_{t_m} = x_k^{(m)})$ and $p^{(m)} = (p_1^{(m)}, ..., p_K^{(m)})$. Hence the probability distribution of $\mathrm{Proj}_{x^{(m)}}(\widetilde{X}_{t_m})$ is $\widehat{\mu}_m = \sum_{k=1}^{K} p_k^{(m)} \delta_{x_k^{(m)}}$.

In the Vlasov case, that is, there exist $\beta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ and $a : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{M}_{d,q}(\mathbb{R})$ such that

$$b(x, \mu) = \int_{\mathbb{R}^d} \beta(x, u)\mu(du) \quad \text{and} \quad \sigma(x, \mu) = \int_{\mathbb{R}^d} a(x, u)\mu(du),$$

the theoretical quantization formulas (E) can be written as

$$\begin{aligned}\widetilde{X}_{t_{m+1}} &= \widehat{X}_{t_m} + b(\widehat{X}_{t_m}, \widehat{\mu}_m)h + \sigma(\widehat{X}_{t_m}, \widehat{\mu}_m)\sqrt{h}Z_{m+1} \\ &= \widehat{X}_{t_m} + h\sum_{k=1}^{K} \beta(\widehat{X}_{t_m}, x_k^{(m)})p_k^{(m)} + \sqrt{h}Z_{m+1} \sum_{k=1}^{K} a(\widehat{X}_{t_m}, x_k^{(m)})p_k^{(m)}.\end{aligned}$$

Thus, given $\widehat{X}_{t_m}$ and $p^{(m)}$, we have

$$\widetilde{X}_{t_{m+1}} \sim \mathcal{N}\Big(\widehat{X}_{t_m}+h\sum_{k=1}^{K} p_k^{(m)}\beta(\widehat{X}_{t_m},x_k^{(m)}),\ \ h\big[\sum_{k=1}^{K} p_k^{(m)}a(\widehat{X}_{t_m},x_k^{(m)})\big]^\top\big[\sum_{k=1}^{K} p_k^{(m)}a(\widehat{X}_{t_m},x_k^{(m)})\big]\ \Big)$$

since $Z_{m+1}\sim\mathcal{N}(0,\mathbf{I}_q)$. Thus, $(\widehat{X}_{t_m},p^{(m)})_{0\le m\le M}$ makes up a Markov chain with transition probability

$$\begin{aligned}\pi_{ij}^{(m)} :=&\mathbb{P}\big(\widehat{X}_{t_{m+1}}=x_j^{(m+1)}\mid\widehat{X}_{t_m}=x_i^{(m)},p^{(m)}\big)\\ =&\mathbb{P}\big(\widetilde{X}_{t_{m+1}}\in C_j(x^{(m+1)})\mid\widehat{X}_{t_m}=x_i^{(m)},p^{(m)}\big)\\ =&\mathbb{P}\Big[\underbrace{\big(x_i^{(m)}+h\sum_{k=1}^{K}p_k^{(m)}\beta(x_i^{(m)},x_k^{(m)})+\sqrt{h}\sum_{k=1}^{K}p_k^{(m)}a(x_i^{(m)},x_k^{(m)})Z_{m+1}\big)}_{\mathcal{E}_i(x^{(m)},p^{(m)},Z_{m+1})}\in C_j(x^{(m+1)})\Big]\end{aligned}\tag{7.3.1}$$

and

$$\begin{aligned}&\mathbb{P}\big(\widehat{X}_{t_{m+1}}=x_j^{(m+1)}\mid p^{(m)}\big)=\mathbb{P}\big(\widetilde{X}_{t_{m+1}}\in C_j(x^{(m+1)})\mid p^{(m)}\big)\\ &=\sum_{i=1}^{K}\mathbb{P}\big(\widehat{X}_{t_{m+1}}=x_j^{(m+1)}\mid\widehat{X}_{t_m}=x_i^{(m)},p^{(m)}\big)\cdot\mathbb{P}(\widehat{X}_{t_m}=x_i^{(m)})\\ &=\sum_{i=1}^{K}\mathbb{P}\Big(\mathcal{E}_i(x^{(m)},p^{(m)},Z_{m+1})\in C_j(x^{(m+1)})\Big)\cdot p_i^{(m)}.\end{aligned}\tag{7.3.2}$$

The formula (7.3.2) is in fact the value of $p_j^{(m+1)}$ given $p^{(m)}$.

7.3.2 Application of Lloyd's algorithm to the recursive quantization

In order to implement Lloyd's algorithm, we need to compute

$$\begin{aligned}&\mathbb{E}\big[\widetilde{X}_{t_{m+1}}\mathbb{1}_{C_j(x^{(m+1)})}(\widetilde{X}_{t_{m+1}})\mid p^{(m)}\big]\\ =&\mathbb{E}\Big[\mathbb{E}\big[\widetilde{X}_{t_{m+1}}\mathbb{1}_{C_j(x^{(m+1)})}(\widetilde{X}_{t_{m+1}})\mid\widehat{X}_{t_m}=x_i^{(m)},p^{(m)}\big]\ \Big|\ p^{(m)}\Big]\\ =&\sum_{i=1}^{K}\mathbb{E}\big[\widetilde{X}_{t_{m+1}}\mathbb{1}_{C_j(x^{(m+1)})}(\widetilde{X}_{t_{m+1}})\mid\widehat{X}_{t_m}=x_i^{(m)},p^{(m)}\big]\cdot\mathbb{P}(\widehat{X}_{t_m}=x_i^{(m)})\\ =&\sum_{i=1}^{K}\mathbb{E}\big[\widetilde{X}_{t_{m+1}}\mathbb{1}_{C_j(x^{(m+1)})}(\widetilde{X}_{t_{m+1}})\mid\widehat{X}_{t_m}=x_i^{(m)},p^{(m)}\big]\cdot p_i^{(m)},\end{aligned}\tag{7.3.3}$$

where

$$\mathbb{E}\big[\widetilde{X}_{t_{m+1}}\mathbb{1}_{C_j(x^{(m+1)})}(\widetilde{X}_{t_{m+1}})\mid \widehat{X}_{t_m}=x_i^{(m)},p^{(m)}\big]=\mathbb{E}\big[Y\mathbb{1}_{C_j(x^{(m+1)})}(Y)\big] \tag{7.3.4}$$

with

$$Y\sim\mathcal{N}\Big(x_i^{(m)}+h\sum_{k=1}^K p_k^{(m)}\beta(x_i^{(m)},x_k^{(m)}),\ \ h\big[\sum_{k=1}^K p_k^{(m)}a(x_i^{(m)},x_k^{(m)})\big]^\top\big[\sum_{k=1}^K p_k^{(m)}a(x_i^{(m)},x_k^{(m)})\big]\Big). \tag{7.3.5}$$

Then, given $p^{(m)}$, we can use (7.3.3) and (7.3.2) to compute the Lloyd iteration (4.0.19) in order to obtain the optimal quantizer of $\widetilde{X}_{t_{m+1}}$.

Remark 7.3.1. The recursive quantization method has a high computing speed in dimension 1 since the the Voronoï cells in dimension 1 are in fact intervals of $\overline{\mathbb{R}}$. For example, let $x=(x_1,...,x_K)\in\mathbb{R}^K$ be a quantizer with $x_i<x_{i+1}$, $i=1,...,K$, one can choose a Voronoï partition as follows:

$$\begin{aligned}
&C_1(x)=\big(-\infty,\frac{x_1+x_2}{2}\big),\\
&C_k(x)=\big[\frac{x_{k-1}+x_k}{2},\frac{x_k+x_{k+1}}{2}\big), \qquad k=2,...,K-1,\\
&C_K(x)=\big[\frac{x_{K-1}+x_K}{2},+\infty\big).
\end{aligned}$$

Let $x^{(m)}=(x_1^{(m)},...,x_K^{(m)})$ be the quantizer of the m-th Euler step. The transition probability $\pi_{ij}^{(m)}$ in (7.3.1) reads

$$F_{m,\sigma^2}\Big(\frac{x_{j+1}^{(m)}+x_j^{(m)}}{2}\Big)-F_{m,\sigma^2}\Big(\frac{x_{j-1}^{(m)}+x_j^{(m)}}{2}\Big),$$

where F_{m,σ^2} denotes the cumulative distribution function of $\mathcal{N}(m,\sigma^2)$ with

$$m=x_i^{(m)}+h\sum_{k=1}^K p_k^{(m)}\beta(x_i^{(m)},x_k^{(m)}) \ \text{ and } \ \sigma=\sqrt{h}\big[\sum_{k=1}^K p_k^{(m)}a(x_i^{(m)},x_k^{(m)})\big]. \tag{7.3.6}$$

Moreover, the Lloyd iteration (7.3.4) depends on

$$\int_{(x_{j-1}^{(m)}+x_j^{(m)})/2}^{(x_{j+1}^{(m)}+x_j^{(m)})/2}\xi\cdot f_{m,\sigma^2}(\xi)d\xi \tag{7.3.7}$$

where $f_{m,\sigma^2}(\xi)$ is the density function of $\mathcal{N}(m,\sigma^2)$ with the same m and σ as in (7.3.6). In fact, to avoid computing the integral, (7.3.7) can be alternatively calculated by the following method,

$$\forall a,b\in\mathbb{R},\quad \int_a^b\xi\cdot f_{m,\sigma^2}(\xi)(\xi)d\xi=\int_a^b\frac{1}{\sqrt{2\pi\sigma^2}}\xi e^{-\frac{(\xi-m)^2}{2\sigma^2}}d\xi,$$

$$= \frac{\sigma}{\sqrt{2\pi}}\left[-e^{-\frac{(\xi-m)^2}{2\sigma^2}}\right]_a^b + m\big[F_{m,\sigma^2}(b) - F_{m,\sigma^2}(a)\big]. \tag{7.3.8}$$

However, in high dimension, there does not exist such an alternative formula as (7.3.8) to accelerate the calculation. We refer to the website *www.qhull.com* for the cubature formulas of the numerical integration over a convex set in low dimensions.

7.4 L^2-error analysis of doubly quantized scheme (H)

Let $x^{(m)} = (x_1^{(m)}, ..., x_K^{(m)})$ denote the quantizer of $\bar{X}_{t_m}$ at m-th Euler step. Recall that the doubly quantized scheme can be written as follows[(1)] (a more detailed version is in Algorithm 3),

$$(H): \quad \begin{cases} \widetilde{X}_0 = X_0, \ \ \widehat{X}_0 = \mathrm{Proj}_{x^{(0)}}(\widetilde{X}_0) \\ \widetilde{X}_{t_{m+1}} = \widehat{X}_{t_m} + h\cdot b(\widehat{X}_{t_m}, \widehat{\mu}_{t_m}) + \sqrt{h}\,\sigma(\widehat{X}_{t_m}, \widehat{\mu}_{t_m})\widehat{Z}_{m+1}, \ \ m = 0, ..., M-1 \\ \text{where } h = \frac{T}{M} \ \text{ and } \ \widehat{\mu}_{t_m} = P_{\widehat{X}_{t_m}} \\ \widehat{X}_{t_{m+1}} = \mathrm{Proj}_{x^{(m+1)}}(\widetilde{X}_{t_{m+1}}), \end{cases},$$

where $z = (z_1, ..., z_J)$ is an L^2-optimal quantizer of $\mathcal{N}(0, \mathbf{I}_q)$ with $J \gg K$[(2)], $w = (w_1, ..., w_J)$ the corresponding weight of z, $\widehat{Z}_m \overset{i.i.d}{\sim} \sum_{j=1}^J \delta_{z_j} w_j$, and $(\widehat{Z}_1, ..., \widehat{Z}_M)$ is independent to X_0.

The reason why we can explicitly represent $\widehat{\mu}_{t_m}$ if we use $\widehat{Z}_m$ instead of Z_m is the following. If we have two independent random variables X and Y with respective discrete distributions $X \sim \sum_{n=1}^N \delta_{x_n} p_n^x$, $Y \sim \sum_{m=1}^M \delta_{y_m} p_m^y$, $M, N \in \mathbb{N}^*$, we can always explicitly write the distribution of $f(X) + g(X)Y$ with f, g Borel function by enumerating all possible occurrences of this random variable, namely

$$f(X) + g(X)Y \ \sim \sum_{\substack{1\le n\le N \\ 1\le m\le M}} \delta_{f(x_n)+g(x_n)y_m} \cdot p_n^x\, p_m^y.$$

The following proposition establish an error bound for the doubly quantization method.

Proposition 7.4.1. *Let $\widehat{X}_{t_m}$, $\widehat{\mu}_{t_m}$ define as in (H) and let $\bar{X}_{t_m}$ and $\bar{\mu}_{t_m}$ define as in*

(1) By a slight abus of notation, we use here the same notation as in (E).
(2) It is a natural recommendation for practitioners but not a mathematical requirement.

(C). Assume that Assumption (I) is satisfied with $p = 2$, then

$$\mathcal{W}_2(\bar{\mu}_{t_m}, \widehat{\mu}_{t_m}) \leq \left\| \bar{X}_{t_m} - \widehat{X}_{t_m} \right\|_2 \leq \sum_{j=0}^{m} \Big[1 + hL(2 + 2hL + q)\Big]^{m-j} \Big[\Xi_j + \sqrt{h} \cdot \tilde{C} \cdot e_{K_2,Z}\Big]^j,$$

where $e^2_{K_2,Z}$ denotes the quantization error of Z on its optimal quantizer z and $\Xi_m = \left\| \widetilde{X}_m - \widehat{X}_m \right\|_2$ denotes the L^2-quantization error of $\widetilde{X}_m$ on $x^{(m)}$.

