Optimal dual quantizers of 1D log-concave distributions: uniqueness and Lloyd like algorithm

Benjamin Jourdain, Gilles Pagès

Version of Record: https://www.sciencedirect.com/science/article/pii/S0021904521000447
Manuscript_8650c43ecf995d74bc18e7921793abe4

Optimal dual quantizers of $1D$ log-concave distributions: uniqueness and Lloyd like algorithm *

Benjamin Jourdain † Gilles Pagès ‡

Abstract

We establish for dual quantization the counterpart of Kieffer's uniqueness result for compactly supported one dimensional probability distributions having a log-concave density (also called strongly unimodal): for such distributions, L^r-optimal dual quantizers are unique at each level N, the optimal grid being the unique critical point of the quantization error. An example of non-strongly unimodal distribution for which uniqueness of critical points fails is exhibited. In the quadratic $r = 2$ case, we propose an algorithm to compute the unique optimal dual quantizer. It provides a counterpart of Lloyd's method I algorithm in a Voronoi framework (see [13, 14]). Finally semi-closed forms of L^r-optimal dual quantizers are established for power distributions on compacts intervals and truncated exponential distributions.

1 Introduction

Optimal Delaunay or dual quantization has been introduced in [19] in a one dimensional setting for probabilistic numerical purposes, in order to produce a fast algorithm for pricing credit derivative products in finance. It was then developed in higher dimension in [20] as a possible alternative to optimal Voronoi (or primal) quantization (see [5, 6, 17, 16] for introduction) to solve various non-linear problems in quantitative finance (American option pricing and δ-hedging, stochastic control for portfolio management, etc). Both quantization modes are spatial discretization methods of probability distributions or random vectors, one relying on Voronoi diagrams and the other on Delaunay triangulation (in 2-dimension). Delaunay quantization is limited to compactly supported distributions but shares a universal "stationarity property" (see further on) which makes it much more flexible when used as a numerical tool. This paper is essentially focused on the 1-dimensional setting. Our aim is to prove for Delaunay quantization some uniqueness and convergence results related to optimal quantizers and their numerical computation for strongly unimodal distributions known in optimal Voronoi quantization as Trushkin's and Kieffer's theorems (see e.g. [23] and [11] respectively).

Let us briefly explain what Delaunay quantization is in a one dimensional setting. It answers the question: how to spatially discretize a compactly supported random variable with support (contained in) $[a, b]$ using a finite subset $\Gamma = \{x_1, \ldots, x_N\} \subset [a, b]$ with $x_1 = a < x_2 < \cdots <$

*This research benefited from the support of the "Chaire Risques Financiers", Fondation du Risque
†CERMICS, Ecole des Ponts, INRIA, Marne-la-Vallée, France. E-mail: benjamin.jourdain@enpc.fr
‡Laboratoire de Probabilités, Statistique et Modélisation, UMR 8001, Campus Pierre et Marie Curie, Sorbonne Université case 158, 4, pl. Jussieu, F-75252 Paris Cedex 5, France. E-mail: gilles.pages@upmc.fr

$x_{N-1} < x_N = b$. The basic idea to discretize a random variable X is the following: when $X : (\Omega, \mathcal{A}, \mathbb{P}) \to [a, b]$ falls into the interval $[x_i, x_{i+1}]$, one replaces the value of X by $\widehat{X}$ which takes values x_i, x_{i+1} with respective probabilities $\frac{x_{i+1}-X}{x_{i+1}-x_i}$ and $\frac{X-x_i}{x_{i+1}-x_i}$. These probabilities come as the coefficients when writing X as of the linear interpolation of x_i and x_{i+1} since

$$X = \frac{x_{i+1} - X}{x_{i+1} - x_i} x_i + \frac{X - x_i}{x_{i+1} - x_i} x_{i+1}.$$

This leads to introduce the so-called *Delaunay* projection (or *splitting operator*) defined for every $\xi \in [a, b]$ and $u \in (0, 1)$ by

$$\mathrm{Proj}_\Gamma^{del}(\xi, u) = a\mathbf{1}_{\{a\}}(\xi) + \sum_{i=1}^{N-1} \left[x_i \cdot \mathbf{1}_{\left\{0< u< \frac{x_{i+1}-\xi}{x_{i+1}-x_i}\right\}} + x_{i+1} \cdot \mathbf{1}_{\left\{\frac{x_{i+1}-\xi}{x_{i+1}-x_i} \le u<1\right\}} \right] \mathbf{1}_{(x_i, x_{i+1}]}(\xi) \quad (1.1)$$

so that

$$\widehat{X} = \widehat{X}^{\Gamma,dual} = \mathrm{Proj}_\Gamma^{del}(X, U) \quad \text{with} \quad U \sim U\big((0,1)\big),\ U \perp\!\!\!\perp X$$

where $\perp\!\!\!\perp$ stands for independence. The above formula can be taken as a definition for a Γ-valued dual quantizer of X.

Note that, owing to the above remark that aimed at its construction

$$\forall\, i \in \{1, \dots, N\}, \quad \int_0^1 \mathrm{Proj}_\Gamma^{del}(\xi, u) du = \xi$$

or, equivalently,

$$\mathbb{E}\big(\widehat{X} \,|\, X\big) = X \quad (1.2)$$

which is a stationarity property dual from that satisfied by quadratic optimal primal (or *Voronoi*) quantization (see (1.3) below). In particular $X \le_{cvx} \widehat{X}$ (convex ordering). Applications of dual quantization were first mostly devoted to provide efficient numerical schemes and fast algorithms to solve non-linear problems arising in numerical probability applied to finance like the pricing and hedging of multi-asset American style options (see e.g. [21]) or the the pricing of credit derivatives (see [19]), basically as a competitor of Voronoi quantization and other methods (regressions, Malliavin Monte Carlo). Its dual behaviour with respect to convex order provides an informal way to provide lower and upper-bounds in various stochastic control problems. More recently, with the development of martingale optimal transport problems in finance, both Voronoi and Delaunay quantization methods have been shown as a systematic tool to design time discretization schemes that preserve convex order (see [10]) and more generally to solve numerically discrete time martingale optimal transport problems (see [9]) which turns out to be a quite challenging problem (see [1], [3], [4], [7], [8]).

The distribution of $\widehat{X}^{\Gamma,dual}$ is entirely characterized by its value set Γ and the weights $p_i^\Gamma = \mathbb{P}(\widehat{X}^{\Gamma,dual} = x_i)$ given for every $i = 1, \dots, N$, by

$$p_i^\Gamma = \mathbf{1}_{\{i=1\}} \mathbb{P}(X = a) + \mathbf{1}_{\{i\neq 1\}} \mathbb{E}\, \tfrac{X - x_{i-1}}{x_i - x_{i-1}} \mathbf{1}_{\{X \in (x_{i-1}, x_i]\}} + \mathbf{1}_{\{i\neq N\}} \mathbb{E}\, \tfrac{x_{i+1} - X}{x_{i+1} - x_i} \mathbf{1}_{\{X \in (x_i, x_{i+1}]\}}.$$

If we introduce the cumulative distribution function (c.d.f.) $F(x) = \mathbb{P}\big(X \in (-\infty, x]\big)$ and the first partial moment $K(x) = \mathbb{E}\, X\mathbf{1}_{X\in(-\infty,x]}$, then these weights write

$$\begin{aligned} p_i^\Gamma &= \mathbf{1}_{\{i=1\}}\mathbb{P}(X=a) + \mathbf{1}_{\{i\neq 1\}}\int_{(x_{i-1},x_i]} \tfrac{\xi-x_{i-1}}{x_i-x_{i-1}}\mu(d\xi) + \mathbf{1}_{\{i\neq N\}}\int_{(x_i,x_{i+1}]} \tfrac{x_{i+1}-\xi}{x_{i+1}-x_i}\mu(d\xi) \\ &= \mathbf{1}_{\{i=1\}}F(a) + \mathbf{1}_{\{i\neq 1\}}\frac{[K]_{x_{i-1}}^{x_i} - x_{i-1}[F]_{x_{i-1}}^{x_i}}{x_i - x_{i-1}} + \mathbf{1}_{\{i\neq N\}}\frac{[F]_{x_i}^{x_{i+1}}x_{i+1} - [K]_{x_i}^{x_{i+1}}}{x_{i+1}-x_i} \end{aligned}$$

where, for simplicity, we will denote for a function $g : \mathbb{R} \to \mathbb{R}$ and two real numbers $x \le y$, $[g]_x^y = g(y) - g(x)$.

The L^r-mean error induced by replacing X by its dual quantization $\widehat{X}^{\Gamma,dual}$ is naturally defined by

$$\begin{aligned} \|X - \widehat{X}^{\Gamma,dual}\|_r^r &= \int_a^b \mathbb{E}\,|\xi - \mathrm{Proj}_x^{del}(\xi, U)|^r \mu(d\xi) \\ &= \sum_{i=1}^{N-1}\int_{(x_i,x_{i+1})}\left((\xi-x_i)^r\frac{x_{i+1}-\xi}{x_{i+1}-x_i} + (x_{i+1}-\xi)^r\frac{\xi-x_i}{x_{i+1}-x_i}\right)\mu(d\xi). \end{aligned}$$

The basic application of dual quantization, like its historical counterpart in the Voronoi sense, is to produce quadrature formulae adapted to the distribution μ of the random variable X since for a function $g : \mathbb{R} \to \mathbb{R}$ and a Γ-quantization $\widehat{X}^\Gamma$ of X

$$\mathbb{E}\, g(X) \simeq \mathbb{E}\, g\big(\widehat{X}^\Gamma\big) = \sum_{i=1}^N p_i^\Gamma\, g(x_i)$$

where the weights p_i^Γ depend on the quantization mode (primal or dual). If $X \in L^2(\mathbb{P})$, and g is C^1 and its gradient is Lipschitz continuous with constant $[\nabla g]_{\mathrm{Lip}}$, writing

$$g(X) - g(\widehat{X}) = \int_0^1 \Big(\nabla g(\widehat{X} + \alpha(X-\widehat{X})) - \nabla g(X)\Big).(X-\widehat{X})d\alpha + \nabla g(X).(X-\widehat{X})$$

and using the stationarity property (1.2) to get rid of the expectation of the second term in the right-hand side, one obtains

$$\big|\mathbb{E}\, g(X) - \mathbb{E}\, g(\widehat{X})\big| \le \tfrac{1}{2}[\nabla g]_{\mathrm{Lip}}\mathbb{E}|X-\widehat{X}|^2$$

(see [21]). Note that the counterpart of such a second order error bound in Voronoi quantization *only holds for optimal quadratic quantizers*. Nevertheless, this quadrature formula emphasizes the need for quantizers inducing, at a given level $N \ge 3$, an as small as possible quantization error. That is a grid Γ such that $\mathbb{E}|X-\widehat{X}|^2$ is minimum. This is the main purpose of optimal quantization and in fact such optimal "minimizing" grids do exist, see Theorem 1.1 below. Let us specify the example of optimal quantizations for $\mathcal{U}([0,1])$. It follows from [20], Section 5.1 (see also the remark after Theorem 2.1), that the L^r-optimal dual quantizer of $\mathcal{U}([0,1])$ (does not depend on r and) is given at level $N \ge 2$ by

$$\Gamma^{(N),del} = \Big\{\tfrac{i-1}{N-1} : i = 1, \dots, N\Big\}$$

with weights $p_1 = p_N = \frac{1}{2(N-1)}$ and $p_2 = p_3 = \cdots = p_{N-1} = \frac{1}{N-1}$ whereas, for Voronoi quantization, the L^r-optimal quantizer of $\mathcal{U}([0,1])$ does not depend on r either and is given (see [16]) at level $N \geq 1$ by

$$\Gamma^{(N),vor} = \Big\{ \tfrac{2i-1}{2N} : i = 1, \ldots, N \Big\}$$

with all weights given by $p_i = \frac{1}{N}$. Note that the optimal Voronoi N-quantizer is made up with the midpoints of the optimal Delaunay $(N+1)$-quantizer. Consequently, in this elementary framework, Voronoi optimal N-quantizers correspond to midpoint quadrature formula for numerical integration over $[0,1]$ whereas Delaunay quantization yields the trapezoid quadrature formula. Such a property no longer holds for general distributions.

When X is an $\mathbb{R}^d$-valued random vector with compactly supported distribution μ, $d \geq 2$, one considers grids $\Gamma = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ such that $\mathrm{supp}(\mu) \subset \mathrm{conv}(\Gamma)$ and the Delaunay projection operator is defined on a hyper-triangulation of $\mathrm{conv}(\Gamma)$ sharing some minimality properties. The main feature of such dual quantization in higher dimension is that it still satisfies for every grid Γ the above dual stationarity property. This has been established in [20] in full generality with a natural extension to unbounded random vectors (to the price of a partial loss of the stationarity property).

Then, for any fixed $r \in [1, +\infty)$, one may define the lowest possible L^r-error induced by replacing X by any of its dual quantization $\widehat{X}^{\Gamma,dual}$ where Γ runs over grids of size (or cardinality) at most N. To keep sense one should assume that $N \geq d_\mu + 1$ where d_μ denotes the dimension of the vector space spanned by $\mathrm{supp}(\mu)$ in $\mathbb{R}^d$. So we define for $N \geq d_\mu + 1$, the L^r dual quantization error modulus by

$$d_{r,N}(X) := \inf \big\{ \big\| X - \widehat{X}^{\Gamma,dual} \big\|_r,\ \mathrm{conv}(\Gamma) \supset \mathrm{supp}(\mu),\ \mathrm{card}(\Gamma) \leq N \big\}.$$

It turns out (see again [20]) that it satisfies the more general bound

$$d_{r,N}(X) = \inf_Y \Big\{ \big\| X - Y \big\|_r : Y : (\Omega \times \Omega_0, \mathcal{A} \otimes \mathcal{A}_0, \mathbb{P} \otimes \mathbb{P}_0) \to \mathbb{R}^d, \\ \mathrm{card}(Y(\Omega \times \Omega_0)) \leq N \text{ and } \mathbb{E}_{\mathbb{P}\otimes\mathbb{P}_0}(Y|X) = X \Big\}.$$

which emphasizes the connections with martingale optimal transport explored in other papers [9, 10] on the one hand and with Voronoi/primal quantization.