Proof. In order to simplify the notation, we write $\bar{X}_m$ (respectively, $\widetilde{X}_m$, $\widehat{X}_m$) instead of $\bar{X}_{t_m}$ (respectively, $\widetilde{X}_{t_m}$, $\widehat{X}_{t_m}$). It follows that

$$\begin{aligned}
&\mathbb{E}\left|\bar{X}_{m+1} - \widetilde{X}_{m+1}\right|^2 \\
&= \mathbb{E}\Big|\big(\bar{X}_m - \widehat{X}_m\big) + h\big[b(\bar{X}_m, \bar{\mu}_m) - b(\widehat{X}_m, \widehat{\mu}_m)\big] \\
&\qquad\qquad + \sqrt{h}\big[\sigma(\bar{X}_m, \bar{\mu}_m)Z_{m+1} - \sigma(\widehat{X}_m, \widehat{\mu}_m)\widehat{Z}_{m+1}\big]\Big|^2 \\
&= \underbrace{\mathbb{E}\Big|\big(\bar{X}_m - \widehat{X}_m\big) + h\big[b(\bar{X}_m, \bar{\mu}_m) - b(\widehat{X}_m, \widehat{\mu}_m)\big]\Big|^2}_{(a)} \\
&\qquad\qquad + \underbrace{\mathbb{E}\left[h\left|\sigma(\bar{X}_m, \bar{\mu}_m)Z_{m+1} - \sigma(\widehat{X}_m, \widehat{\mu}_m)\widehat{Z}_{m+1}\right|^2\right]}_{(b)} \\
&\underbrace{+ 2\sqrt{h}\,\mathbb{E}\left\langle \big(\bar{X}_m - \widehat{X}_m\big) + h\big[b(\bar{X}_m, \bar{\mu}_m) - b(\widehat{X}_m, \widehat{\mu}_m)\big] \;\middle|\; \sigma(\bar{X}_m, \bar{\mu}_m)Z_{m+1} - \sigma(\widehat{X}_m, \widehat{\mu}_m)\widehat{Z}_{m+1} \right\rangle}_{(c)}.
\end{aligned} \tag{7.4.1}$$

Remark that at each step m, we take the optimal quantizer of $\mathcal{N}(0, \mathrm{I}_q)$ so that by Proposition 4.0.1-(a), we have for every $m = 0, ..., M$,

$$\mathbb{E}\left[Z_{m+1} \mid \widehat{Z}_{m+1}\right] = \widehat{Z}_{m+1}. \tag{7.4.2}$$

Hence, $\mathbb{E}\left[\widehat{Z}_{m+1}\right] = \mathbb{E}\left[Z_{m+1}\right] = \mathbf{0}_q$. Consequently, Term (c) of (7.4.1) equals to 0.

For Term (a) of (7.4.1), we have

$$\begin{aligned}
(a) &= \mathbb{E}\Big|\big(\bar{X}_m - \widehat{X}_m\big) + h\big[b(\bar{X}_m, \bar{\mu}_m) - b(\widehat{X}_m, \widehat{\mu}_m)\big]\Big|^2 \\
&= \mathbb{E}\Big[\Big|\bar{X}_m - \widehat{X}_m\Big|^2 + h^2\Big|b(\bar{X}_m, \bar{\mu}_m) - b(\widehat{X}_m, \widehat{\mu}_m)\Big|^2 \\
&\qquad + 2h\big\langle \bar{X}_m - \widehat{X}_m \mid b(\bar{X}_m, \bar{\mu}_m) - b(\widehat{X}_m, \widehat{\mu}_m)\big\rangle\Big]
\end{aligned}$$

and

$$\begin{aligned}\langle \bar{X}_m - \widehat{X}_m \mid b(\bar{X}_m, \bar{\mu}_m) - b(\widehat{X}_m, \widehat{\mu}_m)\rangle &\le \left|\bar{X}_m - \widehat{X}_m\right| \left|b(\bar{X}_m, \bar{\mu}_m) - b(\widehat{X}_m, \widehat{\mu}_m)\right| \\ &\le L\left|\bar{X}_m - \widehat{X}_m\right| \left[\left|\bar{X}_m - \widehat{X}_m\right| + \mathcal{W}_2(\bar{\mu}_m, \widehat{\mu}_m)\right],\end{aligned}$$

so that owing to the fact that $\mathcal{W}_2^2(\bar{\mu}_m, \widehat{\mu}_m) \le \mathbb{E}\left|\bar{X}_m - \widehat{X}_m\right|^2$,

$$\begin{aligned}(a) &\le \mathbb{E}\left[\left|\bar{X}_m - \widehat{X}_m\right|^2 + 4h^2L^2\left|\bar{X}_m - \widehat{X}_m\right|^2 + 4hL\left|\bar{X}_m - \widehat{X}_m\right|^2\right] \\ &\le \left(1 + 4h^2L^2 + 4hL\right)\mathbb{E}\left|\bar{X}_m - \widehat{X}_m\right|^2.\end{aligned}$$

Next, for Part (b) of (7.4.1), we have

$$\begin{aligned}(b) &= \mathbb{E}\left[h\left|\sigma(\bar{X}_m, \bar{\mu}_m)Z_{m+1} - \sigma(\widehat{X}_m, \widehat{\mu}_m)\widehat{Z}_{m+1}\right|^2\right] \\ &= h\mathbb{E}\left[\left|\sigma(\bar{X}_m, \bar{\mu}_m)(Z_{m+1} - \widehat{Z}_{m+1}) + \left[\sigma(\bar{X}_m, \bar{\mu}_m) - \sigma(\widehat{X}_m, \widehat{\mu}_m)\right]\widehat{Z}_{m+1}\right|^2\right] \\ &= h\Big[\mathbb{E}\left|\sigma(\bar{X}_m, \bar{\mu}_m)(Z_{m+1} - \widehat{Z}_{m+1})\right|^2 + \mathbb{E}\left|\left[\sigma(\bar{X}_m, \bar{\mu}_m) - \sigma(\widehat{X}_m, \widehat{\mu}_m)\right]\widehat{Z}_{m+1}\right|^2\Big],\end{aligned}$$

where the last equality is due to the orthogonality between $Z_{m+1} - \widehat{Z}_{m+1}$ and $\widehat{Z}_{m+1}$ by (7.4.2). It follows that

$$\begin{aligned}\mathbb{E}\left|\sigma(\bar{X}_m, \bar{\mu}_m)(Z_{m+1} - \widehat{Z}_{m+1})\right|^2 &\le \mathbb{E}\left\|\sigma(\bar{X}_m, \bar{\mu}_m)\right\|^2 \cdot \mathbb{E}\left|Z_{m+1} - \widehat{Z}_{m+1}\right|^2 \\ &\le C_{b,\sigma,L}\left(1 + \left\|\bar{X}_m\right\|_2^2\right) \cdot e_{K_2,Z}^2 \\ &\le C_{b,\sigma,L}\left(1 + \left\|\sup_{1\le m\le M}\left|\bar{X}_m\right|\right\|_2^2\right) \cdot e_{K_2,Z}^2 \\ &\le C_{b,\sigma,L,T,\|X_0\|_2} \cdot e_{K_2,Z}^2,\end{aligned}$$

where $e_{K_2,Z}^2$ denote the quantization error of Z on its optimal quantizer z and the last inequality is due to Lemma 5.2.2, and

$$\begin{aligned}\mathbb{E}\left|\left[\sigma(\bar{X}_m, \bar{\mu}_m) - \sigma(\widehat{X}_m, \widehat{\mu}_m)\right]\widehat{Z}_{m+1}\right|^2 &\le \mathbb{E}\left\|\sigma(\bar{X}_m, \bar{\mu}_m) - \sigma(\widehat{X}_m, \widehat{\mu}_m)\right\|^2 \cdot \mathbb{E}\left|\widehat{Z}_{m+1}\right|^2 \\ &\le 2Lq\,\mathbb{E}\left|\bar{X}_m - \widehat{X}_m\right|^2,\end{aligned}$$

where the last inequality is due to (7.4.2)

$$\mathbb{E}\left|\widehat{Z}_{m+1}\right|^2 = \mathbb{E}\left[\left|\mathbb{E}\left[Z_{m+1} \mid \widehat{Z}_{m+1}\right]\right|^2\right] \le \mathbb{E}\left[\left|Z_{m+1}\right|^2\right] = q.$$

Consequently,

$$\mathbb{E}\left|\bar{X}_{m+1}-\widetilde{X}_{m+1}\right|^2 \leq \left[1+2hL(2+2hL+q)\right]\mathbb{E}\left|\bar{X}_m-\widehat{X}_m\right|^2 + h\cdot\tilde{C}\cdot e_{K_2,Z}^2,$$

where $\tilde{C}=C_{b,\sigma,L,T,\|X_0\|_2}$. Thus

$$\begin{aligned}\left\|\bar{X}_{m+1}-\widehat{X}_{m+1}\right\|_2 &\leq \sqrt{1+2hL(2+2hL+q)}\left\|\bar{X}_m-\widehat{X}_m\right\|_2 + \sqrt{h}\cdot\tilde{C}\cdot e_{K_2,Z}\\ &\leq \left[1+hL(2+2hL+q)\right]\left\|\bar{X}_m-\widetilde{X}_m\right\|_2 + \sqrt{h}\cdot\tilde{C}\cdot e_{K_2,Z}.\end{aligned}$$

Finally, let $\Xi_m=\left\|\widetilde{X}_m-\widehat{X}_m\right\|_2$ denote the L^2-quantization error of $\widetilde{X}_m$ on $x^{(m)}$, then

$$\left\|\bar{X}_m-\widehat{X}_m\right\|_2 \leq \sum_{j=0}^{m}\left[1+hL(2+2hL+q)\right]^{m-j}\left[\Xi_j+\sqrt{h}\cdot\tilde{C}\cdot e_{K_2,Z}\right]^j$$

and one concludes by using the fact that $\mathcal{W}_2(\bar{\mu}_m,\widehat{\mu}_m)\leq\left\|\bar{X}_m-\widehat{X}_m\right\|_2$. □

Remark 7.4.1. Comparing with the result of Proposition 7.4.1 and Proposition 7.2.1, the doubly quantized scheme adds at each step a quantization error of $\mathcal{N}(0,\mathbf{I}_q)$ in the sum. Here we give a brief comparison between the recursive quantization method and the doubly quantized scheme.

	Doubly quantized scheme	Recursive quantization method
Application scope	McKean-Vlasov equation	Vlasov equation
Computing time	better in dimension 1, acceptable in higher dimension	better in dimension 1, slow for higher dimension
Accuracy	higher L^2-error	lower L^2-error

Table 7.1 A brief comparison between the recursive quantization method and the doubly quantized scheme

7.5 L^2-error analysis of the hybrid particle-quantization scheme (G→D)

For $m=0,...,M$, let $x^{(m)}=(x_1^{(m)},...,x_K^{(m)})$ be the quantizer of $\bar{\mu}_{t_m}^N$ defined in (D) and let $\left(C_k(x^{(m)})\right)_{1\leq k\leq K}$ be a Voronoï partition generated by $x^{(m)}$. Let $\mathrm{Proj}_{x^{(m)}}$ denote

the projection function on $x^{(m)}$.

We recall the definition of the hybrid particle-quantization scheme:

$$(F): \quad \begin{cases} \forall n \in \{1,...,N\}, \\ \widetilde{X}^{n,N}_{t_{m+1}} = \widetilde{X}^{n,N}_{t_m} + b(\widetilde{X}^{n,N}_{t_m}, \widehat{\mu}^K_{t_m})h + \sigma(\widetilde{X}^{n,N}_{t_m}, \widehat{\mu}^K_{t_m})\sqrt{h}Z^n_{m+1} \\ \widehat{\mu}^K_{t_m} = \big(\frac{1}{N}\sum_{n=1}^N \delta_{\widetilde{X}^{n,N}_{t_m}}\big) \circ \mathrm{Proj}^{-1}_{x^{(m)}} = \sum_{k=1}^K \big[\delta_{x^{(m)}_k} \cdot \sum_{n=1}^N \mathbb{1}_{V_k(x^{(m)})}(\widetilde{X}^{n,N}_{t_m})\big] \\ \bar{X}^{n,N}_0 \overset{i.i.d}{\sim} X_0, \;\; Z^n_m \overset{i.i.d}{\sim} \mathcal{N}(0, I_q), \;\; (\bar{X}^{n,N}_0)_{1\le n\le N} \perp\!\!\!\perp (Z^n_m)_{1\le n\le N, 1\le m\le M}. \end{cases}$$

Then we use $\widehat{\mu}^K_{t_m}$ as an estimator of $\bar{\mu}^N_{t_m}$ in (D). The following proposition provides an upper bound of $\mathbb{E}\,\mathcal{W}_2\big(\widehat{\mu}^K_{t_m}, \bar{\mu}^N_{t_m}\big)$.

Proposition 7.5.1. *Assume that the conditions in Assumption (I) is true with $p = 2$. Then for any $m \in \{1,...,M\}$, we have*

$$\mathbb{E}\,\mathcal{W}_2\big(\widehat{\mu}^K_{t_m}, \bar{\mu}^N_{t_m}\big) \le C_2 \sum_{j=0}^{m-1} C_1^j \sqrt{\mathbb{E}\,\hat{\Xi}^2_{m-1-j}} + \mathbb{E}\,\hat{\Xi}_m. \tag{7.5.1}$$

where $\hat{\Xi}_m = \mathcal{W}_2\big(\widehat{\mu}^K_{t_m}, \frac{1}{N}\sum_{i=1}^N \delta_{\widetilde{X}^{n,N}_{t_m}}\big)$ and C_1, C_2 are constants depending on h, L and q.

Remark 7.5.1. For every $m = 1,...,M$, it follows from (1.1.15) that

$$\mathbb{E}\,\hat{\Xi}_m = \mathbb{E}\,\mathcal{W}_2\big(\widehat{\mu}^K_{t_m}, \frac{1}{N}\sum_{i=1}^N \delta_{\widetilde{X}^{n,N}_{t_m}}\big) = \mathbb{E}\,e_{K, \frac{1}{N}\sum_{i=1}^N \delta_{\widetilde{X}^{n,N}_{t_m}}}(x^{(m)}).$$

Thus one can implement Lloyd's algorithm at each Euler step in order to minimize the error bound on the right-hand side of (7.5.1), as what mentioned in Algorithm 4.

Proof of Proposition 7.5.1. For any $m \in \{1,...,M\}$, the measure $\frac{1}{N}\sum_{n=1}^N \delta_{(\widetilde{X}^{n,N}_{t_m}, \bar{X}^{n,N}_{t_m})}$ is a random coupling of $\frac{1}{N}\sum_{n=1}^N \delta_{\widetilde{X}^{n,N}_{t_m}}$ and $\bar{\mu}^N_{t_m} = \frac{1}{N}\sum_{n=1}^N \delta_{\bar{X}^{n,N}_{t_m}}$. Thus, for any $m \in \{1,...,M\}$,

$$\begin{aligned} \mathbb{E}\Big[\mathcal{W}_2^2\big(\frac{1}{N}\sum_{n=1}^N \delta_{\widetilde{X}^{n,N}_{t_m}}, \bar{\mu}^N_{t_m}\big)\Big] &\le \mathbb{E}\Big[\int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^2 \frac{1}{N}\sum_{n=1}^N \delta_{(\widetilde{X}^{n,N}_{t_m}, \bar{X}^{n,N}_{t_m})}(dx,dy)\Big] \\ &= \mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^N \Big|\widetilde{X}^{n,N}_{t_m} - \bar{X}^{n,N}_{t_m}\Big|^2\Big]. \end{aligned} \tag{7.5.2}$$

On the other hand,

$$\begin{aligned} \widetilde{X}^{n,N}_{t_{m+1}} - \bar{X}^{n,N}_{t_{m+1}} =& \widetilde{X}^{n,N}_{t_m} - \bar{X}^{n,N}_{t_m} + \Big[b(\widetilde{X}^{n,N}_{t_m}, \widehat{\mu}^K_{t_m}) - b(\bar{X}^{n,N}_{t_m}, \bar{\mu}^N_{t_m})\Big]h \\ &+ \Big[\sigma(\widetilde{X}^{n,N}_{t_m}, \widehat{\mu}^K_{t_m}) - \sigma(\bar{X}^{n,N}_{t_m}, \bar{\mu}^N_{t_m})\Big]\sqrt{h}Z^n_{m+1}. \end{aligned} \tag{7.5.3}$$

Let $b_m^{\mathrm{Q}} := b(\widetilde{X}_{t_m}^{n,N}, \widehat{\mu}_{t_m}^K)$, $b_m^{\mathrm{Euler}} := b(\bar{X}_{t_m}^{n,N}, \bar{\mu}_{t_m}^N)$, $\sigma_m^{\mathrm{Q}} := \sigma(\widetilde{X}_{t_m}^{n,N}, \widehat{\mu}_{t_m}^K)$ and $\sigma_m^{\mathrm{Euler}} := \sigma(\bar{X}_{t_m}^{n,N}, \bar{\mu}_{t_m}^N)$. Let $\mathcal{F}_m$ be the $\sigma-$algebra generated by X_0, $Z_m^n, n = 1, ..., N, m = 1, ..., M$. Then b_m^{Euler}, b_m^{Q}, $\sigma_m^{\mathrm{Euler}}$, σ_m^{Q} are $\mathcal{F}_m$-measurable and $Z_{m+1}^n, n \in \{1, ..., N\}$ are independent of $\mathcal{F}_m$. Hence, it follows from (7.5.3) that

$$\begin{aligned}
&\mathbb{E}\left|\widetilde{X}_{t_{m+1}}^{n,N} - \bar{X}_{t_{m+1}}^{n,N}\right|^2 \\
&= \mathbb{E}\Big[\Big|(\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}) + \big(b_m^{\mathrm{Q}} - b_m^{\mathrm{Euler}}\big)h\Big|^2\Big] + \mathbb{E}\Big[\Big|\big(\sigma_m^{\mathrm{Q}} - \sigma_m^{\mathrm{Euler}}\big)\sqrt{h}Z_{m+1}^n\Big|^2\Big] \\
&\qquad + \underbrace{2\mathbb{E}\Big[\big\langle(\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}) + (b_m^{\mathrm{Q}} - b_m^{\mathrm{Euler}})h \mid (\sigma_m^{\mathrm{Q}} - \sigma_m^{\mathrm{Euler}})\sqrt{h}Z_{m+1}^n\big\rangle\Big]}_{=0 \text{ since } Z_{m+1}^n,\, n=1,...,N \text{ are independent of } \mathcal{F}_m.}. \qquad (7.5.4)
\end{aligned}$$

Moreover, Assumption (I) implies,

$$\left|b_m^{\mathrm{Q}} - b_m^{\mathrm{Euler}}\right| \vee \left|\sigma_m^{\mathrm{Q}} - \sigma_m^{\mathrm{Euler}}\right| \le L\Big[\left|\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}\right| + \mathcal{W}_2(\widehat{\mu}_{t_m}^K, \bar{\mu}_{t_m}^N)\Big].$$

Hence, we have

$$\mathbb{E}\Big[\Big|b_m^{\mathrm{Q}} - b_m^{\mathrm{Euler}}\Big|^2\Big] \le 2L^2\Big[\mathbb{E}\left|\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}\right|^2 + \mathbb{E}\mathcal{W}_2^2(\widehat{\mu}_{t_m}^K, \bar{\mu}_{t_m}^N)\Big],$$

and

$$\begin{aligned}
\mathbb{E}\Big[\Big|(\sigma_m^{\mathrm{Q}} - \sigma_m^{\mathrm{Euler}})\sqrt{h}\,Z_{m+1}^n\Big|^2\Big] &\le h\mathbb{E}\Big\{\mathbb{E}\big[\big|\big|\big|\sigma_m^{\mathrm{Q}} - \sigma_m^{\mathrm{Euler}}\big|\big|\big|^2 |Z_{m+1}^n|^2 \mid \mathcal{F}_m\big]\Big\} \\
&= h\,\mathbb{E}\Big\{\big|\big|\big|\sigma_m^{\mathrm{Q}} - \sigma_m^{\mathrm{Euler}}\big|\big|\big|^2\mathbb{E}\big[\,|Z_{m+1}^n|^2\,\big]\Big\} = h\,q\,\mathbb{E}\Big[\big|\big|\big|\sigma_m^{\mathrm{Q}} - \sigma_m^{\mathrm{Euler}}\big|\big|\big|^2\Big] \\
&\le 2L^2 h\,q\Big[\mathbb{E}\left|\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}\right|^2 + \mathbb{E}\mathcal{W}_2^2(\widehat{\mu}_{t_m}^K, \bar{\mu}_{t_m}^N)\Big].
\end{aligned}$$

Hence, (7.5.4) becomes

$$\begin{aligned}
&\mathbb{E}\big[\left|\widetilde{X}_{t_{m+1}}^{n,N} - \bar{X}_{t_{m+1}}^{n,N}\right|^2\big] \\
&= \mathbb{E}\Big[\Big|(\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}) + \big(b_m^{\mathrm{Q}} - b_m^{\mathrm{Euler}}\big)h\Big|^2\Big] + \mathbb{E}\Big[\Big|\big(\sigma_m^{\mathrm{Q}} - \sigma_m^{\mathrm{Euler}}\big)\sqrt{h}Z_{m+1}^n\Big|^2\Big] \\
&= \mathbb{E}\big[(\left|\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}\right|^2\big] + \mathbb{E}\Big[\Big|b_m^{\mathrm{Q}} - b_m^{\mathrm{Euler}}\Big|^2 h^2\Big] + \mathbb{E}\Big[\Big|\big(\sigma_m^{\mathrm{Q}} - \sigma_m^{\mathrm{Euler}}\big)\sqrt{h}Z_{m+1}^n\Big|^2\Big] \\
&\quad + 2h\mathbb{E}\big[\langle\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N} \mid b_m^{\mathrm{Q}} - b_m^{\mathrm{Euler}}\rangle\big] \\
&\le \mathbb{E}\big[(\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N})^2\big] + 2L^2(h^2 + hq)\Big[\mathbb{E}\left|\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}\right|^2 + \mathbb{E}\mathcal{W}_2^2(\widehat{\mu}_{t_m}^K, \bar{\mu}_{t_m}^N)\Big] \\
&\quad + h\,\mathbb{E}\Big[\left|\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}\right|^2 + \Big|b_m^{\mathrm{Q}} - b_m^{\mathrm{Euler}}\Big|^2\Big] \\
&\le \mathbb{E}\big[(\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N})^2\big] + 2L^2(h^2 + hq)\Big[\mathbb{E}\left|\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}\right|^2 + \mathbb{E}\mathcal{W}_2^2(\widehat{\mu}_{t_m}^K, \bar{\mu}_{t_m}^N)\Big] \\
&\quad + h\,\mathbb{E}\Big[\left|\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}\right|^2\Big] + 2L^2 h\Big[\mathbb{E}\left|\widetilde{X}_{t_m}^{n,N} - \bar{X}_{t_m}^{n,N}\right|^2 + \mathbb{E}\mathcal{W}_2^2(\widehat{\mu}_{t_m}^K, \bar{\mu}_{t_m}^N)\Big],
\end{aligned}$$

$$\leq \big[1+2L^2(h^2+hq)+h+2L^2h\big]\mathbb{E}\big[\,\big|\widetilde{X}^{n,N}_{t_m}-\bar{X}^{n,N}_{t_m}\big|^2\,\big] + (2L^2(h^2+hq)+2L^2h)\mathbb{E}\mathcal{W}_2^2(\widehat{\mu}^K_{t_m},\bar{\mu}^N_{t_m}).$$

Let $C_1 \coloneqq 1+2L^2(h^2+hq)+h+2L^2h$, $C_2 \coloneqq 2L^2(h^2+hq)+2L^2h$. Let

$$\hat{\Xi}_m = \mathcal{W}_2\big(\widehat{\mu}^K_{t_m}, \frac{1}{N}\sum_{i=1}^N \delta_{\widetilde{X}^{n,N}_{t_m}}\big).$$

It follows that,

$$\begin{aligned}
&\sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\,\big|\widetilde{X}^{n,N}_{t_{m+1}}-\bar{X}^{n,N}_{t_{m+1}}\big|^2\,\big]}\\
&= \sqrt{C_1\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[(\widetilde{X}^{n,N}_{t_m}-\bar{X}^{n,N}_{t_m})^2\big] + C_2\mathbb{E}\big[\mathcal{W}_2^2(\widehat{\mu}^K_{t_m},\bar{\mu}^N_{t_m})\big]}\\
&\leq \sqrt{C_1\cdot\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\,\big|\widetilde{X}^{n,N}_{t_m}-\bar{X}^{n,N}_{t_m}\big|^2\,\big] + C_2\Big[2\mathbb{E}\mathcal{W}_2^2(\widehat{\mu}^K_{t_m},\frac{1}{N}\sum_{i=1}^N\delta_{\widetilde{X}^{n,N}_{t_m}}) + 2\mathbb{E}\mathcal{W}_2^2(\frac{1}{N}\sum_{i=1}^N\delta_{\widetilde{X}^{n,N}_{t_m}},\bar{\mu}^N_{t_m})\Big]}\\
&\leq \sqrt{(C_1+2C_2)\cdot\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\,\big|\widetilde{X}^{n,N}_{t_m}-\bar{X}^{n,N}_{t_m}\big|^2\,\big] + 2C_2\mathbb{E}\,\hat{\Xi}_m^2} \quad \text{(by (7.5.2))}\\
&\leq \sqrt{C_1+2C_2}\sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\,\big|\widetilde{X}^{n,N}_{t_m}-\bar{X}^{n,N}_{t_m}\big|^2\,\big]} + \sqrt{2C_2}\sqrt{\mathbb{E}\,\hat{\Xi}_m^2}. \qquad (7.5.5)
\end{aligned}$$

Let $\bar{C}_1 \coloneqq \sqrt{C_1+2C_2}$ and $\bar{C}_2 = \sqrt{2C_2}$. The inequality (7.5.5) implies

$$\sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\,\big|\widetilde{X}^{n,N}_{t_m}-\bar{X}^{n,N}_{t_m}\big|^2\,\big]} \leq \bar{C}_2\sum_{j=0}^{m-1}\bar{C}_1^j\sqrt{\mathbb{E}\,\hat{\Xi}^2_{m-1-j}}.$$