Indeed if one replaces the above Delaunay projection by a (Borel) *nearest neighbour projection* on the grid Γ, denoted $\mathrm{Proj}_\Gamma^{vor}$ and if we set if $\widehat{X}^{\Gamma,vor} = \mathrm{Proj}_\Gamma^{vor}(X)$ for some L^r-integrable random vector, then

$$\begin{aligned} e_{r,N}(X) &:= \inf \big\{ \big\| X - \widehat{X}^{\Gamma,vor} \big\|_r,\ \mathrm{card}(\Gamma) \leq N \big\} \\ &= \inf_Y \Big\{ \big\| X - Y \big\|_r : Y : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^d,\ \mathrm{card}\big(Y(\Omega)\big) \leq N \Big\}. \end{aligned}$$

One has $e_{r,n}(X) \leq d_{r,N}(X)$ when both moduli make sense since dual quantization takes into account the additional martingale transport property between X and its quantization. Note that in fact both $d_{r,N}(X)$ and $e_{r,N}(X)$ only depend on the distribution, say μ, of X so that we will also denote $d_{r,N}(\mu)$ (and $e_{r,N}(\mu)$).

It is classical background (see [6] or [17]) that the infimum is in fact a minimum and that at each *level* N there exist an optimal grid $\Gamma_N^{r,vor}$ such that $e_{r,N}(X) = \big\|X - \widehat{X}^{\Gamma_N^{r,vor},vor}\big\|_r$. It should be noticed that, whereas all dual quantizations satisfy the above stationarity equation (1.2), only L^r-optimal Voronoi quantizers with $r = 2$ satisfy a stationarity property, namely a reverse one

$$\mathbb{E}\big(X \,|\, \widehat{X}^{\Gamma_N^{2,vor}vor}\big) = \widehat{X}^{\Gamma_N^{2,vor}vor}. \tag{1.3}$$

Likewise, as soon as $d_{r,N}(X) < +\infty$, it is established in [20] that $d_{r,N}(X)$ holds as a minimum i.e. $d_{r,N}(X) = \big\|X - \widehat{X}^{\Gamma_N^{r,dual},dual}\big\|_r$. To be more precise we state the original existence result for dual quantization, see [20].

Theorem 1.1 (Existence of optimal dual quantizers) *Let $r \in [1, +\infty)$ and let μ be a compactly supported distribution on $(\mathbb{R}^d, \mathcal{B}or(\mathbb{R}^d))$. For every level $N \geq d_\mu + 1$, there exists at least one L^r-optimal grid $\Gamma_N^{r,del}$ with size at most N i.e. satisfying*

$$d_{r,N}(\mu) = \left(\int_{\mathbb{R}^d\times[0,1]} |\xi - \mathrm{Proj}^{del}_{\Gamma_N^{r,del}}(\xi,u)|^r \mu(d\xi)du\right)^{1/r} = \big\|X - \mathrm{Proj}^{del}_{\Gamma_N^{r,del}}(X,U)\big\|_r,\ (X,U) \sim \mu\otimes U([0,1]).$$

Moreover $\mathrm{conv}\big(\Gamma_N^{r,del}\big) \supset \mathrm{supp}(\mu)$. *If* $\mathrm{supp}(\mu)$ *has at least N elements, then $\Gamma_N^{r,del}$ has full size N and $d_{r,N}(\mu)$ decreases to 0 as long as it does not vanish, which never occurs if* $\mathrm{supp}(\mu)$ *is infinite.*

Finally, we recall below the main result established in [22] which is counterpart for dual quantization of the celebrated Zador theorem ruling the sharp rate of decay to 0 of the optimal L^r-quantization error and its non-asymptotic version, counterpart of Pierce's lemma. It rules the dual quantization error rate in a quite similar way for bounded random vectors.

Theorem 1.2 (Rate of decay of optimal dual quantization) (a) Sharp rate for dual quantization*: Let $X \in L^\infty_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ be a bounded random vector with distribution $\mathbb{P}_X = \varphi.\lambda_d \overset{\perp}{+} \nu_X$ where λ_d denotes the Lebesgue measure and ν_X denotes its singular component. Then, for every $r \in (0, +\infty)$,*

$$\lim_{N\to+\infty} N^{\frac{1}{d}} d_{r,N}(X) = \widetilde{J}^{del}_{d,r} \left(\int_{\mathbb{R}^d} \varphi^{\frac{d}{d+r}} d\lambda_d\right)^{\frac{1}{d}+\frac{1}{r}}$$

where $\widetilde{J}^{del}_{d,r} = \inf_{N\geq 1} N^{\frac{1}{d}} d_{r,N}\big(\mathcal{U}([0,1]^d)\big) \geq \widetilde{J}^{vor}_{d,r} = \inf_{N\geq 1} N^{\frac{1}{d}} e_{r,N}\big(\mathcal{U}([0,1]^d)\big)$.

When $d = 1$, $\widetilde{J}^{del}_{1,r} = \left(\frac{2}{(r+1)(r+2)}\right)^{1/r}$ whereas $\widetilde{J}^{vor}_{d,r} = \left(\frac{1}{(r+1)2^r}\right)^{1/r}$. Hence, $\frac{\widetilde{J}^{del}_{1,r}}{\widetilde{J}^{vor}_{1,r}} = \left(\frac{2^{r+1}}{r+2}\right)^{1/r} \uparrow 2$ as $r \uparrow +\infty$.

(b) Non-asymptotic bound*: Let $r, \eta > 0$. For every dimension $d \geq 1$, there exists a real constant $\widetilde{C}^{del}_{d,\eta,r} > 0$ such that, for every random vector $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^d$, $L^\infty(\mathbb{P})$-bounded,*

$$d_{r,N}(X) \leq \widetilde{C}^{del}_{d,\eta,r} N^{-\frac{1}{d}} \sigma_{r+\eta}(X) \tag{1.4}$$

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

Remark. Note that claim (b) remains true if the support of $\mathbb{P}_X$ does not span $\mathbb{R}^d$ as an affine space, but A_μ with dimension $d' < d$. However, if such is the case (1.4) is suboptimal since it still holds with $N^{-1/d'}$ replacing $N^{-1/d}$ in the right-hand side.

One of the first striking theoretical results on optimal Voronoi quantization, beyond the existence of optimal quantizers for general distributions in any dimension and at any level, was obtained by Trushkin who proved (see [23], see also [11]) the following uniqueness result for one dimensional strongly unimodal distributions.

Theorem 1.3 (Trushkin, 1982) *Let $r \in \{1, 2\}$. Assume that μ has a finite r-moment and $\mu = f.\lambda$ where $f : \mathbb{R} \to \mathbb{R}_+$ is a* log*-concave density on the real line. For every integer $N \in \mathbb{N}$, there exists a unique L^r-optimal grid $\Gamma_N^{r,vor} = \{x_1, \ldots, x_N\} \subset \mathrm{conv}\big(\mathrm{supp}(\mu)\big)$ of size N such that for $X \sim \mu$,*

$$e_{r,N}(X) = \big\|X - \widehat{X}^{\Gamma_N^{r,vor},vor}\big\|_r.$$

For extensions to more general loss functions see again [23] or [2]. As a second step, Kieffer established in the quadratic case $r = 2$, the convergence of the so-called Lloyd's Method I (or Lloyd's algorithm, see [13]) at an exponential rate for strongly unimodal distributions whose log-density is not piecewise affine i.e. for strongly log-concave densities (see [11]).

The first main result of this paper is to prove that Trushkin's uniqueness theorem remains true for L^r-dual quantization under the same strong unimodal assumption, even for any $r \geq 1$. Then, we propose, still in 1-dimension, a kind of counterpart of the Lloyd's Method 1, to compute optimal quadratic dual quantizers and we prove that this algorithm converges at an exponential rate, uniformly in the starting point, under a strong unimodality property.

Finally, we also provide more specific fast algorithms to compute L^r-optimal dual quantizers for two families of distributions : power distributions over a compact interval and truncated exponential distributions.

2 Uniqueness of optimal scalar L^r-dual quantizers $r \geq 1$

Our aim is to establish uniqueness of L^r-optimal dual quantizers for every $r \geq 1$ under suitable assumptions on μ. We exclude the trivial case when $\mu(d\xi) = \delta_x(d\xi)$ for some $x \in \mathbb{R}$ and the optimal grid at each level N is $\{x\}$. To enable dual quantization, we suppose that μ is compactly supported and denote by $a < b$ the real numbers such that $[a, b] = \mathrm{conv}\big(\mathrm{supp}(\mu)\big)$. For $N \geq 2$, a dual quantization grid Γ with $\mathrm{card}(\Gamma) \leq N$ writes $\Gamma = \{x_1, \ldots, x_N\}$ for $x_1 \leq x_2 \leq \ldots \leq x_N$ satisfying $x_1 \leq a$ and $x_N \geq b$. Then, since $\mu(\mathbb{R} \setminus [a, b]) = 0$, for $X \sim \mu$,

$$\|X - \widehat{X}^{\Gamma,dual}\|_r^r = \sum_{i=1}^{N-1} \int_{(x_i \vee a, x_{i+1} \wedge b)} \left(|\xi - x_i|^r \frac{x_{i+1} - \xi}{x_{i+1} - x_i} + |x_{i+1} - \xi|^r \frac{\xi - x_i}{x_{i+1} - x_i} \right) \mu(d\xi).$$

When $\xi \in (x_i \vee a, x_{i+1} \wedge b)$, then, by convexity of $x \mapsto |\xi - x|^r$, one has

$$|\xi - x_i \vee a|^r \le |\xi - x_i|^r \frac{x_{i+1} - x_i \vee a}{x_{i+1} - x_i} + |x_{i+1} - \xi|^r \frac{x_i \vee a - x_i}{x_{i+1} - x_i}$$

$$|\xi - x_{i+1} \wedge b|^r \le |\xi - x_i|^r \frac{x_{i+1} - x_{i+1} \wedge b}{x_{i+1} - x_i} + |x_{i+1} - \xi|^r \frac{x_{i+1} \wedge b - x_i}{x_{i+1} - x_i} \text{ so that}$$

$$|\xi - x_i \vee a|^r \frac{x_{i+1} \wedge b - \xi}{x_{i+1} \wedge b - x_i \vee a} + |x_{i+1} \wedge b - \xi|^r \frac{\xi - x_i \vee a}{x_{i+1} \wedge b - x_i \vee a} \le |\xi - x_i|^r \frac{x_{i+1} - \xi}{x_{i+1} - x_i} + |x_{i+1} - \xi|^r \frac{\xi - x_i}{x_{i+1} - x_i}$$

with strict first inequality if $x_i < a$ (by strict convexity of $x \mapsto |\xi - x|^r$ when $r > 1$ and, when $r = 1$, since $\xi - x_i$ and $x_{i+1} - \xi$ have opposite signs), strict second inequality if $x_{i+1} > b$ and therefore strict third inequality if $x_i < a$ or $x_{i+1} > b$. Hence $\|X - \widehat{X}^{\Gamma,dual}\|_r^r$ is not smaller than

$$\sum_{i=1}^{N-1} \int_{(x_i \vee a, x_{i+1} \wedge b)} \left(|\xi - x_i \vee a|^r \frac{x_{i+1} \wedge b - \xi}{x_{i+1} \wedge b - x_i \vee a} + |x_{i+1} \wedge b - \xi|^r \frac{\xi - x_i \vee a}{x_{i+1} \wedge b - x_i \vee a} \right) \mu(d\xi),$$

and even larger if there exists $i \in \{1, \dots, N-1\}$ such that $x_i < a < x_{i+1}$ or $x_i < b < x_{i+1}$ (because of the support condition, μ gives positive weight to any interval $[a, x]$ and $[x, b]$ with $a < x < b$). Therefore the optimal grid for $N = 2$ is $\{a, b\}$ and, when $N \ge 3$, the grid $\{a \vee x_1 \wedge b, \dots, a \vee x_N \wedge b\}$ outperforms Γ or performs as well but contains at most $N-1$ points. If $\mathrm{supp}(\mu)$ contains at least N points, by Theorem 1.1, any optimal grid with size at most N contains N points and we deduce that such a grid writes $\{a, x_2, \dots, x_{N-1}, b\}$ with $a < x_2 < \ldots < x_{N-1} < b$. Let $\mathcal{S}_N^{a,b} = \{(x_2, \dots, x_{N-1}) \in (a, b)^{N-2} : x_2 < \ldots < x_{N-1}\}$ with closure $\overline{\mathcal{S}}_N^{a,b} = \{(x_2, \dots, x_{N-1}) \in [a, b]^{N-2} : x_2 \le \ldots \le x_{N-1}\}$ and for $x = (x_2, \dots, x_{N-1}) \in \mathcal{S}_N^{a,b}$,

$$L_N(x) := \|X - \widehat{X}^{\Gamma,dual}\|_r^r \text{ where } \Gamma = \{a, x_2, \dots, x_{N-1}, b\}.$$

We will of course use the natural convention $x_1 = a$ and $x_N = b$ in what follows. The optimal grids, which exist according to Theorem 1.1, are of the form $\{a, x_2, \dots, x_{N-1}, b\}$ with $x = (x_2, \dots, x_{N-1}) \in \mathcal{S}_N^{a,b}$ minimizing L_N over $\mathcal{S}_N^{a,b}$ when $\mathrm{supp}(\mu)$ contains at least N points.