Hence, it follows from (7.5.2) that

$$\begin{aligned}
\mathbb{E}\,\mathcal{W}_2\big(\frac{1}{N}\sum_{i=1}^N\delta_{\widetilde{X}^{n,N}_{t_m}},\,\bar{\mu}^N_{t_m}\big) &\leq \sqrt{\mathbb{E}\,\mathcal{W}_2^2\big(\frac{1}{N}\sum_{i=1}^N\delta_{\widetilde{X}^{n,N}_{t_m}},\,\bar{\mu}^N_{t_m}\big)} \leq \sqrt{\frac{1}{N}\sum_{i=1}^N\mathbb{E}\big[\,\big|\widetilde{X}^{n,N}_{t_m}-\bar{X}^{n,N}_{t_m}\big|^2\,\big]}\\
&\leq \bar{C}_2\sum_{j=0}^{m-1}\bar{C}_1^j\sqrt{\mathbb{E}\,\hat{\Xi}^2_{m-1-j}}.
\end{aligned}$$

Consequently,

$$\mathbb{E}\,\mathcal{W}_2\big(\widehat{\mu}^K_{t_m},\bar{\mu}^N_{t_m}\big) \leq \mathbb{E}\,\mathcal{W}_2\big(\frac{1}{N}\sum_{i=1}^N\delta_{\widetilde{X}^{n,N}_{t_m}},\,\bar{\mu}^N_{t_m}\big) + \mathbb{E}\,\mathcal{W}_2\big(\frac{1}{N}\sum_{i=1}^N\delta_{\widetilde{X}^{n,N}_{t_m}},\,\widehat{\mu}^K_{t_m}\big)$$

$$\leq \bar{C}_2 \sum_{j=0}^{m-1} \bar{C}_1^j \sqrt{\mathbb{E}\hat{\Xi}^2_{m-1-j}} + \mathbb{E}\,\hat{\Xi}_m.$$

□

7.6 Simulation examples

In this section, we illustrate our theoretical results by two simulations. The first one is the Burgers equation introduced and already considered for numerical tests in Bossy and Talay (1997). This is a one-dimensional example with an explicit solution and we use this example to compare the accuracy and computational time of the different simulation methods under consideration. The second example is the Network of FitzHugh-Nagumo neurons already numerically investigated in Baladron et al. (2012) (also in Reis et al. (2018)), which is a 3-dimensional example. All examples are written in *Python 3.7*.

7.6.1 Simulation of the Burgers equation, comparison of the three algorithms

In Bossy and Talay (1997), the authors analyse the solution and investigate the particle method of the Burgers equation

$$\begin{cases} dX_t = \int_{\mathbb{R}} H(X_t - y)\mu_t(dy)dt + \sigma dB_t \\ \forall t \in [0,T], \mu_t = P_{X_t} \\ X_0 : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})) \end{cases}, \tag{7.6.1}$$

where H is the Heaviside function ($H(z) = 1$, if $z \geq 0$, $H(z) = 0$, if $z < 0$) and σ is a real constant. If we denote by $V(t,x)$ the cumulative distribution function of μ_t, then $V(t,x)$ satisfies

$$\begin{cases} \frac{\partial V}{\partial t} = \frac{1}{2}\sigma^2 \frac{\partial^2 V}{\partial x^2} - V\frac{\partial V}{\partial x} \\ V(0,x) = V_0(x) \end{cases}. \tag{7.6.2}$$

Moreover, if the initial cumulative distribution function V_0 satisfies $\int_0^x V_0(y)dy = \mathcal{O}(x)$, then the function V has a closed form given by (see Hopf (1950))

$$V(t,x) = \frac{\int_{\mathbb{R}} V_0(y)\exp\Big(-\frac{1}{\sigma^2}\big[\frac{(x-y)^2}{2t} + \int_0^y V_0(z)dz\big]\Big)dy}{\int_{\mathbb{R}} \exp\Big(-\frac{1}{\sigma^2}\big[\frac{(x-y)^2}{2t} + \int_0^y V_0(z)dz\big]\Big)dy}, \quad (t,x) \in [0,T]\times\mathbb{R}. \tag{7.6.3}$$

Hence, if we consider $X_0 = 0$, then the cumulative distribution function at time $T = 1$ is

$$F_{T=1}(x) = \frac{\int_{\mathbb{R}_+} \exp\Big(- \frac{1}{\sigma^2}\big[\frac{(x-y)^2}{2} + y\big]\Big) dy}{\int_{\mathbb{R}} \exp\Big(- \frac{1}{\sigma^2}\big[\frac{(x-y)^2}{2} + y\mathbb{1}_{y\geq 0}\big]\Big) dy}. \tag{7.6.4}$$

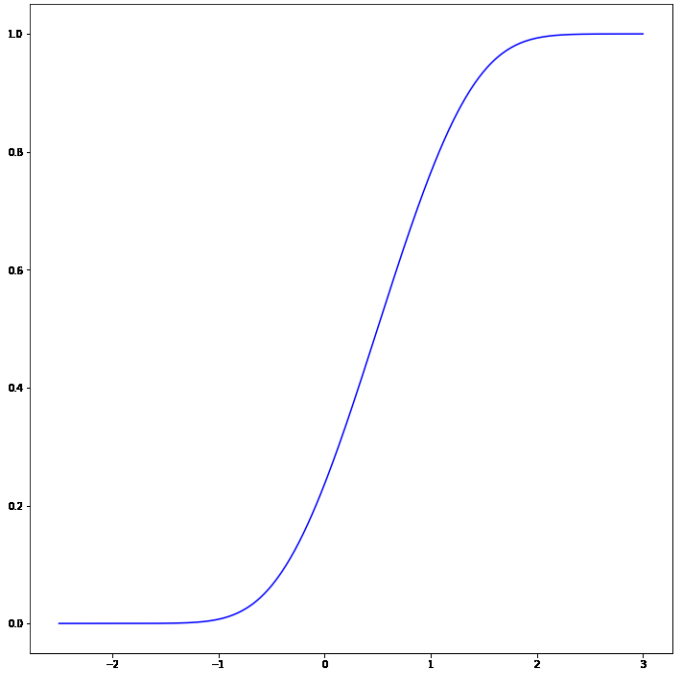


Figure 7.1 True cumulative distribution function

As there exists an explicit formula of the cumulative distribution function at time $T = 1$, we can compute the accuracy of the different numerical methods proposed in the former sections by computing

$$\|F_{\text{simu}} - F_{\text{true}}\|_{\text{sup}}\,, \tag{7.6.5}$$

where F_{simu} represents the simulated cumulative distribution function by different numerical methods and F_{true} is the true cumulative distribution function (7.6.4). We know that for two probability distributions $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ with respective cumulative distribution function F and G, the Wasserstein distance $\mathcal{W}_p(\mu, \nu)$ can be computed by

$$\mathcal{W}_p^p(\mu, \nu) = \int_0^1 \left|F^{-1}(u) - G^{-1}(u)\right|^p du,\ p \geq 1. \tag{7.6.6}$$

However, it is computationally extremely costly to directly compute the inverse function of the cumulative distribution function (7.6.4) and if we compute (7.6.6) by using

Monte-Carlo simulation, it will induce its own statistical error which may disturb our comparisons. Thus, instead of considering (7.6.6), we preferred to compute (7.6.5) by

$$\|F_{\text{simu}}(x) - F_{\text{true}}(x)\|_{\sup} \simeq \sup_{x \in Unifset} |F_{\text{simu}}(x) - F_{\text{true}}(x)|\,,$$

where $Unifset$ is a uniformly spaced point set in $[-2.5, 3.5]$. One may consider that this measure of the errors is more stringent than the Wasserstein distance, at least if $Unifset$ contains a great number of points.

In the following simulation, we choose $\sigma^2 = 0.2$ and $M = 50$ so that we have the same time step $h = \frac{T}{M} = 0.02$ for each method.

We first give a preliminary illustration of the simulated cumulative distribution function by Algorithm 1,2 and 4. The Burgers equation (7.6.1) is a one-dimensional Vlasov equation so that Algorithm 2 based on the recursive quantization method outperforms Algorithm 3 (see Remark 7.4.1). Hence, we omit the simulation by the doubly quantized scheme (Algorithm 3) in this example.

In a second phase, we will precisely present the decreasing rate of the error (7.6.5) of the particle method (Algorithm 1) and of the recursive quantization method without Lloyd quantizer optimization (Algorithm 2) respectively according to N and K. At the end of this section, we will give some comments of the numerical performance of different methods mainly through two aspects: the accuracy and the computing time.

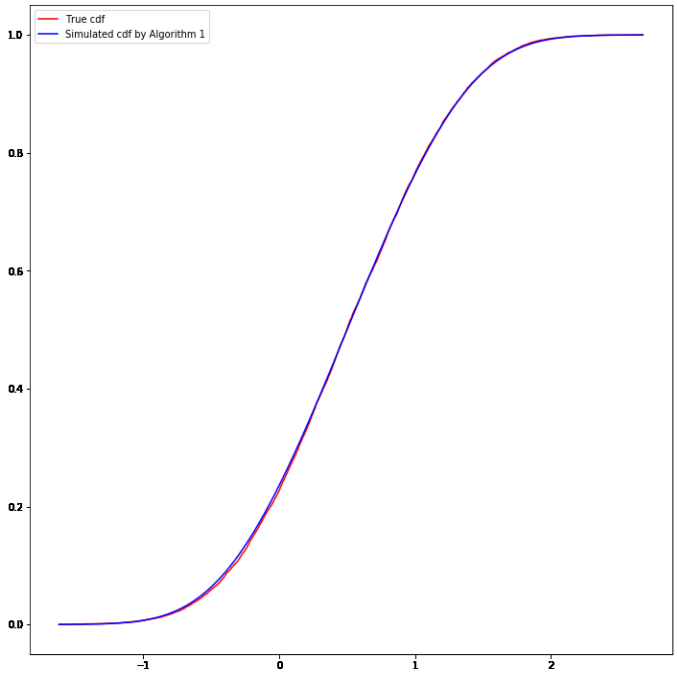


Figure 7.2 Simulated cumulative distribution function by the particle method (Algorithm 1)

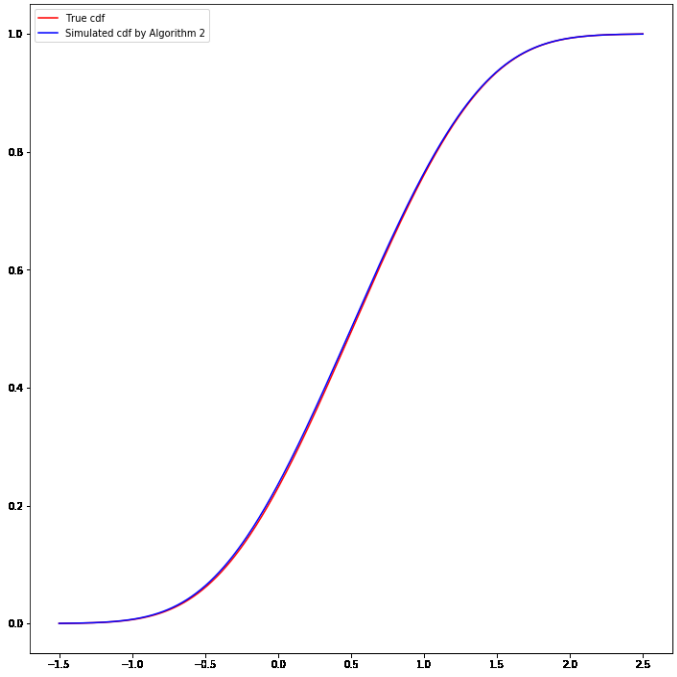


Figure 7.3 Simulated cumulative distribution function by the recursive quantization method without Lloyd iteration (Algorithm 2)

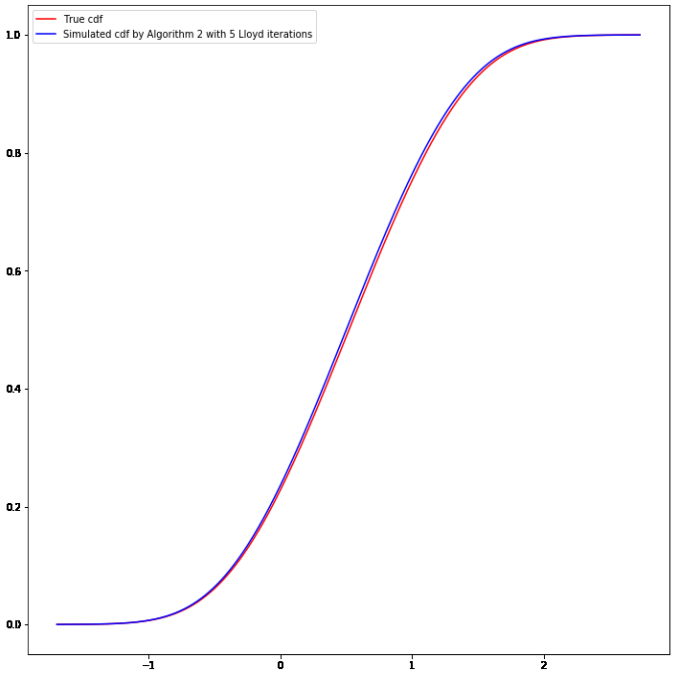


Figure 7.4 Simulated cumulative distribution function by the recursive quantization method with 5 Lloyd iterations at each Euler step (Algorithm 2)