Theorem 2.1 (Uniqueness of critical points of L_N) *Let $N \ge 3$ and $r \in [1, +\infty)$. Assume that $\mu([a, b]) = 1$ and μ is atomless. Then the function $L_N : \mathcal{S}_N^{a,b} \to \mathbb{R}_+$ defined just above is differentiable and its gradient ∇L_N admits a continuous extension on $\overline{\mathcal{S}}_N^{a,b}$. If μ admits a density f with respect to the Lebesgue measure which is positive and* log*-concave on the interval (a, b) and vanishes outside, then the continuous extension of ∇L_N has a unique zero $x^\star$ and $x^\star \in \mathcal{S}_N^{a,b}$.*

We deduce the following result.

Corollary 2.2 (Uniqueness of L^r-quantizers) *Let $N \ge 3$ and $r \in [1, +\infty)$. Assume that μ is atomless and $\mathrm{conv}\big(\mathrm{supp}(\mu)\big) = [a, b]$ with $-\infty < a < b < +\infty$. Then the L^r-optimal dual quantization grids at the level N are of the form $\{a, x_2, \dots, x_{N-1}, b\}$ with $(x_2, \dots, x_{N-1}) \in \mathcal{S}_N^{a,b}$ solving the master equation $\nabla L_N(x_2, \dots, x_{N-1}) = 0$. If, moreover, μ admits a density f with respect to the Lebesgue measure which is positive and* log*-concave on the interval (a, b) and vanishes outside, then the unique L^r-optimal dual quantization grid at level N of μ is $\{a, x_2^\star, \dots, x_{N-1}^\star, b\}$ where $x^\star$ is the unique critical point of L_N in $\mathcal{S}_N^{a,b}$.*

Notice that by concavity of $x \mapsto \log f(x)$ on (a,b), this function is continuous on (a,b) and admits limits in $\{-\infty\} \cup \mathbb{R}$ as $x \to a+$ and $x \to b-$ so that f is continuous on (a,b) and admits limits in $\mathbb{R}_+$ as $x \to a+$ and $x \to b-$. The density f is continuous on $\mathbb{R}$ iff both limits are equal to 0. In any case, f is bounded on $\mathbb{R}$.

To prove Theorem 2.1, we will rely on the following two classical results, tailored variants of the celebrated Mountain pass lemma and Gershgorin's lemma respectively.

Theorem 2.3 (Mountain pass Lemma) Compact case *(see [15]) Assume that $K \subset \mathbb{R}^n$ is the closure of a nonempty compact convex open set O (then $\overset{\circ}{K} = O$) and that $L : K \longrightarrow \mathbb{R}$ is C^1 on $\overset{\circ}{K}$, ∇L admits a continuous extension on K satisfying $\{\nabla L = 0\} \subset \overset{\circ}{K}$ and for every small enough $\varepsilon > 0$, $(Id - \varepsilon \nabla L)(\overset{\circ}{K}) \subset \overset{\circ}{K}$. If two distinct zeros of ∇L are (strict) local minima, then ∇L has a third zero which can in no case be a local minimum.*

Proposition 2.4 (À la Gershgorin Lemma) *(a) Let $A = [a_{ij}]$ be an $n \times n$ symmetric matrix with dominating diagonal i.e.*

$$\forall\, i \in \{1,\dots,n\},\ \forall\, j \in \{1,\dots,n\} \setminus \{i\},\ a_{ij} \le 0 \ \text{ and } \ \Lambda_i := \sum_{j=1}^{n} a_{ij} \ge 0.$$

Then all eigenvalues of A are non-negative.

(b) If moreover, A is tridiagonal with $a_{ii\pm1} < 0$, for $i = 2,\dots,n-1$, a_{12}, $a_{nn-1} < 0$, Λ_1 or $\Lambda_n > 0$, then all eigenvalues of A are positive.

Proof. Let $\lambda \in \mathbb{R}$ be an eigenvalue of the symmetric matrix A and $x = (x_1,\dots,x_n)$ one of its eigenvectors. Let $i_0 \in \mathrm{argmax}_{1\le i\le n}|x_i|$. We may assume without loss of generality that $x_{i_0} = 1$. Also note that $a_{ii} \ge 0$, $i = 1,\dots,n$.

(a) We have $|x_j| \le 1$ for all $j = 1,\dots,n$ so that

$$\lambda = a_{i_0 i_0} + \sum_{j \ne i_0} a_{i_0 j} x_j \ge a_{i_0 i_0} + \sum_{j \ne i_0} a_{i_0 j}|x_j| \ge \Lambda_{i_0} \ge 0.$$

(b) If $\lambda = 0$, then under the convention $a_{10} = a_{nn+1} = 0$ and $x_0 = x_{n+1} = 1$, we have

$$0 = \lambda = a_{i_0 i_0} + a_{i_0 i_0-1} x_{i_0-1} + a_{i_0 i_0+1} x_{i_0+1} \ge a_{i_0 i_0} + a_{i_0 i_0-1}|x_{i_0-1}| + a_{i_0 i_0+1}|x_{i_0+1}| \ge \Lambda_i = 0$$

so that $x_{i_0\pm1} = |x_{i_0\pm1}| = 1$. Then, by induction, we show that $x_i = 1$ for all $i = 1,\dots,n$ i.e. $x = \mathbf{1}$. But then, if $L_1 = a_{11} + a_{12} > 0$,

$$\lambda = \lambda x_1 = a_{11} x_1 + a_{12} x_2 = a_{11} + a_{12} > 0$$

which yields a contradiction. One concludes likewise if $\Lambda_n > 0$. Hence $\lambda > 0$. □

In order to prove Theorem 2.1 we will follow the strategy originally developed in [15] for Voronoi optimal quantizers.

Proof. In the proof, the hypotheses on μ will only be gradually reinforced to those made in the statement.

We first assume that $\mu([a,b])=1$. We know that, for every $\xi\in[x_{i-1},x_i]$ and every $u\in[0,1]$,

$$\mathrm{Proj}_x^{del}(\xi,u)=\sum_{i=2}^{N}\mathbf{1}_{\{0\le u<\frac{\xi-x_{i-1}}{\Delta x_i}\}}x_i+\mathbf{1}_{\{\frac{\xi-x_{i-1}}{\Delta x_i}\le u\le 1\}}x_{i-1},\qquad \Delta x_i=x_i-x_{i-1},\ i=2,\dots,N.$$

Now

$$\begin{aligned}L_N(x)&=\int_{x_{i-1}}^{x_i}\Big[|\xi-x_{i-1}|^r\frac{x_i-\xi}{\Delta x_i}+|x_i-\xi|^r\frac{\xi-x_{i-1}}{\Delta x_i}\Big]\mu(d\xi)\\&=\sum_{i=2}^{N}(\Delta x_i)^r\int_{x_{i-1}}^{x_i}\varpi_r\Big(\frac{\xi-x_{i-1}}{\Delta x_i}\Big)\mu(d\xi)\end{aligned}\tag{2.5}$$

where

$$\varpi_r(u)=u^r(1-u)+(1-u)^r u,\quad u\in[0,1],$$

(See Figure 2). Note that $\varpi_r\Big(\frac{x_i-x_{i-1}}{\Delta x_i}\Big)\mu(\{x_i\})=0=\varpi_r(\frac{x_{i-1}-x_{i-1}}{\Delta x_i})\mu(\{x_{i-1}\})$ so that the notation $\int_{x_{i-1}}^{x_i}\varpi_r\Big(\frac{\xi-x_{i-1}}{\Delta x_i}\Big)\mu(d\xi)$ makes sense even if μ weights points. The function ϖ_r when extended under the same notation by the value 0 outside the interval $[0,1]$ is continuous and bounded by $1/2$ on the real line. As a consequence, Lebesgue's theorem ensures that $(y,z)\mapsto\int_y^z\varpi_r(\frac{\xi-y}{z-y})\mu(d\xi)=\int_{\mathbb{R}}\varpi_r(\frac{\xi-y}{z-y})\mu(d\xi)$ is continuous on $\{(y,z)\in\mathbb{R}^2:\ y<z\}$. Moreover, for $y<z$, $0\le(z-y)^r\int_y^z\varpi_r(\frac{\xi-y}{z-y})\mu(d\xi)\le\frac{(z-y)^r}{2}$. We deduce that L_N is continuous on $\mathcal{S}_N^{a,b}$ and can be continuously extended to its closure $K=\overline{\mathcal{S}}_N^{a,b}$.

When μ has a density f, an elementary change of variable in each integral yields the alternative formulation

$$L_N(x)=\sum_{i=2}^{N}(\Delta x_i)^{r-1}\int_0^1\varpi_r(z)f\big(x_{i-1}+z\Delta x_i\big)dz.\tag{2.6}$$

Since the extended function ϖ_r is differentiable outside $\{0,1\}$ with a bounded derivative, Lebesgue's theorem ensures that $(y,z)\mapsto\int_{\mathbb{R}}\varpi_r(\frac{\xi-y}{z-y})\mu(d\xi)$ admits a partial derivative with respect to its first (resp. second) variable equal to $\int_{\mathbb{R}}\frac{\xi-z}{(z-y)^2}\varpi_r'(\frac{\xi-y}{z-y})\mu(d\xi)$ (resp. $\int_{\mathbb{R}}\frac{y-\xi}{(z-y)^2}\varpi_r'(\frac{\xi-y}{z-y})\mu(d\xi)$) at each point (y,z) such that $y<z$ and $\mu(\{y\})=0$ (resp. $\mu(\{z\})=0$). Hence, for $i=2:N-1$ and $x\in\mathcal{S}_N^{a,b}$ such that $\mu(\{x_i\})=0$, L_N admits a partial derivative with respect to x_i at x given by

$$\partial_{x_i}L_N(x)=(\Delta x_i)^{r-1}\int_{(x_{i-1},x_i]}\Psi_r\Big(\frac{\xi-x_{i-1}}{\Delta x_i}\Big)\mu(d\xi)-(\Delta x_{i+1})^{r-1}\int_{[x_i,x_{i+1})}\Psi_r\Big(\frac{x_{i+1}-\xi}{\Delta x_{i+1}}\Big)\mu(d\xi),\tag{2.7}$$

where the function Ψ_r is defined by

$$\forall u\in(0,1),\ \Psi_r(u)=r\varpi_r(u)-u\varpi_r'(u)=(r-1)u(1-u)^r+u^{r+1}+ru^2(1-u)^{r-1}>0\tag{2.8}$$

and by $\Psi_r(0)=0$ and $\Psi_r(1)=1+\mathbf{1}_{\{r=1\}}$.

By Fubini's theorem and since $\Psi_r(0) = 0$,

$$\int_{(x_{i-1},x_i]} \Psi_r\Big(\frac{\xi - x_{i-1}}{\Delta x_i}\Big)\mu(d\xi) = \int_{(x_{i-1},x_i]} \int_0^1 \mathbf{1}_{\{z<\frac{\xi-x_{i-1}}{\Delta x_i}\}} \Psi_r'(z)dz\mu(d\xi)$$
$$= \int_0^1 \Psi_r'(z)\big(F(x_i) - F(x_{i-1} + z\Delta x_i)\big)dz.$$

Hence, dealing in a similar way with the second term in the right-hand side of (2.7), we obtain the following second form of $\partial_{x_i} L_N(x)$:

$$\partial_{x_i} L_N(x) = (\Delta x_i)^{r-1} \int_0^1 \Psi_r'(z)\big(F(x_i) - F(x_{i-1} + z\Delta x_i)\big)dz$$
$$- (\Delta x_{i+1})^{r-1} \int_0^1 \Psi_r'(1-z)\big(F(x_i + z\Delta x_{i+1}) - F(x_i)\big)dz. \tag{2.9}$$

When μ is atomless, then L_N is differentiable on $\mathcal{S}_N^{a,b}$ and we easily deduce from the continuity of F and (2.9) that ∇L_N is continuous on $\mathcal{S}_N^{a,b}$ and admits a continuous extension on $\overline{\mathcal{S}}_N^{a,b}$.

When μ has a density f, then the partial derivative is also equal to

$$\partial_{x_i} L_N(x) = (\Delta x_i)^r \int_0^1 \Psi_r(z) f\big(x_{i-1} + z\Delta x_i\big)dz - (\Delta x_{i+1})^r \int_0^1 \Psi_r(1-z) f(x_i + z\Delta x_{i+1}\big)dz. \tag{2.10}$$

Note that each of these three forms of partial derivatives $\partial_{x_i} L_N(x)$ yields a version of the *master equation* for dual quantization at level N, $\nabla L_N(x) = 0$.

From now on, we assume that μ has a density f which is du a.e. positive on (a, b). If $x \in \overline{\mathcal{S}}_N^{a,b}$ solves the master equation then by (2.10), $(\Delta x_i)^r = 0 \Leftrightarrow (\Delta x_{i+1})^r = 0$ for $i = 2, \ldots, N-1$ and since $x_1 = a < b = x_N$, necessarily $\Delta x_i > 0$ for $i = 2, \ldots, N$ i.e. $x \in \mathcal{S}_N^{a,b} = \overset{\circ}{K}$.

When the density f is continuous on (a, b), the cumulative distribution function F is continuously differentiable on this interval and it follows from (2.9) that the Hessian of L_N does exist on $\mathcal{S}_N^{a,b}$ and has a symmetric tridiagonal structure. However in order to apply the refined Gershgorin Lemma (Lemma 2.4(b) with $n = N-2$), we need to show that the sub-diagonal terms are non-positive and the sum of its lines i.e. $\sum_{\ell=0,\pm1} \partial^2_{x_i x_{i+\ell}} L_N(x)$, $i = 2 : N-1$ (with the obvious convention that $\partial_{x_1}[\ldots] = \partial_{x_N}[\ldots] = 0$). We assume that the density f is also continuous and rely on (2.9) to derive that

$$\partial^2_{x_{i-1}x_i} L_N(x) = -(r-1)(\Delta x_i)^{r-2} \int_0^1 \Psi_r'(z)\big(F(x_i) - F(x_{i-1} + z\Delta x_i)\big)dz$$
$$- (\Delta x_i)^{r-1} \int_0^1 \Psi_r'(z)(1-z) f(x_{i-1} + z\Delta x_i)dz$$
$$= -(\Delta x_i)^{r-1} \int_0^1 \vartheta_r(z) f(x_{i-1} + z\Delta x_i)dz \tag{2.11}$$

where, for every $z \in (0, 1)$,

$$\vartheta_r(z) = (r-1)\Psi_r(z) + (1-z)\Psi_r'(z) = (r+1)\big(z(1-z)^r + z^r(1-z)\big) + (r-1)\big(z^{r+1} + (1-z)^{r+1}\big) > 0.$$

Hence, f being du-a.e. positive on (a,b), $\partial^2_{x_{i-1}x_i} L_N(x) < 0$.

Similar computations show, still assuming that f is continuous, that

$$\partial^2_{x_i^2} L_N(x) = \Psi_r(1) f(x_i)\big((\Delta x_i)^{r-1} + (\Delta x_{i+1})^{r-1}\big) + (\Delta x_i)^{r-1}\int_0^1 \widetilde{\vartheta}_r(z) f\big(x_{i-1} + z\Delta x_i\big)dz + (\Delta x_{i+1})^{r-1}\int_0^1 \widetilde{\vartheta}_r(1-z) f\big(x_i + z\Delta x_{i+1}\big)dz \tag{2.12}$$

($\Psi_r(1) = 1 + \mathbf{1}_{\{r=1\}}$) where

$$\widetilde{\vartheta}_r(z) = (r-1)\Psi_r(z) - z\Psi_r'(z).$$

Let us introduce for every $i = 2, \ldots, N-1$, the quantity

$$S_i = \Psi_r(1)\big((\Delta x_i)^{r-1} + (\Delta x_{i+1})^{r-1}\big) f(x_i) - (\Delta x_i)^{r-1}\int_0^1 \Psi_r'(z) f(x_{i-1} + z\Delta x_i)dz - (\Delta x_{i+1})^{r-1}\int_0^1 \Psi_r'(1-z) f(x_i + z\Delta x_{i+1})dz.$$

One derives from (2.11), (2.12) and the obvious fact $\widetilde{\vartheta}_r - \vartheta_r = -\Psi_r'$ that

$$S_i = \sum_{\ell=0,\pm 1} \partial^2_{x_i x_{i+\ell}} L_N(x) = \sum_{j=2}^{N-1} \partial^2_{x_i x_j} L_N(x) \quad \text{for} \quad i = 3 : N-2.$$

(We could have taken advantage of the anti-symmetries induced by the fact that $\partial_{x_i}\Delta x_i + \partial_{x_{i-1}}\Delta x_i = 0$ to compute S_i without computing $\partial^2_{x_i^2} L_N(x)$ but we will need a closed form of the diagonal term of the Hessian for the counterexample below).

Moreover, one checks that by positivity of ϑ_r on $(0,1)$ and of f on (a,b),

$$\sum_{j=2}^{N-1} \partial^2_{x_2 x_j} L_N(x) = \sum_{\ell=0,1} \partial^2_{x_2 x_{i+\ell}} L_N(x) = S_2 + (\Delta x_2)^{r-1}\int_0^1 \vartheta_r(z) f(x_1 + z\Delta x_2)dz > S_2 \text{ and}$$

$$\sum_{j=2}^{N-1} \partial^2_{x_{N-1} x_j} L_N(x) = \sum_{\ell=-1,0} \partial^2_{x_{N-1} x_{N-1+\ell}} L_N(x) = S_{N-1} + (\Delta x_N)^{r-1}\int_0^1 \vartheta_r(z) f(x_{N-1} + z\Delta x_N)dz > S_{N-1}.$$

Assume now that f is positive and right differentiable on (a,b) with right derivative f_r'. Then, by an integration by part, one shows that

$$\begin{aligned} S_i &= (\Delta x_i)^r \int_0^1 \Psi_r(z) f_r'(x_{i-1} + z\Delta x_i)dz - (\Delta x_{i+1})^r \int_0^1 \Psi_r(1-z) f_r'(x_i + z\Delta x_{i+1})dz \\ &= (\Delta x_i)^r \int_0^1 \Psi_r(z)\frac{f_r'(x_{i-1} + z\Delta x_i)}{f(x_{i-1} + z\Delta x_i)} f(x_{i-1} + z\Delta x_i)dz \\ &\quad - (\Delta x_{i+1})^r \int_0^1 \Psi_r(1-z)\frac{f_r'(x_i + z\Delta x_{i+1})}{f(x_i + z\Delta x_{i+1})} f(x_i + z\Delta x_{i+1})dz. \end{aligned}$$

Now if, furthermore, f is log-concave then f is right differentiable and $\frac{f_r'}{f}$ is non-increasing so that $\frac{f_r'}{f} \le \frac{f_r'}{f}(x_i)$ on (x_i, x_{i+1}) and $\frac{f_r'}{f} \ge \frac{f_r'}{f}(x_i)$ on (x_{i-1}, x_i). Since Ψ_r is positive on $(0,1)$, we deduce that

$$S_i \ge \frac{f_r'}{f}(x_i)\left((\Delta x_i)^r\int_0^1 \Psi_r(z) f(x_{i-1}+z\Delta x_i)dz - (\Delta x_{i+1})^r\int_0^1 \Psi_r(1-z) f(x_i+z\Delta x_{i+1})dz\right)$$

where the second factor in the right-hand side is equal to 0 when $x \in \mathcal{S}_N^{a,b}$ is solution to the master equation (derived from) (2.10). Consequently, $\sum_{j=2}^{N-1} \partial^2_{x_j x_i} L_N(x) = S_i \ge 0$ for $i = 3 : N-2$ and $\sum_{j=2}^{N-1} \partial^2_{x_j x_i} L_N(x) > S_i \ge 0$ for $i \in \{2, N-1\}$. It follows from Proposition 2.4 that the Hessian $\nabla^2 L_N(x)$ at an equilibrium point $x \in \mathcal{S}_N^{a,b}$ has a strictly positive spectrum and x is consequently a strict local minimum of L_N on $\mathcal{S}_N^{a,b}$. If we can prove that $(I_d - \varepsilon \nabla L)(\mathcal{S}_N^{a,b}) \subset \mathcal{S}_N^{a,b}$ for small enough ε, then we may apply (the variant of) the Mountain Pass Lemma (Theorem 2.3) to the convex compact $\overline{\mathcal{S}}_N^{a,b}$ of $\mathbb{R}^{N-2}$ with non empty interior to conclude that L_N admits at most one equilibrium point $x \in \mathcal{S}_N^{a,b}$. This is the purpose of the next lemma, the hypothesis of which is satisfied when the density f is positive and log-concave on (a,b) since, according to the remark just after Corollary 2.2, f is then bounded.

Lemma 2.5 *Let $r \in [1, +\infty)$. If the density f satisfies*

$$f \text{ bounded if } r = 1 \text{ or } \int_a^b f^{\frac{1}{r-1}}(\xi)d\xi < +\infty \text{ if } r \in (1,2),$$

then, for $\varepsilon > 0$ small enough, $(Id - \varepsilon \nabla L_N)(\mathcal{S}_N^{a,b}) \subset \mathcal{S}_N^{a,b}$.

Proof. Let $x \in \mathcal{S}_N^{a,b}$.

Assume $r \in (1,2)$. Then, for every $u, v \in [a,b]$, Hölder's inequality implies

$$\int_u^v f(\xi)d\xi \le \left(\int_a^b f^{\frac{1}{r-1}}(\xi)d\xi\right)^{r-1}(v-u)^{2-r}.$$

On the other hand, for $i = 2, \dots, N-2$, it follows from (2.10) that

$$\begin{aligned}
\partial_{x_{i+1}} L_N(x) - \partial_{x_i} L_N(x) &\le (\Delta x_{i+1})^r \int_0^1 \big(\Psi_r(z) + \Psi_r(1-z)\big) f(x_i + z\Delta x_{i+1})dz \\
&\le \Delta x_{i+1}\, C_r (\Delta x_{i+1})^{r-2} \int_{x_i}^{x_{i+1}} f(\xi)d\xi \\
&\le \Delta x_{i+1}\, C_r \left(\int_a^b f^{\frac{1}{r-1}}(\xi)d\xi\right)^{r-1}
\end{aligned}$$

where $C_r = \sup_{z\in[0,1]} \big(\Psi_r(z) + \Psi_r(1-z)\big) < +\infty$ according to (2.8). Consequently for $\varepsilon < \left(C_r \left(\int_a^b f^{\frac{1}{r-1}}(\xi)d\xi\right)^{r-1}\right)^{-1}$

$$x_i - \varepsilon \partial_{x_i} L_N(x) < x_{i+1} - \varepsilon \partial_{x_{i+1}} L_N(x), \quad i = 2 : N-1.$$

If $r = 1$, this inequality follows likewise by replacing $\left(\int_a^b f^{\frac{1}{r-1}}(\xi)d\xi\right)^{r-1}$ by $\|f\|_\infty$.

If $r \geq 2$, just note that $(\Delta x_{i+1})^{r-2}\int_{x_i}^{x_{i+1}} f(\xi)d\xi \leq (b-a)^{r-2}$ and choose $\varepsilon \leq \left(C_r(b-a)^{r-2}\right)^{-1}$. It remains to prove that $x_2 - \varepsilon\,\partial_{x_2}L(x) > a$ and $x_{N-1} - \varepsilon\,\partial_{x_{N-1}}L(x) < b$. In fact

$$\begin{aligned} x_2 - \varepsilon\,\partial_{x_2}L(x) &> x_2 - \varepsilon(\Delta x_2)^r \int_0^1 \Psi_r(z) f(a + z\Delta x_2)dz \\ &\geq x_2 - \Delta x_2 \varepsilon \|\Psi_r\|_{\sup}(\Delta x_2)^{r-2}\int_a^{x_2} f(\xi)d\xi. \end{aligned}$$

Inspecting the same cases as above, one shows under the assumptions made on f for $r \in [1,2]$, that for $\varepsilon \in (0, \varepsilon_r']$ small enough (independently of x),

$$x_2 - \varepsilon\,\partial_{x_2}L(x) > x_2 - (x_2 - a) = a.$$

The second inequality follows likewise. This completes the proof of the lemma. □

The following counterexample shows that uniqueness of critical points of L_N may fail when the density f is continuous, *du-a.e.* positive and (left and) right differentiable, but not log-concave.

Counter–example. The idea to devise this counter-example is to find a distribution μ with a periodic density on the interval $[0,1]$ that trivially makes $x^\star = \left(\frac{k-1}{N-1}\right)_{k=1,\dots,N}$ an equilibrium at level N but which assigns much mass in between the codewords $x_k^* = \frac{k-1}{N-1}$ so that this equilibrium cannot be a local minimum of L_N. As a consequence there will be at least one further equilibrium point: the minimum of L_N known to lie in $S_N^{0,1}$.

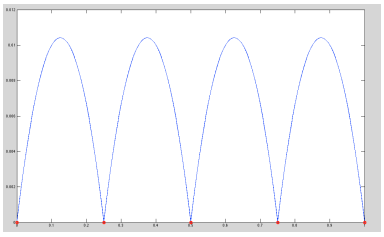

Figure 1 – *The probability density g for $r = 2$ and $N = 5$ (see (2.14)). Red bulllets are the 5 codewords x_k^*, $k = 1:5$.*

Let $r \geq 1$ and let $g \in \mathcal{C}([0,1], \mathbb{R}_+)$ a probability density function satisfying $g(z) = g(1-z)$ and $g(0) = 0$. We define for a fixed $N \geq 2$ the probability measure

$$\mu(du) := g\big(\{(N-1)u\}\big)du$$

where $\{\cdot\}$ denotes the fractional part function. This defines an absolutely continuous probability measure μ on the unit interval with an a.e. positive continuous density.

One checks that $x^\star := \left(\frac{k-1}{N-1}\right)_{k=1,\dots,N}$ satisfies $\nabla L_N(x^\star) = 0$ using the master equation derived from (2.10) and the obvious facts that $\Delta x_i = \frac{1}{N-1}$, $i = 2,\dots,N$, and $\int_0^1 \Psi_r(z)g(z)dz = \int_0^1 \Psi_r(1-z)g(z)dz$.

Elementary computations starting from (2.11) and (2.12) show that the Hessian $\nabla^2 L_{_N}(x^\star)$ is a symmetric tridiagonal $(N-2)\times(N-2)$ matrix of the form

$$\nabla^2 L_{_N}(x^\star) = (N-1)^{-(r-1)}A \quad \text{with} \quad A = \begin{bmatrix} 2a & b & 0 & \cdots & \cdots & 0 \\ b & 2a & b & 0 & \cdots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ldots & 0 \\ \vdots & \ddots & b & \ddots & 2a & b \\ 0 & \ldots & \ldots & 0 & b & 2a \end{bmatrix} \in \mathcal{S}(N-2,\mathbb{R}).$$

with

$$a = \int_0^1 \widetilde{\vartheta}_r(z)g(z)dz \quad \text{and} \quad b = -\int_0^1 \vartheta_r(z)g(z)dz < 0$$

(since both ϑ_r and g are positive on $(0,1)$). It is classical background that (real) eigenvalues of such a symmetric tridiagonal matrix A are

$$\lambda_k = 2\Big(a + b\cos\big(\tfrac{k\pi}{N-1}\big)\Big),\ k = 1,\ldots,N-2, \tag{2.13}$$

so that its lowest eigenvalue is $\lambda_{\min}(N) = 2\big(a + b\cos(\frac{\pi}{N-1})\big)$ (obtained with $k=1$).