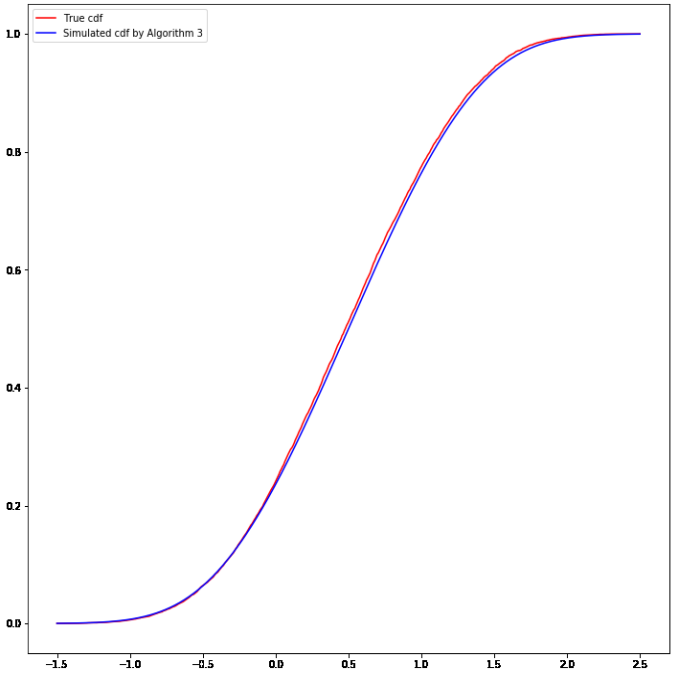


Figure 7.5 Simulated cumulative distribution function by the hybrid particle-quantization scheme without Lloyd iteration (Algorithm 4)

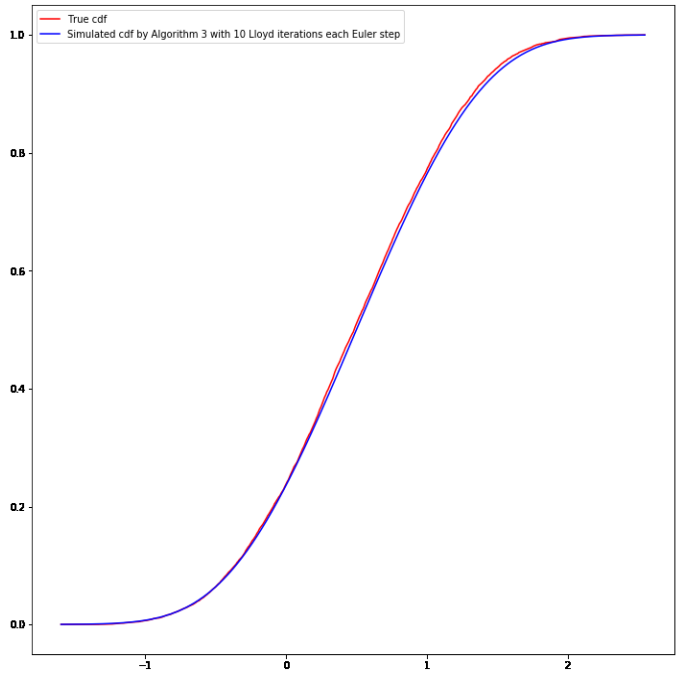


Figure 7.6 Simulated cumulative distribution function by the hybrid particle-quantization scheme with 5 Lloyd iterations at each Euler step (Algorithm 4)

A detailed comparison of the different methods is displayed in the following table. Remind that the particle method (Algorithm 1) and the hybrid particle-quantization scheme (Algorithm 4) are random algorithms so that their accuracy are computed by taking an average error computed over 50 independent identical simulations.

		Particle size and quantizer size	Computing time for each Euler step	Error $\|F_{\text{simu}}(x) - F_{\text{true}}(x)\|_{\text{sup}}$
Algorithm 1	Particle method	Particle size $N = 10000$	0.00320s	0.01021
Algorithm 2	Recursive quantization method without Lloyd iterations	Quantizer size $K = 500$	0.00205s	0.01054
Algorithm 2	Recursive quantization method with 5 Lloyd iterations at each Euler step	Quantizer size $K = 500$	8.21598s	0.01029
Algorithm 4	Hybrid particle-quantization scheme without Lloyd iterations	Particle size $N = 10000$ Quantizer size $K = 500$	6.09480s	0.01626
Algorithm 4	Hybrid particle-quantization scheme with 5 Lloyd iterations at each Euler step	Particle size $N = 10000$ Quantizer size $K = 500$	9.37229s	0.01013

Table 7.2 Detailed comparison of three algorithms

Now we present the convergence rate of the error of the particle method (Algorithm 1) with respect to the particle size $N = 2^8, 2^9, 2^{10}, 2^{11}, 2^{12}, 2^{13}$ for a fixed $M = 50$. As the particle method (Algorithm 1) is a random algorithm, the simulation results are also random, including the error $\|F_{\text{simu}} - F_{\text{true}}\|_{\text{sup}}$. Consequently, we will rerun independently and identically 500 times for each value of N.

N	2^8	2^9	2^{10}	2^{11}	2^{12}	2^{13}
Error $\|F_{\text{simu}} - F_{\text{true}}\|_{\text{sup}}$	0.04691	0.03409	0.02438	0.01785	0.01407	0.01131
Standard deviation	0.01207	0.00939	0.00687	0.00469	0.00408	0.00294

Table 7.3 Error of the particle method (Algorithm 1) with respect to the particle size N

In the following figure we show the curve of the error with respect to N and the log-error with respect to $\log_2(N)$.

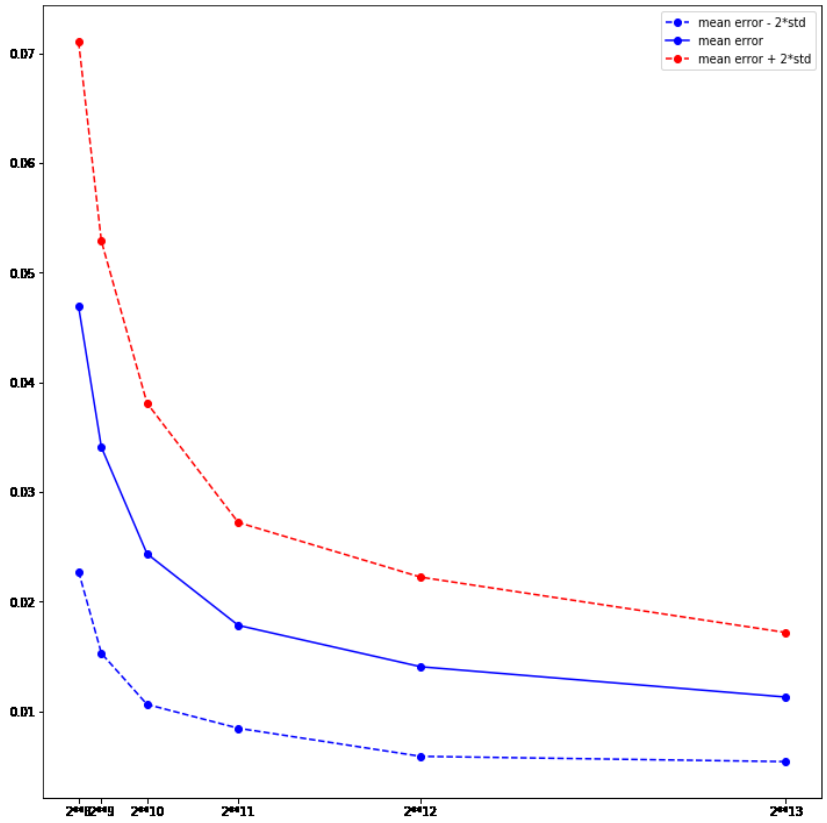


Figure 7.7 Error of the particle method (Algorithm 1) with respect to the particle size N

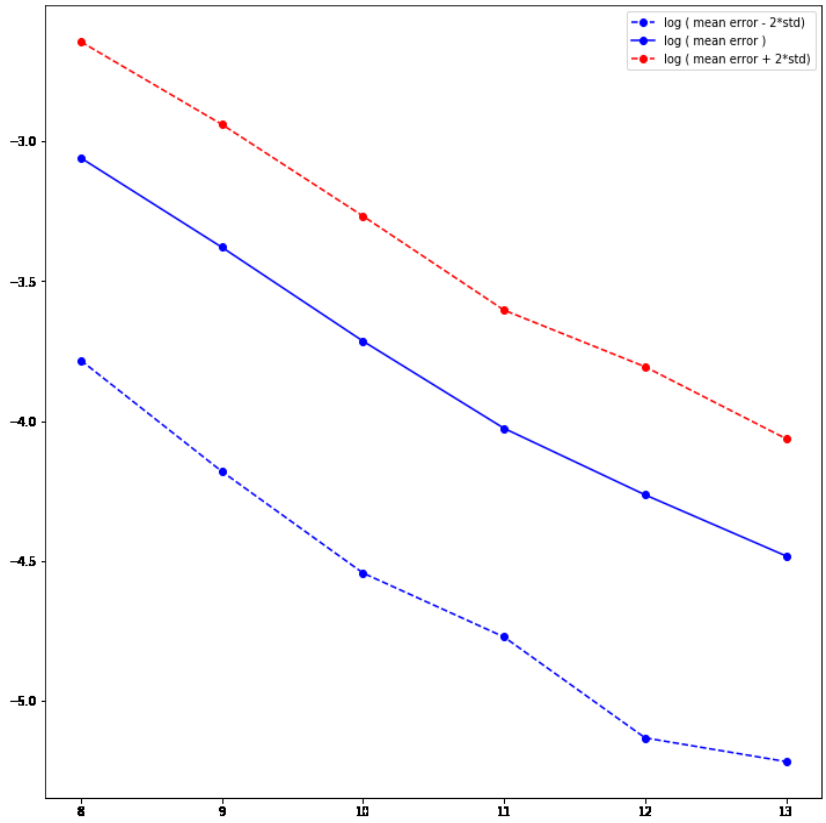


Figure 7.8 Log-error of the particle method (Algorithm 1) with respect to $\log_2(N)$. The slope is approximately equal to -0.28451.

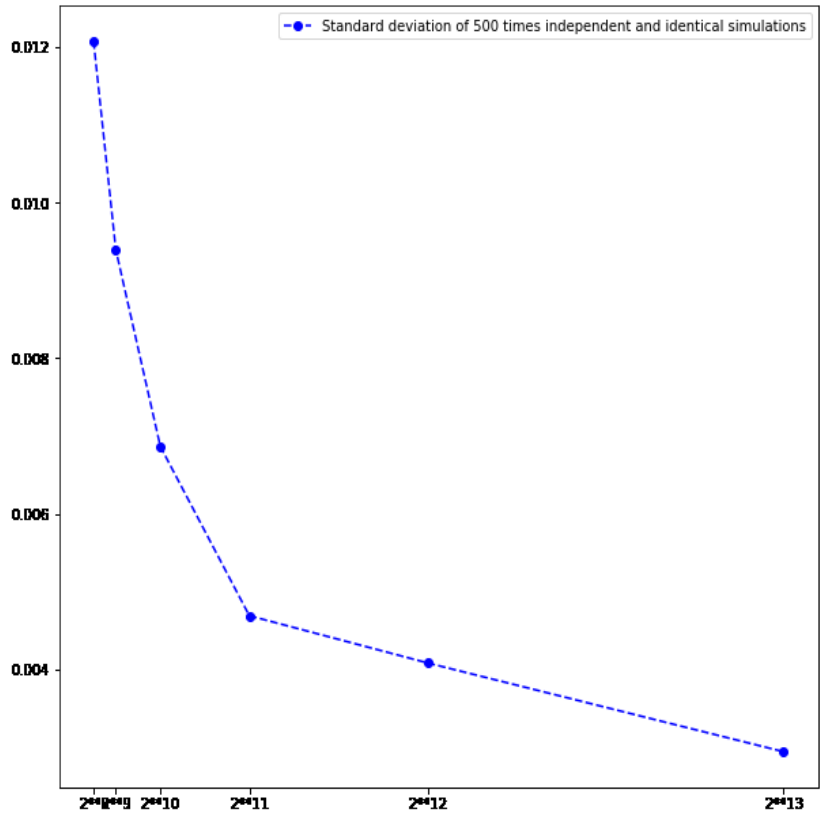


Figure 7.9 The standard deviation of the error of the particle method (Algorithm 1) with respect to N

Now we present the convergence rate of the error of the recursive quantization method (Algorithm 2) with respect to the quantizer size K for a fixed $M = 50$. We will take $K = 2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}$. Remind that here we use a fixed quantizer sequence which is a uniformly spaced point set in [-2.5, 3.5] without Lloyd I algorithm for the quantizer optimization.

K	2^5	2^6	2^7	2^8	2^9	2^{10}
Error $\|F_{\text{simu}} - F_{\text{true}}\|_{\text{sup}}$	0.07347	0.04176	0.02360	0.01471	0.01043	0.00829

Table 7.4 Error of the recursive quantization method (Algorithm 2) with respect to the quantizer size K

In the following figure we show the curve of the error with respect to K and the log

error with respect to $\log_2(K)$.