Using again that $\vartheta_r - \widetilde{\vartheta}_r = -\Psi_r'$,

$$a + b = \int_0^1 (\widetilde{\vartheta}_r - \vartheta_r)(z)g(z)dz = -\int_0^1 \Psi_r'(z)g(z)dz.$$

Now, we note that $\Psi_r'(z) + \Psi_r'(1-z) = -\varpi_r''(z)$, $z\in[0,1]$, so that, taking advantage of the fact that $g(1-z) = g(z)$, we derive

$$\int_0^1 \Psi_r'(z)g(z)dz = \int_0^1 \frac{\Psi_r'(z)+\Psi_r'(1-z)}{2}g(z)dz = -\frac{1}{2}\int_0^1 \varpi_r''(z)g(z)dz = -\int_0^{1/2} \varpi_r''(z)g(z)dz$$

since both g and ϖ_r'' are symmetric (w.r.t. $1/2$) on $[0,1]$. If we assume that g is also differentiable on $(0,1)$, then an integration by part yields

$$a+b = \int_0^{1/2} \varpi_r''(z)g(z)dz = g(1/2)\varpi_r'(1/2) - g(0)\varpi'(0) - \int_0^{1/2} \varpi_r'(z)g'(z)dz = -\int_0^{1/2} \varpi_r'(z)g'(z)dz$$

since $g(0) = 0$ and, by symmetry w.r.t $1/2$, $\varpi_r'(1/2) = 0$. Finally setting

$$g = c_r\varpi_r \tag{2.14}$$

so that g is a probability density, one has

$$a + b = -c_r\int_0^{1/2} (\varpi_r'(z))^2 dz < 0,$$

hence $\frac{a+b}{b} > 0$ since $b<0$. Consequently, $\lambda_{\min}(N) < 0$ for any N large enough such that

$$\cos\big(\tfrac{\pi}{N-1}\big) > 1 - \frac{a+b}{b}.$$

Then, as some of the eigenvalues of the Hessian of L_N at $x^\star$ are negative, this point cannot be a local minima of L_N. The function L_N also has a local minima lying in $\mathcal{S}_N^{a,b}$ since $\mathrm{supp}(\mu) = [0,1]$, namely any L^r-optimal dual N-quantizer. Hence, uniqueness of the solution to the master equation fails.

However, note that this only stands as a counter-example to uniqueness of solutions of the master equation: the set of local minima may still be reduced to a single N-tuple.

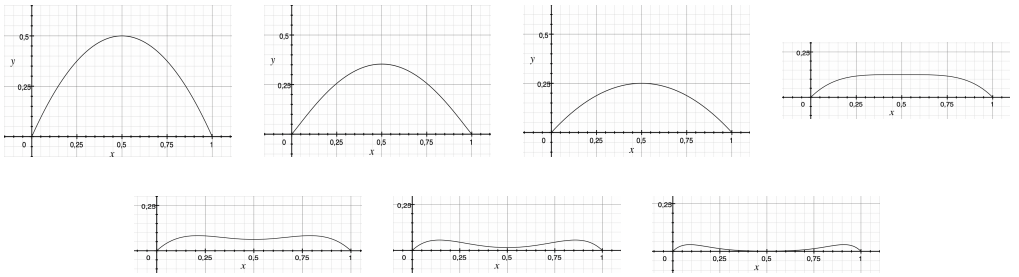

Figure 2 – *Functions ϖ_r for (from left to right)* $r = 1, 1.5, 2, 3, 4, 6, 10$.

Remarks • The conclusion of the above counterexample holds true for any symmetric probability density g on $[0,1]$ such that $g(0) = 0$ and

$$\int_0^{1/2} g'(z)\varpi_r'(z)dz > 0.$$

• If $\mu = U([a,b])$ then the master equation e.g. derived from (2.9) reads $(\Delta x_i)^r = (\Delta x_{i+1})^r$, $i = 2 : N-1$ since $\int_0^1 \Psi_r(z)dz = \int_0^1 \Psi_r(1-z)dz$. Hence $\Delta x_i = \frac{b-a}{N-1}$, $i = 2 : N$, so that one retrieves the fact that the unique L^r-optimal dual N-quantizer of $U([0,1])$ is always $x^{\star,N} = \big(\frac{i-1}{N-1}\big)_{i=1:N}$ for all $r \geq 1$.

• If r is an integer, then $\Psi_r(z)$ is a polynomial function with degree at most $r+1$. To be more precise one checks that its term of degree r is always 0 and that the coefficient $(-1)^r\big((-1)^r - 1\big)z^{r+1}$ of its term of degree $r+1$ is 0 if and only if r is even. Hence $d^0\,\Psi_r = r-1$ if r is an even integer and $r+1$ if r is an odd integer.

3 A Lloyd like algorithm for dual quantization in the quadratic case

A fixed point formulation of the master equation. In this section, we specialize to the quadratic $r = 2$ case and take advantage of this specialization to derive a more convenient expression

of the distortion of the dual grid $\Gamma = \{a, x_2, \dots, x_{N-1}, b\}$ parametrized by $(x_2, \dots, x_{N-1}) \in \mathcal{S}_N^{a,b}$:

$$\begin{aligned}
L_N(x) &= \int_{[a,b]} \mu(d\xi) \int_0^1 du |\xi - \mathrm{Proj}_\Gamma^{del}(\xi, u)|^2 = \int_{[a,b]} \mu(d\xi) \int_0^1 du \big(\mathrm{Proj}_\Gamma^{del}(\xi, u)\big)^2 - \int_{\mathbb{R}} \xi^2 \mu(d\xi) \\
&= \sum_{i=1}^{N-1} \int_{(x_i, x_{i+1}]} \mu(d\xi) \Big[\tfrac{x_{i+1}-\xi}{x_{i+1}-x_i} x_i^2 + \tfrac{\xi - x_i}{x_{i+1}-x_i} x_{i+1}^2 \Big] - \int_{\mathbb{R}} \xi^2 \mu(d\xi) \\
&= \sum_{i=1}^{N-1} \int_{(x_i, x_{i+1}]} \big((x_{i+1} + x_i)\xi - x_{i+1} x_i \big) \mu(d\xi) - \int_{\mathbb{R}} \xi^2 \mu(d\xi) \qquad (3.15) \\
&= \sum_{i=1}^{N-1} \big((x_i + x_{i+1})[K]_{x_i}^{x_{i+1}} - x_i x_{i+1} [F]_{x_i}^{x_{i+1}} \big) - \int_{\mathbb{R}} \xi^2 \mu(d\xi), \qquad (3.16)
\end{aligned}$$

where, we recall that $F(x) = \mu((-\infty, x])$ and $K(x) = \int_{(-\infty,x]} \xi \mu(d\xi)$ respectively denote the cumulative distribution function and the first partial moment of μ and for a function $g : \mathbb{R} \to \mathbb{R}$ and two real numbers $x \le y$, $[g]_x^y = g(y) - g(x)$. Then, it follows from (3.15) that the mapping $x \mapsto L_N(x)$ is continuously differentiable at x when the distribution μ is atomless (i.e. F and K are continuous) with

$$\partial_{x_i} L_N(x) = [K]_{x_{i-1}}^{x_{i+1}} - \big(x_{i+1} [F]_{x_i}^{x_{i+1}} + x_{i-1} [F]_{x_{i-1}}^{x_i} \big), \quad i = 2 : N-1, \qquad (3.17)$$

The master equation for optimal quadratic quantizers reads $x_1 = a$, $x_N = b$ and

$$\nabla_{x_{2:N-1}} L_N(x) = 0 \qquad (3.18)$$

that is

$$[K]_{x_{i-1}}^{x_{i+1}} = x_{i+1} [F]_{x_i}^{x_{i+1}} + x_{i-1} [F]_{x_{i-1}}^{x_i}, \quad i = 2 : N-1. \qquad (3.19)$$

Using Fubini's theorem for the second equality, we obtain that

$$\begin{aligned}
[K]_{x_{i-1}}^{x_{i+1}} &= \int_{(x_{i-1}, x_{i+1}]} \int_{(x_{i-1}, x_{i+1}]} \mathbf{1}_{\{y < \xi\}} dy \, \mu(d\xi) + x_{i-1} [F]_{x_{i-1}}^{x_{i+1}} \\
&= \int_{(x_{i-1}, x_{i+1}]} \big(F(x_{i+1}) - F(y) \big) dy + x_{i-1} [F]_{x_{i-1}}^{x_{i+1}} \\
&= - \int_{x_{i-1}}^{x_{i+1}} F(y) dy + x_{i+1} F(x_{i+1}) - x_{i-1} F(x_{i-1}).
\end{aligned}$$

We deduce the following more synthetic form for the master equation

$$(x_{i+1} - x_{i-1}) F(x_i) = \int_{x_{i-1}}^{x_{i+1}} F(\xi) d\xi, \quad i = 2 : N-1$$

which may also be deduced from the case $r = 2$ in (2.9) using $\Psi_2' = 1$ and performing a change of variables in each integral. Equivalently, we have

$$F(x_i) = \frac{\int_{x_{i-1}}^{x_{i+1}} F(\xi) d\xi}{x_{i+1} - x_{i-1}}, \quad i = 2 : N-1. \qquad (3.20)$$

Assume from now on that the distribution μ is atomless with support $[a,b]$. Then $F:[a,b]\to[0,1]$ is an increasing homeomorphism and we may define its inverse F^{-1} so that the above equation (3.20) can also be written as the fixed point equation

$$(x_2,\ldots,x_{N-1}) = T(x) = \big(T_2(x),\ldots,T_{N-1}(x)\big)$$

where, for every $x\in\mathcal{S}_N^{a,b}$,

$$T(x) = \left(F^{-1}\left(\frac{\int_{x_{i-1}}^{x_{i+1}}F(\xi)d\xi}{x_{i+1}-x_{i-1}}\right)\right)_{i=2:N-1}. \tag{3.21}$$

Since F is continuous and increasing from $F(a)=0$ to $F(b)=1$, $T(x)\in\mathcal{S}_N^{a,b}$ for each $x\in\mathcal{S}_N^{a,b}$. By Corollary 2.2, any quadratic optimal dual quantization grid at level N is of the form $\{a,x_2,\ldots,x_{N-1},b\}$ with $(x_2,\ldots,x_{N-1})\in\mathcal{S}_N^{a,b}$ solution to the master equation $\nabla L_N(x)=0$ or equivalently fixed point of T. If μ admits a density f with respect to the Lebesgue measure which is positive and log-concave on the interval (a,b) and vanishes outside, then there is a unique such point $(x_2^\star,\ldots,x_{N-1}^\star)$.

One checks that T *can be continuously extended to the closure* $\overline{\mathcal{S}}_N^{a,b}$ *of* $\mathcal{S}_N^{a,b}$. Indeed $x_{i-1}=x_i<x_{i+1}$ or $x_{i-1}<x_i=x_{i+1}$, the extension of $T_i(x)$ is straightforward and if $x_{i-1}=x_i=x_{i+1}$, set $T_i(x)=x_i$ (in both cases with $x_1=a$ and $x_N=b$).

From such a fixed point identity, one can devise an iterative fixed point procedure which can be seen as the counterpart of so-called Lloyd's method I procedure for dual quantization:

$$x^{[\ell+1]} = T\big(x^{[\ell]}\big),\ \ell\geq 0,\quad x^{[0]}\in\overline{\mathcal{S}}_N^{a,b}. \tag{3.22}$$

When μ admits a density f positive and log-concave on (a,b) and vanishing outside, while proving that this procedure converges at a geometric rate to $x^\star$, we are going to check that T admits a unique fixed point in $\mathcal{S}_N^{a,b}$ thus providing an alternative argument for the uniqueness statement in Theorem 2.1.

Convergence of the dual Lloyd algorithm. First note that if F is continuously differentiable on (a,b) i.e. μ has a continuous density f on (a,b), then the mapping T is itself continuously differentiable at any $x\in\mathcal{S}_N^{a,b}$ with a Jacobian matrix $J_T(x)=\big[\frac{\partial T_i}{\partial x_j}(x)\big]_{2\leq i,j\leq N-1}$ where $\frac{\partial T_i}{\partial x_j}(x)=0$ if $|i-j|\geq 2$ and

$$\frac{\partial T_i}{\partial x_{i-1}}(x) = \frac{1}{f\circ F^{-1}\Big(\frac{\int_{x_{i-1}}^{x_{i+1}}F(\xi)d\xi}{x_{i+1}-x_{i-1}}\Big)}\times\frac{\int_{x_{i-1}}^{x_{i+1}}(F(\xi)-F(x_{i-1}))d\xi}{(x_{i+1}-x_{i-1})^2}>0,\quad i=3,\ldots,N-1, \tag{3.23}$$

and

$$\frac{\partial T_i}{\partial x_{i+1}}(x) = \frac{1}{f\circ F^{-1}\Big(\frac{\int_{x_{i-1}}^{x_{i+1}}F(\xi)d\xi}{x_{i+1}-x_{i-1}}\Big)}\times\frac{\int_{x_{i-1}}^{x_{i+1}}(F(x_{i+1})-F(\xi))d\xi}{(x_{i+1}-x_{i-1})^2}>0,\quad i=2,\ldots,N-2. \tag{3.24}$$

The main result of this section is the following.