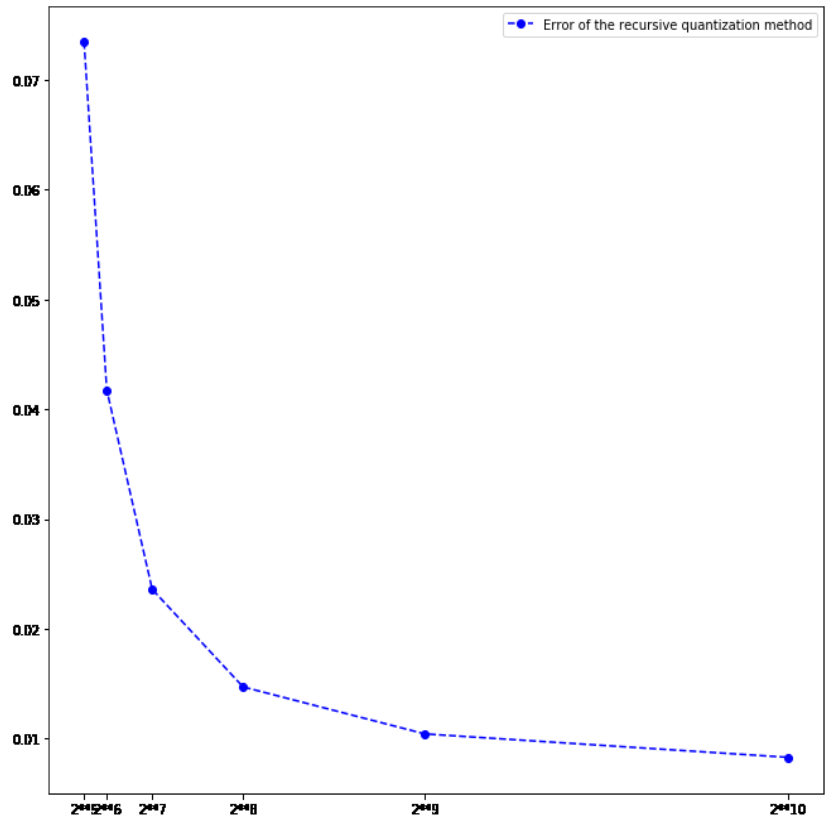


Figure 7.10 Error of the recursive quantization method (Algorithm 2) with respect to the quantizer size K

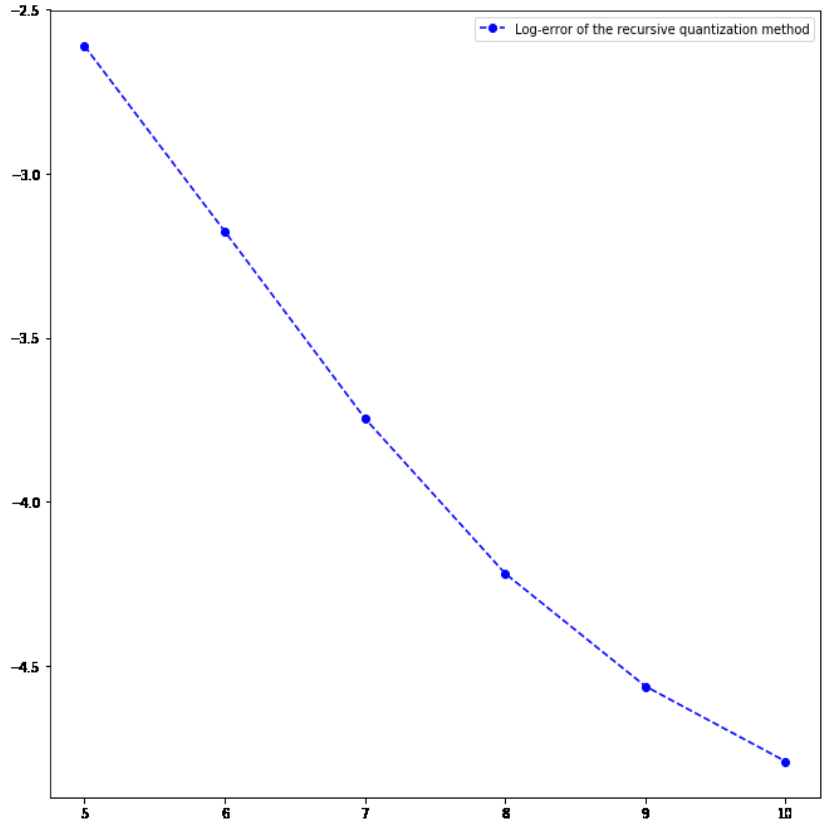


Figure 7.11 Log-error of the recursive quantization method (Algorithm 2) with respect to $\log_2(K)$. The slope is approximately equal to -0.43626.

Now we provide some comments on the performance of the numerical methods.

- Comparison of the computing time.

  The particle method (Algorithm 1) and the recursive quantization method (Algorithm 2) without Lloyd iteration are the two fastest methods. In fact, these two methods are essentially computing a Markov chain in $\mathbb{R}^N$ and $\mathbb{R}^K$ respectively. The application of the Lloyd procedure in Algorithm 2 is a little faster than in Algorithm 4 since we used the formulas showed in (7.3.8). However, in a higher dimension, the Lloyd procedure in Algorithm 4 will be faster than in Algorithm 2.

- Comparison of the accuracy computed by $\|F_{\text{simu}}(x) - F_{\text{true}}(x)\|_{\text{sup}}$.

  – Algorithm 1 and Algorithm 4 are "random" algorithms whose simulation results, including the error $\|F_{\text{simu}}(x) - F_{\text{true}}(x)\|_{\text{sup}}$, depend on ω in $(\Omega, \mathcal{F}, \mathbb{P})$. In Figure 7.7 and Figure 7.9, we display the standard deviation of errors of Algorithm 1 comparing with the errors themselves. Comparing with these two algorithms, Algorithm 2 is more robust and deterministic.

– Comparing with the particle size N in Table 7.3 and the quantizer size K in Table 7.4, one can remark that to achieve the same accuracy, we need fewer points in the quantizer than in the particle. So if we need a discrete representation of the cumulative distribution function F (or equivalently, a discrete representation of the probability distribution μ) to compute a further functional of μ, such as an integral with respect to μ, the recursive quantization based scheme provides a smaller data set (K-size quantizer and K-size weight vector) than the particle method.

– The error of Algorithm 2, especially when we implement without the Lloyd I quantizer optimization, much depends on the choice of quantizer. Generally, a practical way to choose the initial quantizer of a probability distribution μ is to use self-quantization technique for which we refer to Delattre et al. (2006), Graf and Luschgy (2000)[Section 7.1 and Section 14], Pagès and Printems (2003) and Pages et al. (2004). Another efficient trick to improve this optimization phase is to rely on a so-called "splitting method" which uses the trained quantizer of Euler step l as a initial quantizer of Euler step $l+1$.

In this one dimensional case, we did not remark the obvious advantage of the hybrid particle quantization scheme (Algorithm 4) comparing with other methods. However, in the next section, we will show that the hybrid method provides a fair balance between the accuracy and the obtained data size.

7.6.2 Simulation of the network of FitzHugh-Nagumo neurons in dimension 3

We consider the network of FitzHugh-Nagumo neurons introduced in Baladron et al. (2012):

$$dX_t = b(X_t, \mu_t)dt + \sigma(X_t, \mu_t)dB_t \tag{7.6.7}$$

with $b : \mathbb{R}^3 \times \mathcal{P}(\mathbb{R}^3) \to \mathbb{R}^3$ and $\sigma : \mathbb{R}^3 \times \mathcal{P}(\mathbb{R}^3) \to \mathbb{M}_{3\times 3}$ defined by

$$b(x,\mu) := \begin{pmatrix} x_1 - (x_1)^3/3 - x_2 + I - \int_{\mathbb{R}^3} J(x_1 - V_{rev}) z_3 \, \mu(dz) \\ c(x_1 + a - bx_2) \\ a_r \frac{T_{\max}(1-x_3)}{1+\exp\left(-\lambda(x_1 - V_T)\right)} - a_d x_3 \end{pmatrix},$$

$$\sigma(x,\mu) := \begin{pmatrix} \sigma_{ext} & 0 & -\int_{\mathbb{R}^3} \sigma_J (x_1 - V_{rev}) z_3 \, \mu(dz) \\ 0 & 0 & 0 \\ 0 & \sigma_{32}(x) & 0 \end{pmatrix},$$

with

$$\sigma_{32} \coloneqq \mathbb{1}_{x_3\in(0,1)}\sqrt{a_r\frac{T_{\max}(1-x_3)}{1+\exp\big(-\lambda(x_1-V_T)\big)}+a_d\,x_3}\,\Gamma\exp\big(-\frac{\Lambda}{1-(2x_3-1)^2}\big).$$

The probability distribution of X_0 is

$$X_0\sim\mathcal{N}\left(\begin{pmatrix}V_0\\ \omega_0\\ y_0\end{pmatrix},\begin{pmatrix}\sigma_{V_0}&0&0\\ 0&\sigma_{\omega_0}&0\\ 0&0&\sigma_{y_0}\end{pmatrix}\right)$$

with the following parameter values

$V_0=0$	$\sigma_{V_0}=0.4$	$a=0.7$	$b=0.8$	$c=0.08$	$I=0.5$	$\sigma_{ext}=0.5$
$\omega_0=0.5$	$\sigma_{\omega_0}=0.4$	$V_{rev}=1$	$a_r=1$	$a_d=1$	$T_{max}=1$	$\lambda=0.2$
$y_0=0.3$	$\sigma_{y_0}=0.05$	$J=1$	$\sigma_J=0.2$	$V_T=2$	$\Gamma=0.1$	$\Lambda=0.5.$

In this section, we compare the performance of the particle method (introduced in Section 7.1) and the hybrid method (introduced in Section 7.5) in two aspects. First, we intuitively compare these two methods by simulating the density function of (x_1,x_2) for $T=1.5$, as in the original paper Baladron et al. (2012)[Page 31, Figure 4, the third one in the right]. In this step, we choose the Euler step number $M=5000$ to reduce (as much as possible) the error of the discretization in time. In Figures 7.12, 7.13, 7.15 and 7.16, we display the images of the density function simulated by these two methods. Next, as the particle method and the hybrid method are both random methods, we take

$$\varphi(\mu_T^{\text{simu}})\coloneqq\int_{\mathbb{R}^3}|\xi|^2\,\mu_T^{\text{simu}}(d\xi)=\mathbb{E}\left|X_T^{\text{simu}}\right|^2$$

as a test function for the simulated distribution μ_T^{simu} at time T, rerun 200 times for each method and compare the mean and the standard deviation of $\varphi(\mu_T^{\text{simu}})$. As this network example is a 3-dimensional example, the doubly quantization method (introduced in Section 7.4) and the recursive quantization method (introduced in Section 7.3) are costly in the computing time (for a laptop) at present, due to the quantizer size of the normal distribution to obtain $\widehat{Z}_m$ in (H) and the integral of (7.3.5) over a Voronoï cell.

The images of the density function simulated respectively by the particle method and the hybrid method are as follows.

Particle method (Algorithm 1):

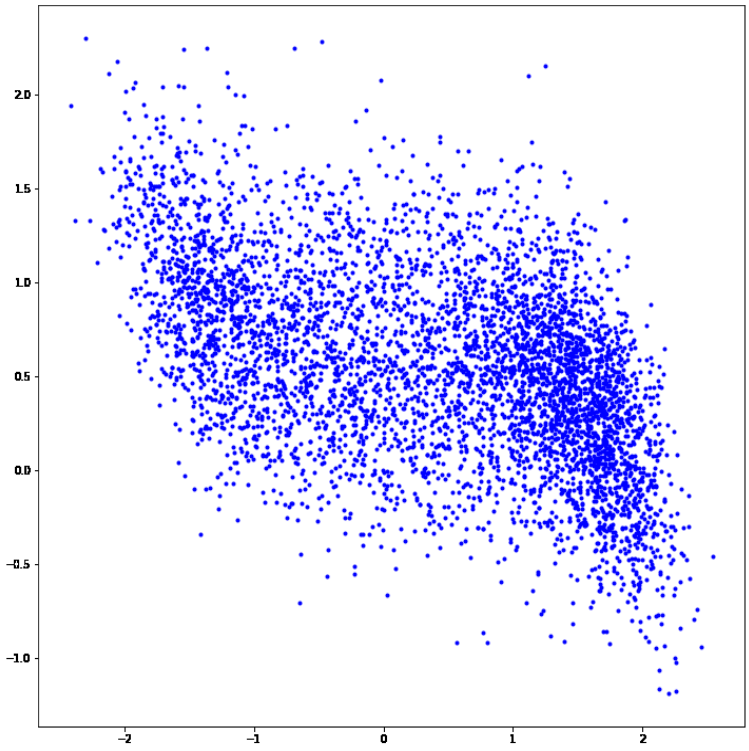


Figure 7.12 The first and second coordinates of 5000 particles at time $T = 1.5$

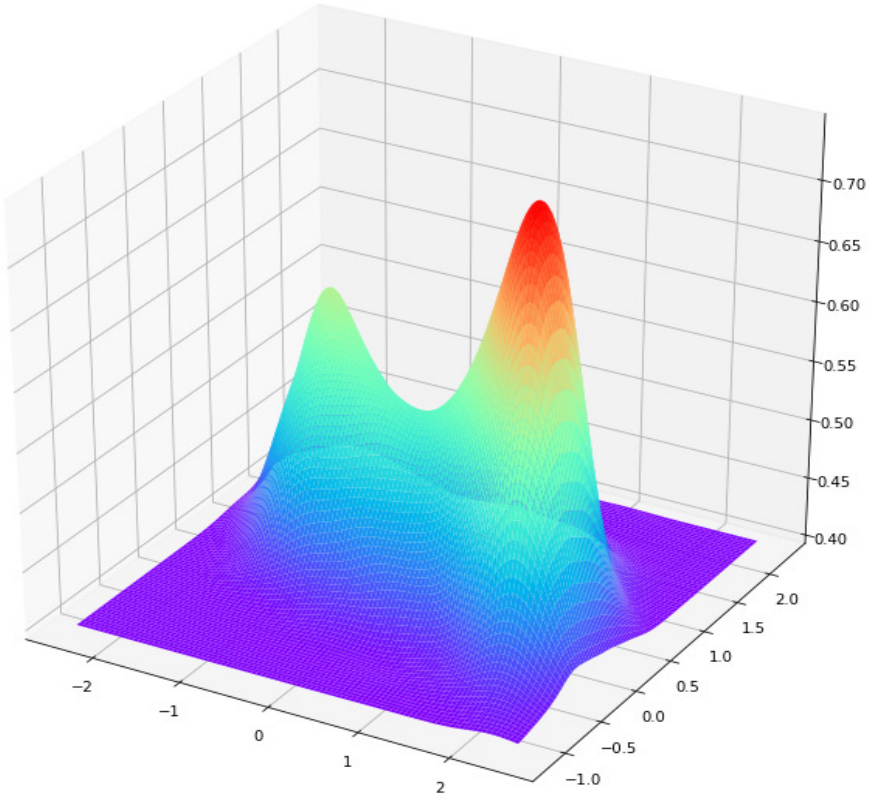


Figure 7.13 The simulated density function smoothened by the Gaussian kernel method (bandwith = 0.241)

The hybrid particle-quantization scheme (Algorithm 4):

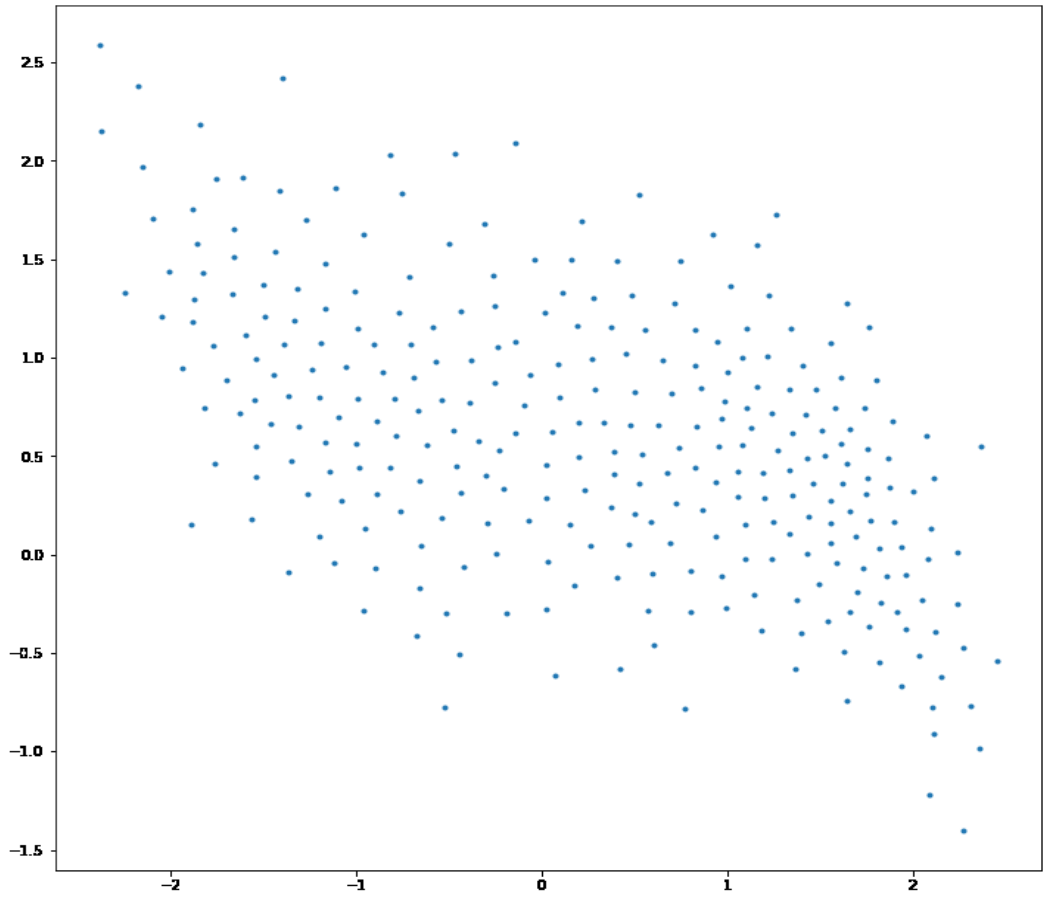


Figure 7.14 The quantizer of (x_0, x_1), simulated with particle number $N = 5000$, quantizer size $K = 300$ and 10 Lloyd iterations at each Euler step

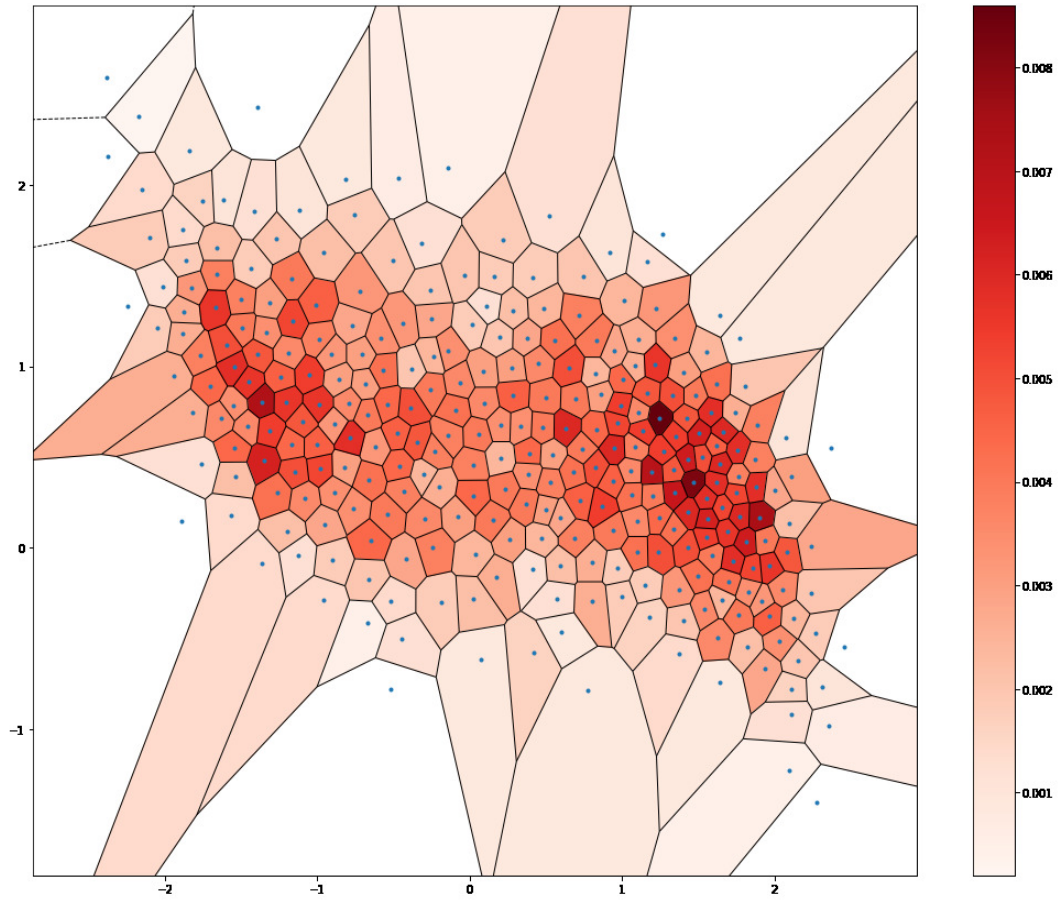


Figure 7.15 The Voronoï cells of the above quantizer. The color of each Voronoï cell represents the weight of this cell (the darker the heavier).

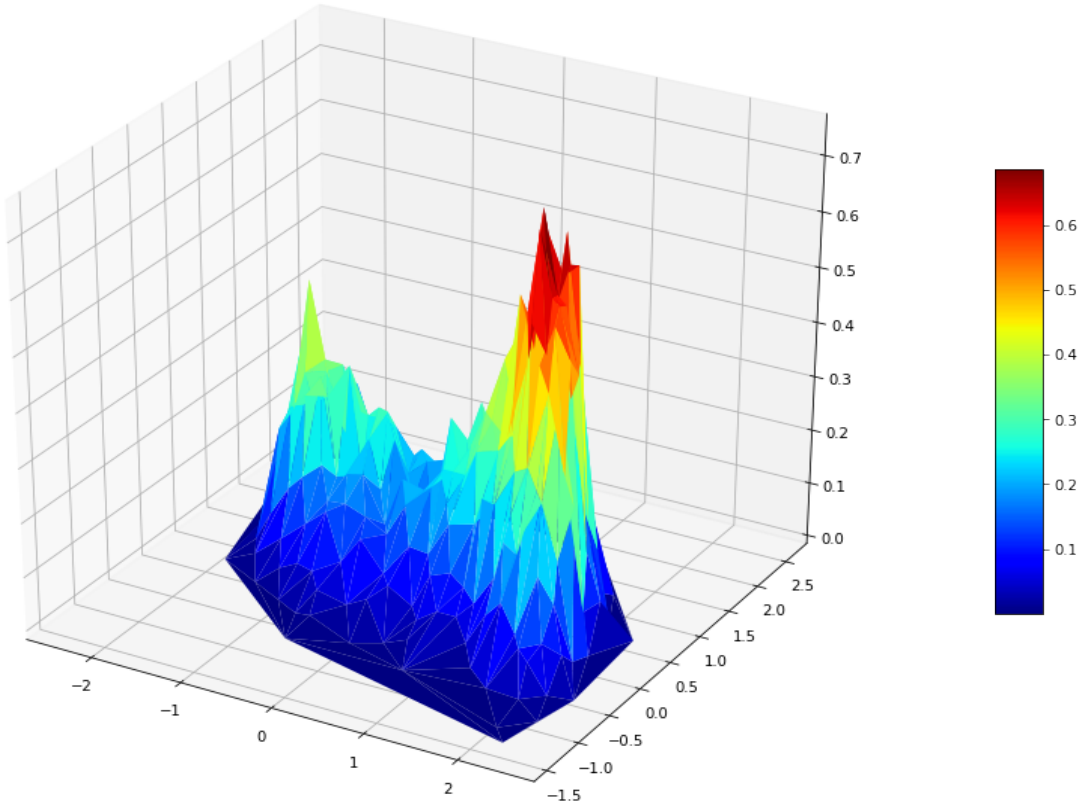


Figure 7.16 The density function simulated by Algorithm 4. The vertical axis is the weight divided by the area of the corresponding Voronoï cell.

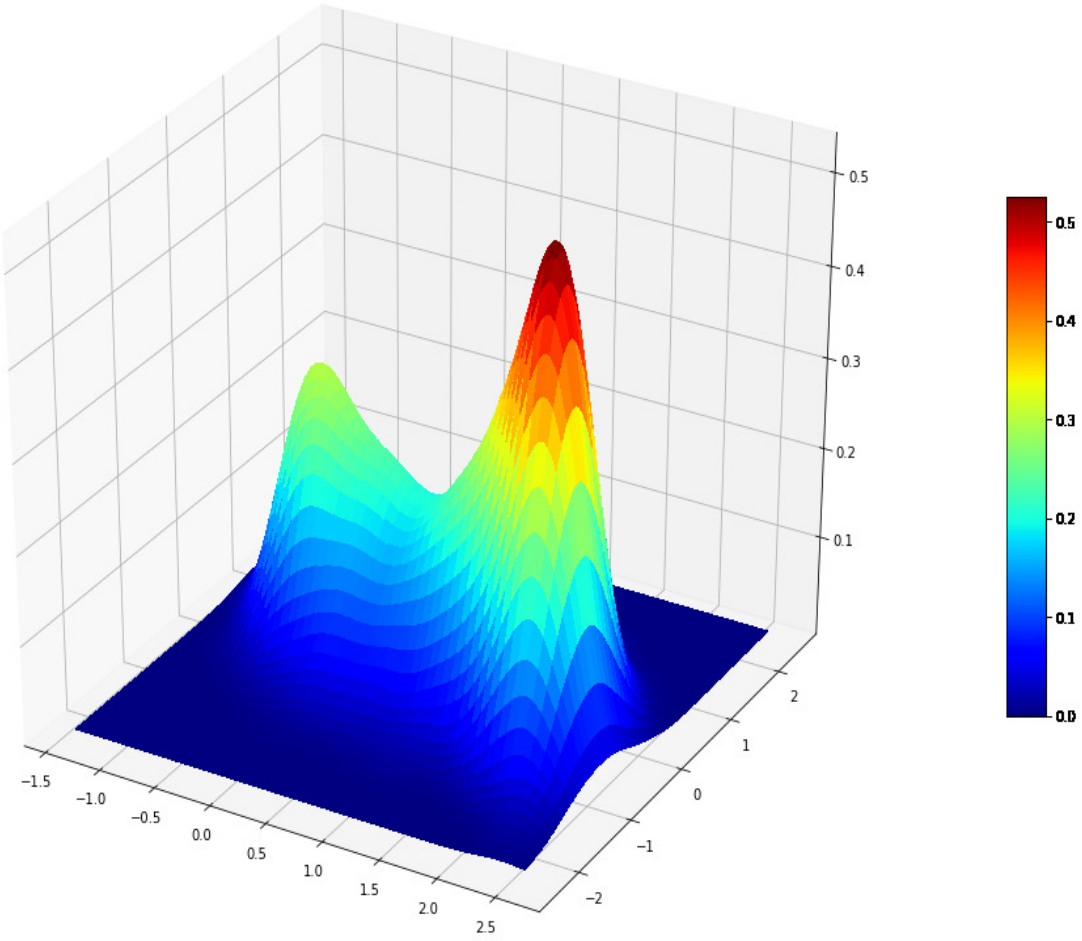


Figure 7.17 The smoothen density function of Figure 7.16 by the Gaussian kernal method (bandwidth = 0.22).