Theorem 3.1 *Let $-\infty < a < b < +\infty$. Assume that μ admits a density f with respect to the Lebesgue measure which is positive and* log*-concave on the interval (a,b) and vanishes outside. Then T admits a unique fixed point $x^\star$ in $\overline{\mathcal{S}}_N^{a,b}$ and $x^\star \in \mathcal{S}_N^{a,b}$. Moreover, $\{a, x_2^\star, \ldots, x_{N-1}^\star, b\}$ is the unique quadratic optimal dual quantization grid at level N of μ and*

$$\exists \rho \in [0,1),\ \forall x^{[0]} \in \overline{\mathcal{S}}_N^{a,b}, \forall \ell \in \mathbb{N},\ |x^{[\ell]} - x^\star|_{\ell^\infty} \leq \begin{cases} |x^{[0]} - x^\star|_{\ell^\infty} \rho^{\lfloor \ell / \widetilde{N} \rfloor} \text{ where } \widetilde{N} = \lceil \frac{N}{2} \rceil - 1 \\ |x^{[0]} - x^\star|_{\ell^\infty} \rho^{\ell} \text{ if } f \text{ is strictly log-concave on } (a,b). \end{cases}$$

As a preamble, we first establish an equivalence between two characterizations of *strong unimodality*, one being the log-concavity of the density of the distribution, whereas the other (see below) will be extensively used in what follows. Then we establish a general result about fixed point of locally contracting transforms from a compact set in itself.

Lemma 3.2 *Let $-\infty < a < b < +\infty$ and let $f : (a,b) \to (0,+\infty)$ be a positive probability density on (a,b) with cumulative distribution function $[a,b] \ni x \mapsto F(x) = \int_a^x f(\xi)d\xi$ and quantile function F^{-1}. The function f is* log*-concave (resp. strictly* log*-concave) on (a,b) iff $f \circ F^{-1}$ is concave (resp. strictly concave) on $(0,1)$.*

Proof. The cumulative distribution function F being continuous and increasing on $[a,b]$ with $F(a) = 0$ and $F(b) = 1$, it admits a continuous and increasing inverse $F^{-1} : [0,1] \to [a,b]$.

Let us suppose that f is log-concave (resp. strictly log-concave). Then $\log f$ is continuous and admits a non-increasing (resp. decreasing) right-hand derivative $(\log f)'_r$ as a real-valued concave (resp. strictly concave) function. By composition with the exponential, $f = \exp \circ \log f$ also admits a right-hand derivative equal to $f \times (\log f)'_r$. Since f is continuous and positive, the function F and its inverse F^{-1} are continuously differentiable with respective derivatives f and $\frac{1}{f \circ F^{-1}}$. We conclude that $f \circ F^{-1}$ admits a right-hand derivative equal to $\frac{f \times (\log f)'_r}{f} \circ F^{-1} = (\log f)'_r \circ F^{-1}$ which is non-increasing (resp. decreasing) as the composition of the non-increasing (resp. decreasing) function $(\log f)'_r$ with the increasing function F^{-1}. Therefore $f \circ F^{-1}$ is concave (resp. strictly concave).

When $f \circ F^{-1}$ is concave (resp. strictly concave), then this function is continuous and, by composition with the continuous function F, f is continuous so that F and F^{-1} are continuously differentiable with respective derivatives f and $\frac{1}{f \circ F^{-1}}$. Moreover $f = (f \circ F^{-1}) \circ F$ admits a right-hand derivative equal to $f'_r = (f \circ F^{-1})'_r \circ F \times f$. Then $\log f$ admits a right-hand derivative equal to $\frac{f'_r}{f} = (f \circ F^{-1})'_r \circ F$ which is non-increasing (resp. decreasing) as the composition of the non-increasing (resp. decreasing) function $(f \circ F^{-1})'_r$ with the increasing function F. Therefore $\log f$ is concave (resp. strictly concave). □

Proposition 3.3 *Let K be a convex compact subset of $\mathbb{R}^d$, equipped with a norm $\|\cdot\|$, and let $T : K \to K$ be a $\|\cdot\| - 1$-Lipschitz continuous mapping such that for some $k \in \mathbb{N}^*$, the mapping T^k obtained by iterating T k-times satisfies*

$$\rho_k := \sup_{y \in K,\, y \neq y^\star} \frac{\|T^k(y) - y^\star\|}{\|y - y^\star\|} < 1$$

for some fixed point $y^\star$ of T (the set of fixed points is non empty by Brouwer's theorem). Then $y^\star$ is the unique fixed point of T and for every $y_0 \in K$, the sequence recursively defined for $n \in \mathbb{N}$ by $y_{n+1} = T(y_n)$ geometrically converges to $y^\star$:

$$\forall n \in \mathbb{N},\ \|y_n - y^\star\| \leq \|y_0 - y^\star\| \rho_k^{\lfloor n/k \rfloor}.$$

Proof. The inequality $\|T^k(y) - y^\star\| \le \rho_k \|y - y^\star\|$ valid for each $y \in K$ with $\rho_k < 1$ implies that every fixed point of T^k and therefore of T is equal to $y^\star$. Moreover, using that $y^\star$ is a fixed point of T then the 1-Lipschitz property of T and last the previous inequality, we obtain that

$$\begin{aligned}\|y_n - y^\star\| &= \|T^{n-\lfloor n/k\rfloor k}(y_{\lfloor n/k\rfloor k}) - T^{n-\lfloor n/k\rfloor k}(y^\star)\| \le \|T^k(y_{(\lfloor n/k\rfloor-1)k}) - y^\star\| \\ &\le \rho_k \|T^k(y_{(\lfloor n/k\rfloor-2)k}) - y^\star\| \le \rho_k^{\lfloor n/k\rfloor}\|y_0 - y^\star\|.\end{aligned}$$

□

Proof of Theorem 3.1. Since $K = \overline{\mathcal{S}}_N^{a,b}$ is a convex and compact subset of $\mathbb{R}^{N-2}$, by Brouwer's fixed-point theorem, the set of fixed points of the continuous map $T : \overline{\mathcal{S}}_N^{a,b} \to \overline{\mathcal{S}}_N^{a,b}$ is non empty. Let $x \in \overline{\mathcal{S}}_N^{a,b}$ and $x_1 = a$, $x_N = b$. If $T_i(x) = x_{i-1}$ or $T_i(x) = x_{i+1}$ for some $i = 2, \ldots, N-1$, then, since F is increasing and continuous on $[a, b]$, $x_{i-1} = x_i = x_{i+1}$. If moreover $T(x) = x$, then one deduces that $T_{i+1}(x) = x_i$ if $i \le N-2$ and $T_{i-1}(x) = x_i$ if $i \ge 3$, so that by induction $x_1 = x_2 = \ldots = x_{N-1} = x_N$ which contradicts $x_1 = a < b = x_N$. Hence the non-empty set of fixed points of T is included in $\mathcal{S}_N^{a,b}$. Let $x^\star \in \mathcal{S}_N^{a,b}$ be one of these fixed points. We are going to check that the assumptions of Proposition 3.3 are satisfied with $y^\star = x^\star$ and with $k = 1$ in the strictly log-concave case and $k = \widetilde{N}$ is the log-concave case. The conclusions but the link between $x^\star$ and the unique quadratic optimal dual quantization grid at level N of μ then follow from this proposition. This link is a consequence of Corollary 2.2 and the fact that $x \in \mathcal{S}_N^{a,b}$ is a critical point of L_N iff it is a fixed point of T.

– *Strictly* log-*concave setting.* It follows from (3.23) and (3.24) that, for every $x \in \mathcal{S}_N^{a,b}$ and $i = 3, \ldots, N-2$,

$$\frac{\partial T_i}{\partial x_{i-1}}(x) + \frac{\partial T_i}{\partial x_{i+1}}(x) = \frac{1}{f \circ F^{-1}\Big(\frac{\int_{x_{i-1}}^{x_{i+1}} F(\xi)d\xi}{x_{i+1}-x_{i-1}}\Big)} \times \frac{F(x_{i+1}) - F(x_{i-1})}{x_{i+1} - x_{i-1}}.$$

As f is strictly log-concave, $f \circ F^{-1}$ is strictly concave by Lemma 3.2, hence it follows from Jensen's inequality that, under the convention $x_1 = a$ and $x_N = b$, for every $i = 2, \ldots, N-1$

$$f \circ F^{-1}\left(\frac{\int_{x_{i-1}}^{x_{i+1}} F(\xi)d\xi}{x_{i+1} - x_{i-1}}\right) > \frac{\int_{x_{i-1}}^{x_{i+1}} f(\xi)d\xi}{x_{i+1} - x_{i-1}} = \frac{F(x_{i+1}) - F(x_{i-1})}{x_{i+1} - x_{i-1}} \tag{3.25}$$

with a strict inequality since the probability measure $\mathbf{1}_{[x_{i-1}, x_{i+1}]}(\xi)d\xi$ is not a Dirac mass and F is not constant over $[x_{i-1}, x_{i+1}]$. As a consequence, for every $x \in \mathcal{S}_N^{a,b}$,

$$\forall i = 3, \ldots, N-2,\ \frac{\partial T_i}{\partial x_{i-1}}(x) + \frac{\partial T_i}{\partial x_{i+1}}(x) < 1, \tag{3.26}$$

$$\frac{\partial T_2}{\partial x_3}(x) = \frac{\int_a^{x_3}(F(x_3) - F(\xi))d\xi}{(x_3 - a)^2 f \circ F^{-1}\Big(\frac{\int_a^{x_3} F(\xi)d\xi}{x_3 - a}\Big)} < \frac{1}{f \circ F^{-1}\Big(\frac{\int_a^{x_3} F(\xi)d\xi}{x_3 - a}\Big)} \times \frac{F(x_3) - F(a)}{x_3 - a} < 1, \tag{3.27}$$

$$\frac{\partial T_{N-1}}{\partial x_{N-2}}(x) = \frac{\int_{x_{N-2}}^{b}(F(\xi) - F(x_{N-2}))d\xi}{(b - x_{N-2})^2 f \circ F^{-1}\Big(\frac{\int_{x_{N-2}}^{b} F(\xi)d\xi}{b - x_{N-2}}\Big)} < \frac{1}{f \circ F^{-1}\Big(\frac{\int_{x_{N-2}}^{b} F(\xi)d\xi}{b - x_{N-2}}\Big)} \times \frac{F(b) - F(x_{N-2})}{b - x_{N-2}} < 1. \tag{3.28}$$

Since $\frac{\partial T_i}{\partial x_j}(x) \geq 0$ with equality if $|i-j| \geq 2$, it is easy to deduce that, if $\mathbb{R}^{N-2}$ is equipped with the ℓ^∞-norm $|\cdot|_{\ell^\infty}$,

$$\forall\, x \in \mathcal{S}_N^{a,b}, \quad |||J_T(x)|||_{\ell^\infty} < 1 \tag{3.29}$$

where $|||\cdot|||_{\ell^\infty}$ denotes the operator norm with respect to the ℓ^∞-norm.

Now let $x \in \overline{\mathcal{S}}_N^{a,b}$ and $y \in \mathcal{S}_N^{a,b}$ with $y \neq x$. Then for each $t \in [0,1)$, $tx + (1-t)y \in \mathcal{S}_N^{a,b}$ and $t \mapsto T(tx+(1-t)y)$ is continuous on $[0,1]$ and differentiable on $[0,1)$ so that, with the integrability consequence of (3.29),

$$T(x) - T(y) = \int_0^1 J_T\big(tx + (1-t)y\big)(x-y)dt, \ \forall (x,y) \in \overline{\mathcal{S}}_N^{a,b} \times \mathcal{S}_N^{a,b}. \tag{3.30}$$

By the triangle inequality for integrals, the definition of the $|||\cdot|||_{\ell^\infty}$-norm and (3.29), one deduces that

$$\big|T(x) - T(y)\big|_{\ell^\infty} \leq \int_0^1 \big|\big|\big|J_T\big(tx + (1-t)y\big)\big|\big|\big|_{\ell^\infty} dt\, |x-y|_{\ell^\infty} < |x-y|_{\ell^\infty}. \tag{3.31}$$

Approximating $y \in \overline{\mathcal{S}}_N^{a,b}$ by a sequence of elements in $\mathcal{S}_N^{a,b}$, we deduce that T is $|.|_{\ell^\infty}$ 1-Lipschitz continuous. Moreover, for the choice y equal to the fixed point $x^\star$ of T, (3.30) writes

$$\forall x \in \overline{\mathcal{S}}_N^{a,b}, \quad T\big(x\big) - x^\star = A_x^{[1]}\big(x - x^\star\big) \quad \text{with} \quad A_x^{[1]} := \int_0^1 J_T\big(x^\star + t(x-x^\star)\big)dt. \tag{3.32}$$

Since $x \mapsto A_x^{[1]}$ is continuous on the compact set $\overline{\mathcal{S}}_N^{a,b}$ and

$$\forall x \in \overline{\mathcal{S}}_N^{a,b}, \ \Big|\Big|\Big|A_x^{[1]}\Big|\Big|\Big|_{\ell^\infty} \leq \int_0^1 \big|\big|\big|J_T\big(tx + (1-t)x^\star\big)\big|\big|\big|_{\ell^\infty} dt < 1,$$

we have $\sup_{x \in \overline{\mathcal{S}}_N^{a,b}} \Big|\Big|\Big|A_x^{[1]}\Big|\Big|\Big|_{\ell^\infty} < 1$. With (3.32), we deduce that the hypotheses of Proposition 3.3 are satisfied with $k=1$ and $\rho_1 = \sup_{x \in \overline{\mathcal{S}}_N^{a,b}} \Big|\Big|\Big|A_x^{[1]}\Big|\Big|\Big|_{\ell^\infty}$.

– log-*concave setting.*

The main difference is that the inequality in (3.25) is no longer strict. As a consequence, in (3.26)-(3.28), < 1 should now be replaced by ≤ 1 so that $|||J_T(x)|||_{\ell^\infty} \leq 1$. This still ensures that T is $|\cdot|_{\ell^\infty}$-1-Lipschitz continuous on $\overline{\mathcal{S}}_N^{a,b}$ and (3.32) still holds. To overcome the lack of strict contraction of $\mathbb{R}^{N-2} \ni u \mapsto J_T(x)u$, we are going to take advantage of the inequalities $\frac{\partial T_2}{\partial x_3} < 1$ and $\frac{\partial T_{N-1}}{\partial x_{N-2}} < 1$ still valid on $\mathcal{S}_N^{a,b}$ since the first inequalities in (3.27)-(3.28) remain strict.

For $k \geq 1$ we can iterate (3.32) to obtain

$$\forall x \in \overline{\mathcal{S}}_N^{a,b}, \ T^k\big(x\big) - x^\star = A_x^{[k]} A_x^{[k-1]} \dots A_x^{[1]}\big(x - x^\star\big) \tag{3.33}$$

where, for $\ell \geq 1$, $A_x^{[\ell]} = \int_0^1 J_T\big(x^\star + t(T^{l-1}(x) - x^\star)\big)dt$ is a tridiagonal matrix $\big[a_{ij}^{[\ell]}\big]_{2 \leq i,j \leq N-1}$ satisfying

$$\begin{aligned} &a_{ii}^{[\ell]} = 0, \ \ i = 2, \dots, N-1, \qquad a_{ii\pm1}^{[\ell]} > 0,\ a_{ii-1}^{[\ell]} + a_{ii+1}^{[\ell]} \leq 1, \ \ i = 3, \dots, N-2, \\ \text{and} \quad &0 < a_{23}^{[\ell]},\ a_{N-1N-2}^{[\ell]} < 1. \end{aligned} \tag{3.34}$$

Let $u = (u_2, \dots, u_{N-2}) \in \mathbb{R}^{N-2}$ with $|u|_{\ell^\infty} > 0$. Let $I_1 = \big\{i : |A_x^{[1]}u| = |u|_{\ell^\infty}\big\}$. It is clear that $2, N-1 \notin I_1$ since $\big|(A_x^{[1]}u)_2\big| = a_{23}^{[1]}|u_3| < |u_3| \le |u|_{\ell^\infty}$ (idem for the other term). Now let $I_2 = \big\{i : \big|(A_x^{[2]}A_x^{[1]}u)_i\big| = |u|_{\ell^\infty}\big\}$. It is still clear that $2, N-1 \notin I_2$. Now

$$\big(A_x^{[2]}A_x^{[1]}u\big)_3 = a_{32}^{[2]}\big(A_x^{[1]}u\big)_2 + a_{34}^{[2]}\big(A_x^{[1]}u\big)_4.$$

As $a_{32}^{[2]}$, $a_{34}^{[2]} > 0$ and $a_{32}^{[2]} + a_{34}^{[2]} \le 1$ and $2 \notin I_1$, $|(A_x^{[2]}A_x^{[1]}u)_3| < |u|_{\ell^\infty}$. One shows likewise that $N-2 \notin I_2$. Then one shows the same way round by induction that, $2, \dots, k+1$, $N-(k+1), \dots, N-1 \notin I_k = \big\{i : \big|(A_x^{[k]} \cdots A_x^{[1]}u)_i\big| = |u|_{\ell^\infty}\big\}$ which implies that $I_{\lceil \frac{N}{2} \rceil - 1} = \varnothing$ i.e. $\big|A_x^{[\lceil \frac{N}{2} \rceil - 1]} \cdots A_x^{[1]}u\big|_{\ell^\infty} < |u|_{\ell^\infty}$. Consequently setting $\widetilde{N} = \lceil \frac{N}{2} \rceil - 1$, we have

$$\forall x \in \overline{\mathcal{S}}_N^{a,b},\ \Big|\!\Big|\!\Big| A_x^{[\widetilde{N}]} \cdots A_x^{[1]} \Big|\!\Big|\!\Big|_{\ell^\infty} < 1.$$

Note that a more quantitative bound in terms of the coefficients of the matrices is derived in Proposition 3.4 below. With the continuity of $x \mapsto A_x^{[\widetilde{N}]} \cdots A_x^{[1]}$ over the compact set $\overline{\mathcal{S}}_N^{a,b}$ and (3.33), we deduce that the hypotheses of Proposition 3.3 are satisfied with $k = \widetilde{N}$ and $\rho_{\widetilde{N}} = \sup_{x \in \overline{\mathcal{S}}_N^{a,b}} \Big|\!\Big|\!\Big| A_x^{[\widetilde{N}]} \cdots A_x^{[1]} \Big|\!\Big|\!\Big|_{\ell^\infty} < 1$. □

To evaluate in sharper way $\rho_{\widetilde{N}}$, one might rely on the following Proposition which provides a quantitative bound for the $|||\cdot|||_{\ell^\infty}$-norm of a product of our tridiagonal matrices of interest.

Proposition 3.4 (Quantitative bound) *Let $N \ge 3$ and set $\tilde{N} = \lceil \frac{N}{2} \rceil - 1$. Let $A^{[\ell]} = [a_{ij}^{[\ell]}]_{2 \le i,j \le N-1}$, $l \in \{1, \dots, \tilde{N}\}$, be tridiagonal matrices whose entries satisfy the above conditions* (3.34)*. Then*

$$\begin{aligned}\Big|\!\Big|\!\Big| A^{[\tilde{N}]} \cdots A^{[1]} \Big|\!\Big|\!\Big|_{\ell^\infty} \le & \max_{2 \le i \le \tilde{N}} \Big(1 - a_{ii-1}^{[\tilde{N}]} a_{i-1i-2}^{[\tilde{N}-1]} \cdots a_{32}^{[\tilde{N}+3-i]}(1 - a_{23}^{[\tilde{N}+2-i]})\Big) \\ & \vee \max_{\lfloor \frac{N}{2} \rfloor \le i \le N-1} \Big(1 - a_{ii+1}^{[\tilde{N}]} a_{ii+2}^{[\tilde{N}-1]} \cdots a_{N-2N-1}^{[\tilde{N}+i+2-N]}(1 - a_{N-1N-2}^{[\tilde{N}+1+i-N]})\Big) < 1.\end{aligned}$$

Proof. For a matrix $B = [b_{ij}]_{1 \le i \le n,\, 1 \le j \le d} \in \mathbb{R}_+^{n \times d}$ and for $y, z \in \mathbb{R}^d$ such that $|y_i| \le z_i$, $i = 1 : d$, we have

$$\|By\|_{\ell^\infty} = \max_{1 \le i \le n} \Big| \sum_{j=1}^d b_{ij} y_j \Big| \le \max_{1 \le i \le n} \sum_{j=1}^d b_{ij} z_j = \|Bz\|_{\ell^\infty}.$$

Therefore denoting by $\mathbf{1}$ the vector in $\mathbb{R}^{N-2}$ with all coordinates equal to 1, we have $\Big|\!\Big|\!\Big| A^{[\tilde{N}]} \cdots A^{[1]} \Big|\!\Big|\!\Big|_{\ell^\infty} \le \|A^{[\tilde{N}]} \cdots A^{[1]}\mathbf{1}\|_{\ell^\infty}$. To conclude, we check by induction on $k \in \{1, \cdots, \tilde{N}\}$ that the entry $(A^{[k]} \cdots A^{[1]}\mathbf{1})_i$ is nonnegative and not greater than

$$\begin{cases} a_{ii+1}^{[k]} + a_{ii-1}^{[k]}\Big(a_{i-1i}^{[k-1]} + a_{i-1i-2}^{[k-1]}\big(a_{i-2i-1}^{[k-2]} + a_{i-2i-3}^{[k-2]}(\dots + a_{32}^{[k+3-i]}(a_{23}^{[k+2-i]} + 0))\big)\Big) & \text{if } 2 \le i \le k+1 \\ 1 & \text{if } \quad k+2 \le i \le N-(k+1) \\ a_{ii-1}^{[k]} + a_{ii+1}^{[k]}\Big(a_{i+1i}^{[k-1]} + a_{i+1i+2}^{[k-1]}\big(a_{i+2i+1}^{[k-2]} + a_{i+2i+3}^{[k-2]}(\dots + a_{N-2N-1}^{[k+i+2-N]}(a_{N-1N-2}^{[k+i+1-N]} + 0))\big)\Big) & \text{if } N-k \le i \le N-1 \end{cases}$$

where, by the assumptions made on the entries of the matrices, by induction on $i \in \{2, \cdots, k+1\}$,

$$a_{ii+1}^{[k]} + a_{ii-1}^{[k]}\Big(a_{i-1i}^{[k-1]} + a_{i-1i-2}^{[k-1]}\big(a_{i-2i-1}^{[k-2]} + a_{i-2i-3}^{[k-2]}(\ldots + a_{32}^{[k+3-i]}(a_{23}^{[k+2-i]} + 0))\big)\Big)$$
$$\leq 1 - a_{ii-1}^{[k]} a_{i-1i-2}^{[k-1]} \cdots a_{32}^{[k+3-i]}\big(1 - a_{23}^{[k+2-i]}\big) < 1$$

and, by backward induction on $i \in \{N-k, \cdots, N-1\}$,

$$a_{ii-1}^{[k]} + a_{ii+1}^{[k]}\Big(a_{i+1i}^{[k-1]} + a_{i+1i+2}^{[k-1]}\big(a_{i+2i+1}^{[k-2]} + a_{i+2i+3}^{[k-2]}(\ldots + a_{N-2N-1}^{[k+i+2-N]}(a_{N-1N-2}^{[k+i+1-N]} + 0))\big)\Big)$$
$$\leq 1 - a_{ii+1}^{[k]} a_{i+1i+2}^{[k-1]} \cdots a_{N-2N-1}^{[k+i+2-N]}\big(1 - a_{N-1N-2}^{[k+i+1-N]}\big) < 1.$$

In the induction on k, we use the first bound to get that $(A^{[k+1]} \cdots A^{[1]}\mathbf{1})_i \leq a_{ii+1}^{[k+1]} + a_{ii-1}^{[k+1]}(A^{[k]} \cdots A^{[1]}\mathbf{1})_{i-1}$ for $3 \leq i \leq k+1$ and the second one to get that $(A^{[k+1]} \cdots A^{[1]}\mathbf{1})_i \leq a_{ii-1}^{[k+1]} + a_{ii+1}^{[k+1]}(A^{[k]} \cdots A^{[1]}\mathbf{1})_{i+1}$ for $N-k \leq i \leq N-2$. □

Remarks. • Note that, as all the entries of the matrix $A_x^{[1]}$ such that $T(x) - x^\star = A_x^{[1]}(x - x^\star)$ are non negative, it is clear that if $x \geq x^\star$ (resp. $x \leq x^\star$) *componentwise* then $T(x) \geq x^\star$ (resp. $T(x) \leq x^\star$) *componentwise* so that if $x^{[0]} \geq x^\star$ (resp. $x^{[0} \leq x^\star$) *componentwise* then the whole sequence $x^{[\ell]}$ will satisfy the same inequality.

• This theorem shows the geometric convergence of this dual Lloyd procedure for the (log-concave) truncated exponential distributions toward its unique *quadratic* optimal dual quantizer. A specific family of procedures which works for the search of the L^r-optimal dual quantizer of power distributions for any $r \geq 1$ and truncated exponential distributions for $r \in \{1, 2\}$ is developed in the next section.

• In the log-concave example of the uniform distribution, say on the unit interval $[0, 1]$, one checks that the mapping T is affine and reads on $\overline{\mathcal{S}}_N^{a,b}$

$$T_i(x) = \frac{x_{i-1} + x_{i+1}}{2}, \quad i = 2, \ldots, N-1,$$

with the convention $x_1 = 0$ and $x_N = 1$ so that $T(x) = Ax + b$ with $A = [a_{ij}]_{2 \leq i,j \leq N-1}$ satisfying $a_{i,i\pm1} = \frac{1}{2}$, $i = 3, \ldots, N-2$, $a_{2,3} = \frac{1}{2} = a_{N-1,N-2}$ and $a_{ij} = 0$ otherwise and $b = \frac{1}{2}(0, \cdots, 0, 1)^*$ (with $N-2$ components). The eigenvalues of the symmetric matrix A are $\big(\cos(\frac{k\pi}{N-1})\big)_{1 \leq k \leq N-2}$ (see (2.13)). Therefore, for the Euclidean norm,

$$\forall x \in \mathbb{R}^{N-2},\ |Ax| \leq \cos\left(\frac{\pi}{N-1}\right)|x|.$$

Hence the sequence $x^{[\ell+1]} = T\big(x^{[\ell]}\big)$, $\ell \geq 0$, converges toward the unique equilibrium point $\big(\frac{k-1}{N-1}\big)_{k=2,\ldots,N-1}$ with the geometric rate $\cos\left(\frac{\pi}{N-1}\right)$ uniformly with respect to $x^{[0]} \in \overline{\mathcal{S}}_N^{a,b}$.