The obtained density functions have a similar form by these two methods but the data size obtained by the particle method is

$$5000 \text{ (the number of particle)} \times 3 \text{ (dimension)},$$

while the data size obtained by the hybrid method is

$$300 \text{ (the quantizer size)} \times 4 \text{ (dimension + weight for each quantizer)}.$$

For a more precise comparing, we fix now the time discretization number $M = 150$,

consider the following test function for the simulated distribution μ_T^{simu} at $T = 1.5$

$$\varphi(\mu_T^{\text{simu}}) \coloneqq \int_{\mathbb{R}^3} |\xi|^2 \, \mu_T^{\text{simu}}(d\xi) = \mathbb{E}\left|X_T^{\text{simu}}\right|^2$$

and rerun 200 times for each method. We obtain the following results.

		Particle method		Hybrid method
		N=300	N=3000	$N = 5000$ and $K = 300$
Obtained data size for μ_T^{simu}		300×3	3000×3	300×4
Average computing time for each Euler step		0.00184s	0.01097s	0.43060s
Test function $\varphi(\mu_T^{\text{simu}})$	Mean	1.43673	1.41452	1.42159
	Standard deviation	0.01907	0.00574	0.00534

Table 7.5 Comparison of the simulation result $\varphi(\mu_T^{\text{simu}})$

Intuitively, the hybrid method can be considered as adding a "feature extraction" step on the particle method. Comparing the third and fourth columns of the above table, one can remark that this added step needs more computing time but highly reduces the size of the output data size for the further computing of the test function $\varphi(\mu_T^{\text{simu}})$ without enlarging the standard deviation. However, the second column of the above table shows that if we implement the particle method with a similar data size, the computing results of $\varphi(\mu_T^{\text{simu}})$ provides a much larger standard deviation.

References

Alfonsi, A., Corbetta, J., and Jourdain, B. (2019). Sampling of one-dimensional probability measures in the convex order and computation of robust option price bounds. *International Journal of Theoretical and Applied Finance (IJTAF)*, 22(03):1–41.

Baladron, J., Fasoli, D., Faugeras, O., and Touboul, J. (2012). Mean-field description and propagation of chaos in networks of Hodgkin-Huxley and FitzHugh-Nagumo neurons. *J. Math. Neurosci.*, 2:Art. 10, 50.

Bally, V. and Pagès, G. (2003). A quantization algorithm for solving multi-dimensional discrete-time optimal stopping problems. *Bernoulli*, 9(6):1003–1049.

Bally, V., Pagès, G., and Printems, J. (2005). A quantization tree method for pricing and hedging multidimensional American options. *Math. Finance*, 15(1):119–168.

Berti, P., Pratelli, L., and Rigo, P. (2015). Gluing lemmas and Skorohod representations. *Electron. Commun. Probab.*, 20:no. 53, 11.

Biau, G., Devroye, L., and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory*, 54(2):781–790.

Bolley, F. (2008). Separability and completeness for the Wasserstein distance. In *Séminaire de probabilités XLI*, pages 371–377. Springer.

Bormetti, G., Callegaro, G., Livieri, G., and Pallavicini, A. (2018). A backward Monte Carlo approach to exotic option pricing. *European Journal of Applied Mathematics*, 29(1):146–187.

Bossy, M. and Talay, D. (1997). A stochastic particle method for the McKean-Vlasov and the Burgers equation. *Math. Comp.*, 66(217):157–192.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Bouleau, N. (1988). *Processus stochastiques et applications*. Hermann Paris.

Callegaro, G., Fiorin, L., and Grasselli, M. (2015). Quantized calibration in local volatility. *Risk Magazine*, 28(4):62–67.

Callegaro, G., Fiorin, L., and Grasselli, M. (2017). Pricing via recursive quantization in stochastic volatility models. *Quantitative Finance*, 17(6):855–872.

Chassagneux, J.-F., Szpruch, L., and Tse, A. (2019). Weak quantitative propagation of chaos via differential calculus on the space of measures. *arXiv preprint arXiv:1901.02556*.

Cuesta, J. A. and Matrán, C. (1988). The strong law of large numbers for k-means and best possible nets of Banach valued random variables. *Probab. Theory Related Fields*, 78(4):523–534.

Delattre, S., Graf, S., Luschgy, H., and Pagès, G. (2006). Quantization of probability distributions under norm-based distortion measures ii: Self-similar distributions. *Journal of mathematical analysis and applications*, 318(2):507–516.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification*. Wiley-Interscience, New York, second edition.

El Amri, M. R., Helbert, C., Lepreux, O., Zuniga, M. M., Prieur, C., and Sinoquet, D. (2019). Data-driven stochastic inversion via functional quantization. *Statistics and Computing*, pages 1–17.

El-Mikkawy, M. (2003). A note on a three-term recurrence for a tridiagonal matrix. *Appl. Math. Comput.*, 139(2-3):503–511.

Fadili, A. (2019). Ordre convexe pour les diffusions multidimensionnelles. Application aux modèles à volatilité locale. Ph.D thesis, In progress.

Fort, J.-C. and Pagès, G. (1995). On the a.s. convergence of the Kohonen algorithm with a general neighborhood function. *Ann. Appl. Probab.*, 5(4):1177–1216.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738.

Funaki, T. (1984). A certain class of diffusion processes associated with nonlinear parabolic equations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 67(3):331–348.

Gärtner, J. (1988). On the McKean-Vlasov limit for interacting diffusions. *Math. Nachr.*, 137:197–248.

Gersho, A. and Gray, R. M. (2012). *Vector quantization and signal compression*, volume 159. Springer Science & Business Media.

Gobet, E., Pagès, G., Pham, H., and Printems, J. (2005). *Discretization and simulation for a class of SPDEs with applications to Zakai and McKean-Vlasov equations*. Preprint LPMA.

Gobet, E., Pagès, G., Pham, H., and Printems, J. (2006). Discretization and simulation of the Zakai equation. *SIAM J. Numer. Anal.*, 44(6):2505–2538.

Graf, S. and Luschgy, H. (2000). *Foundations of quantization for probability distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin.

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

Grafakos, L. (2014). *Classical Fourier analysis*, volume 249 of *Graduate Texts in Mathematics*. Springer, New York, third edition.

Hopf, E. (1950). The partial differential equation $u_t + uu_x = \mu u_{xx}$. *Comm. Pure Appl. Math.*, 3:201–230.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.

IEEE Transactions on Information Theory (1982). *IEEE Trans. Inform. Theory*, 28(2).

Jourdain, B. (2000). Diffusion processes associated with nonlinear evolution equations for signed measures. *Methodol. Comput. Appl. Probab.*, 2(1):69–91.

Jourdain, B., Méléard, S., and Woyczynski, W. A. (2008). Nonlinear SDEs driven by Lévy processes and related PDEs. *ALEA Lat. Am. J. Probab. Math. Stat.*, 4:1–29.

Jourdain, B. and Pagès, G. (2019). Convex order, quantization and monotone applications of ARCH models. *In progress*.

Kelley, J. L. (1975). *General topology*. Springer-Verlag, New York-Berlin. Reprint of the 1955 edition [Van Nostrand, Toronto, Ont.], Graduate Texts in Mathematics, No. 27.

Kieffer, J. C. (1982). Exponential rate of convergence for Lloyd's method. I. *IEEE Trans. Inform. Theory*, 28(2):205–210.

Kieffer, J. C. (1983). Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function. *IEEE Trans. Inform. Theory*, 29(1):42–47.

Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

Lacker, D. (2018). Mean field games and interacting particle systems. *Preprint*.

Lacković, I. B. (1982). On the behaviour of sequences of left and right derivatives of a convergent sequence of convex functions. *Univerzitet u Beogradu. Publikacije Elektrotehničkog Fakulteta. Serija Matematika i Fizika*, pages 19–27.

Lejay, A. and Reutenauer, V. (2012). A variance reduction technique using a quantized Brownian motion as a control variate. *The Journal of Computational Finance*, 16(2):61.

Linder, T. (2002). Learning-theoretic methods in vector quantization. In *Principles of nonparametric learning (Udine, 2001)*, volume 434 of *CISM Courses and Lect.*, pages 163–210. Springer, Vienna.

Liu, Y. and Pagès, G. (2018). Convergence rate of optimal quantization grids and application to empirical measure. *arXiv preprint arXiv:1811.08351*.

Liu, Y. and Pagès, G. (2019). Characterization of probability distribution convergence in Wasserstein distance by L^p-quantization error function. *to appear in Bernoulli*.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, 28(2):129–137.

Lucchetti, R. (2006). *Convexity and well-posed problems*, volume 22 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer, New York.

Luschgy, H. and Pagès, G. (2002). Functional quantization of Gaussian processes. *J. Funct. Anal.*, 196(2):486–531.

Luschgy, H. and Pagès, G. (2008). Functional quantization rate and mean regularity of processes with an application to Lévy processes. *Ann. Appl. Probab.*, 18(2):427–469.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pages Vol. I: Statistics, pp. 281–297. Univ. California Press, Berkeley, Calif.

McKean, H. P. (1967). Propagation of chaos for a class of non-linear parabolic equations. *Stochastic Differential Equations (Lecture Series in Differential Equations, Session 7, Catholic Univ., 1967)*, pages 41–57.

Pagès, G. (1998). A space quantization method for numerical integration. *J. Comput. Appl. Math.*, 89(1):1–38.

Pagès, G. (2008). Quadratic optimal functional quantization of stochastic processes and numerical applications. In *Monte Carlo and quasi-Monte Carlo methods 2006*, pages 101–142. Springer, Berlin.

Pagès, G. (2015). Introduction to vector quantization and its applications for numerics. *CEMRACS 2013—modelling and simulation of complex systems: stochastic and deterministic approaches*, 48:29–79.

Pagès, G. (2016). Convex order for path-dependent derivatives: a dynamic programming approach. In *Séminaire de Probabilités XLVIII*, pages 33–96. Springer.

Pagès, G. (2018). *Numerical Probability: An Introduction with Applications to Finance*. Springer.

Pages, G., Pham, H., and Printems, J. (2004). An optimal markovian quantization algorithm for multi-dimensional stochastic control problems. *Stochastics and dynamics*, 4(04):501–545.

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

Pagès, G. and Sagna, A. (2012). Asymptotics of the maximal radius of an L^r-optimal sequence of quantizers. *Bernoulli*, 18(1):360–389.

Pagès, G. and Sagna, A. (2015). Recursive marginal quantization of the Euler scheme of a diffusion process. *Appl. Math. Finance*, 22(5):463–498.

Pagès, G. and Sagna, A. (2018). Improved error bounds for quantization based numerical schemes for BSDE and nonlinear filtering. *Stochastic Process. Appl.*, 128(3):847–883.

Pagès, G. and Yu, J. (2016). Pointwise convergence of the Lloyd I algorithm in higher dimension. *SIAM J. Control Optim.*, 54(5):2354–2382.

Pollard, D. (1981). Strong consistency of k-means clustering. *Ann. Statist.*, 9(1):135–140.

Pollard, D. (1982a). A central limit theorem for k-means clustering. *Ann. Probab.*, 10(4):919–926.

Pollard, D. (1982b). Quantization and the method of k-means. *IEEE Trans. Inform. Theory*, 28(2):199–205.

Reis, G. d., Engelhardt, S., and Smith, G. (2018). Simulation of McKean-Vlasov SDEs with super linear growth. *arXiv preprint arXiv:1808.05530*.

Rudin, W. (1991). *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York, second edition.

Sznitman, A.-S. (1991). Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX—1989*, volume 1464 of *Lecture Notes in Math.*, pages 165–251. Springer, Berlin.

Topsoe, F. (1974). Compactness and tightness in a space of measures with the topology of weak convergence. *Math. Scand.*, 34:187–210.

Trushkin, A. V. (1982). Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions. *IEEE Trans. Inform. Theory*, 28(2):187–198.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

Villani, C. (2003). *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI.

Villani, C. (2009). *Optimal transport, Old and new*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.