4 Computation of one-dimensional dual grids for specific distributions

4.1 Power distributions on compact intervals

Let μ be a probability measure compactly supported on $[0, 1]$ with density f and such that $0 < \mu([0, x]) < 1$ for all $x \in (0, 1)$. Then $x_1^{(N)} = 0$ and $x_N^{(N)} = 1$. To characterize the other points

$(x_i^{(N)})_{2\le i\le N-1}$ in the optimal dual grid when $N\ge 3$, we use the master equations written with the expression (2.10) of the gradient of the distortion :

$$(\Delta x_{i+1}^{(N)})^r\int_0^1\Psi_r(1-z)f(x_i^{(N)}+z\Delta x_{i+1}^{(N)})dz=(\Delta x_i^{(N)})^r\int_0^1\Psi_r(z)f\big(x_{i-1}^{(N)}+z\Delta x_i^{(N)}\big)dz,\ i=2:N-1$$

where $\Psi_r(z)=(r-1)z(1-z)^r+z^{r+1}+rz^2(1-z)^{r-1}$. For power distributions, $f(x)=\alpha x^{\alpha-1}$ for some $\alpha>0$, so that dividing the master equation by $\alpha(x_i^{(N)})^{r+\alpha-1}$ yields

$$\begin{aligned}\left(\frac{\Delta x_{i+1}^{(N)}}{x_i^{(N)}}\right)^r\int_0^1\Psi_r(1-z)\left(1+\frac{z\Delta x_{i+1}^{(N)}}{x_i^{(N)}}\right)^{\alpha-1}dz \qquad (4.35)\\ =\left(\frac{\Delta x_i^{(N)}}{x_i^{(N)}}\right)^r\int_0^1\Psi_r(z)\left(1+\frac{(z-1)\Delta x_i^{(N)}}{x_i^{(N)}}\right)^{\alpha-1}dz,\ i=2:N-1.\end{aligned}$$

We are going to check that the ratios $\lambda_i=\frac{x_i^{(N)}}{x_{i+1}^{(N)}}$ do not depend on $N\ge i+1$ (they of course depend on $r\ge 1$ but we do not make this dependence explicit in the notation). This is a consequence of the equality $\frac{x_1^{(N)}}{x_2^{(N)}}=0$ valid for each $N\ge 2$ and which yields $\lambda_1=0$. Since $x_N^{(N)}=1$, we then have $x_i^{(N)}=\prod_{j=i}^{N-1}\lambda_j$ for $i=1:N-1$ and even for $i=N$ under the usual convention $\prod_{j=N}^{N-1}\lambda_j=1$. Performing the change of variable $y=1-z$ in the integral in the right-hand side of (4.35), we obtain

$$\chi_r(\lambda_i^{-1}-1)=\chi_r(\lambda_{i-1}-1) \qquad (4.36)$$

where

$$\chi_r(x)=\begin{cases}x^r\int_0^1\Psi_r(1-z)(1+zx)^{\alpha-1}dz\ \text{ if }\ x\ge 0\\ (-x)^r\int_0^1\Psi_r(1-z)(1+zx)^{\alpha-1}dz\ \text{ if }\ x\in[-1,0].\end{cases}$$

To conclude that starting from $\lambda_1=0$, the values of λ_i can be computed inductively for $i=2:N-1$ from this equation, it is enough to check that χ_r is one to one on the interval $(0,+\infty)$ where $x_{i+1}^{(N)}/x_i^{(N)}-1$ stands. For $x\in(0,+\infty)$, we have

$$\chi_r'(x)=x^{r-1}\int_0^1\Psi_r(1-z)(r+(r+\alpha-1)zx)(1+zx)^{\alpha-2}dz$$

where the right-hand side is positive since $r\ge 1$ and Ψ_r is non-negative. Notice that we obtain uniqueness for the master equation and therefore uniqueness of the L^r-optimal dual quantization grid at level N even when $\alpha\in(0,1)$ and the density is not log-concave. In the quadratic $r=2$ case, since $\Psi_2(z)=z$, $\alpha\chi_2(x)=\frac{(1+x)^{1+\alpha}}{1+\alpha}-x-\frac{1}{1+\alpha}$. Of course when for $a<b$, μ admits the density $\mathbf{1}_{[a,b]}(x)\frac{\alpha(x-a)^{\alpha-1}}{(b-a)^\alpha}$ (resp. $\mathbf{1}_{[a,b]}(x)\frac{\alpha(b-x)^{\alpha-1}}{(b-a)^\alpha}$) then for $i=1:N$, $x_i^{(N)}=a+(b-a)\prod_{j=i}^{N-1}\lambda_j$ (resp. $x_{N+1-i}^{(N)}=b-(b-a)\prod_{j=i}^{N-1}\lambda_j$).

Numerical example. The optimal quadratic dual 10-quantizer of $\mu(dx)=\mathbf{1}_{[0,1]}(x)\frac{dx}{2\sqrt{x}}$ is given by

$$\{0,0.0744614,0.1675381,0.2704687,0.3804786,0.4961058,0.6164311,0.7408177,0.868795,1\}.$$

Notice that even if the density is log-convex on the interval $(0,1)$, the Lloyd-like iterative algorithm introduced in Section 3 still numerically converges to the corresponding unique solution to the master equation in $\mathcal{S}_{10}^{0,1}$.

The derivation of an equation not depending on $x_i^{(N)}$ relating λ_i to λ_{i-1} was permitted by the key structure condition

$$\forall (x,y) \in (0,1] \times [0,1],\ f(y) = g(x) h\left(\frac{y}{x}\right) \tag{4.37}$$

satisfied by the power density for $g(x) = \alpha x^{\alpha-1}$ and $h(z) = z^{\alpha-1}$. Under this structure condition, (4.36) remains valid for the following generalized definition of χ_r :

$$\chi_r(x) = \begin{cases} x^r \int_0^1 \Psi_r(1-z) h(1+zx) dz \text{ if } x \geq 0 \\ (-x)^r \int_0^1 \Psi_r(1-z) h(1+zx) dz \text{ if } x \in [-1,0]. \end{cases}$$

When the functions g and h are differentiable, the structure condition is only satisfied by power distributions. Indeed, differentiating with respect to x in (4.37), we obtain $\frac{xg'(x)}{g(x)} = \frac{\frac{y}{x} h'(\frac{y}{x})}{h(\frac{y}{x})}$. We deduce that the two functions $\frac{xg'(x)}{g(x)}$ and $\frac{zh'(z)}{h(z)}$ are both equal to some constant $\alpha - 1$. Then $g(x) \propto x^{\alpha-1}$ and $h(z) \propto z^{\alpha-1}$ so that $f(y) \propto y^{\alpha-1}$.

We could also assume that $f(y) = g(x)h(y-x)$, which is typically satisfied when $f(x) = \frac{|\lambda| e^{\lambda x}}{e^{\lambda} - 1}$ for $\lambda \in \mathbb{R}^*$ (we may then choose $h(z) = e^{\lambda z}$) but then it is not so easy to decouple the use of the two boundary conditions $x_1^{(N)} = 0$ and $x_N^{(N)} = 1$ which permits an inductive resolution of the master equation under the former structure condition (4.37). We are nevertheless able to design an almost explicit procedure for these truncated exponential distributions, at least when $\lambda > 0$ and $r \in \{1,2\}$.

4.2 Truncated exponential distributions

Let $\mu(dx) = \mathbf{1}_{[a,b]}(x) \frac{\lambda e^{-\lambda(x-a)}}{1-e^{-\lambda b}} dx$ be a truncated exponential distribution with parameter $\lambda > 0$ on $[a,b]$, $\infty < a < b < +\infty$. Note that if $\lambda < 0$, it suffices to solve the problem for the image $\tilde{\mu} = \mathbf{1}_{[-b,-a]}(x) \frac{|\lambda| e^{-|\lambda|(x+a)}}{e^{|\lambda| b} - 1} dx$ of μ by the linear transform $x \mapsto -x$ and transport the resulting dual quantizer by this involution.

The distribution μ is a log-concave distribution (though not strictly log-concave) so that, for every $r \geq 1$, the L^r-optimal dual quantizer, solution to the L^r-master equation (2.10) is unique at every level $N \geq 3$. Let $N \geq 3$ and let $x = (x_1, \ldots x_N)$, $x_1 = a$, $x_N = b$ and $\Delta x_i = x_i - x_{i-1}$, $i = 2, \ldots, N$. The master equation (2.10) reads

$$\Phi_r(\lambda \Delta x_{i+1}) = \Phi_r(-\lambda \Delta x_i),\ i = 2, \ldots, N-1,\ x_1 = a,\ x_N = b, \tag{4.38}$$

with

$$\Phi_r(x) = |x|^r e^{-x} \int_0^1 \Psi_r(z) e^{xz} dz, \quad x \in \mathbb{R}.$$

If $x^{N,\lambda,a,b}$ denotes the solution to this equation (where the dependence on $r \geq 1$ is not made explicit), one easily checks, taking advantage of uniqueness of the solution, that

$$x^{N,\lambda,a,b} = a \cdot \mathbf{1} + \frac{1}{\lambda} x^{N,1,0,\lambda(b-a)}$$

so we only need to solve the equation with $\lambda = 1$ and $a = 0$ with limit condition $x_N^{N,1,0,b} = b$.

– *Quadratic case* ($r = 2$). We first consider the quadratic case $r = 2$, most commonly (sic) used in applications. Then, $\Psi_2(z) = z$, so that

$$\Phi_2(x) = x^2 e^{-x} \int_0^1 z e^{xz} dz = e^{-x} - 1 + x, \quad x \in \mathbb{R}.$$

The function Φ_2 is C^1 and if we set $\check{\Phi}_2(x) = \Phi_2(-x)$, then $\Phi_{2|\mathbb{R}_+}$ and $\check{\Phi}_{2|\mathbb{R}_+}$ are both increasing C^1-diffeomorphisms of $\mathbb{R}_+$ and (with an obvious abuse of notation) Equation (4.38) reads in a forward way on $\mathbb{R}_+$

$$\Delta x_{i+1} = \theta_2(\Delta x_i),\ i = 2, \ldots, N-1 \quad \text{with} \quad \theta_2 = \Phi_2^{-1} \circ \check{\Phi}_2.$$

As $\Phi_2'(x) = x - \Phi_2(x)$, one checks that

$$\theta_2' = \frac{\check{\Phi}_2'}{\Phi_2'(\Phi_2^{-1}(\check{\Phi}_2))} = \frac{Id_{|\mathbb{R}_+} + \check{\Phi}_2}{\theta_2 - \check{\Phi}_2}$$

so that θ_2 satisfies the ordinary differential equation (*ODE*)

$$\tfrac{1}{2}(\theta_2^2)' = Id_{|\mathbb{R}_+} + \check{\Phi}_2 \cdot (1 + \theta_2').$$

At this stage, noting that, for every $x \in \mathbb{R}_+$,

$$\check{\Phi}_2(x) = e^x - 1 - x = \sum_{k \geq 2} \frac{x^k}{k!},$$

we aim at solving this *ODE* by power series i.e. we assume that

$$\theta_2(x) = \sum_{k \geq 1} a_k x^k$$

since $\theta_2(0) = \Phi_2^{-1} \circ \check{\Phi}_2(0) = 0$. By standard arguments, we see that

$$\tfrac{1}{2}(\theta_2^2)'(x) = \sum_{k \geq 1} b_k x^k \quad \text{with} \quad b_k = \tfrac{k+1}{2} \sum_{\ell=1}^{k} a_{k+1-\ell} a_\ell$$

and

$$x + \check{\Phi}_2(x)\big(1 + \theta_2'(x)\big) = \sum_{k \geq 1} c_k x^k \quad \text{with} \quad c_k = \Big(\sum_{\ell=0}^{k-2} \frac{\ell+1}{(k-\ell)!} a_{\ell+1} + \frac{1}{k!} \Big) \mathbf{1}_{\{k \geq 2\}} + \mathbf{1}_{\{k=1\}}.$$

One derives that $a_1^2 = 1$ so that $a_1 = 1$ since θ_2 is non-decreasing, $\frac{a_1+1}{2} = 3a_1 a_2$ which implies $a_2 = \frac{1}{3}$ and

$$a_k = \frac{1}{k+1} \left(\frac{2}{k!} + \sum_{\ell=2}^{k-1} \frac{\ell}{(k+1-\ell)!} a_\ell \right) - \frac{1}{2} \sum_{\ell=2}^{k-1} a_\ell a_{k+1-\ell}, \quad k \geq 3.$$

As a consequence, we can compute $\theta_2(x)$ with an arbitrary accuracy. Then, the master equation reduces to the scalar boundary condition

$$\sum_{k=0}^{N-2} \theta_2^{\circ k}(\Delta x_2) = b$$

which can be solved numerically by various elementary zero search methods like dichotomy, Newton-Raphson algorithm, etc, since for $k \geq 1$ the k-fold composition $\theta_2^{\circ k}$ of θ_2 is continuous and increasing on $\mathbb{R}_+$ from $\theta_2^{\circ k}(0) = 0$ to $\theta_2^{\circ k}(+\infty) = +\infty$. Then $x_k^{N,1,0,b} = \sum_{j=0}^{k-2} \theta_2^{\circ j}(\Delta x_2)$ for $k = 1, \ldots, N$.

Numerical example. The optimal quadratic dual 11-quantizer of the truncated exponential distribution with parameter $\lambda = 1$ over the unit interval ($a = 0$, $b = 1$) is given by

$$x^{11,1,0,1} = (0, 0.086271, 0.17510, 0.26663, 0.36105, 0.45853, 0.55929, 0.66355, 0.77156, 0.88361, 1).$$

CPU time on a 1.8 MHz processor with Matlab: $6.10^{-3}\, s$ (using a dichotomy algorithm to determine Δx_2).

– *General case* ($r \in (1, +\infty)$). Set $\check{\Phi}_r(x) = \Phi_r(-x)$ for every $x \geq 0$. By an obvious change of variable, one has

$$\check{\Phi}_r(x) = x^r \int_0^1 \Psi_r(1-z) e^{zx} dz, \quad x \geq 0,$$

so that $\check{\Phi}_r(0) = 0$, $\check{\Phi}_r$ is increasing on $\mathbb{R}_+$, goes to infinity at infinity by the monotone convergence theorem since $\Psi_r > 0$ on $(0, 1)$. As a consequence it is a C^1 homeomorphism of $(0, +\infty)$ (in fact a diffeomorphism since $\check{\Phi}_r'$ is never 0 on $(0, +\infty)$). Equation (4.38) can be written in a backward way

$$\Delta x_i = \widetilde{\theta}_r(\Delta x_{i+1}),\ i = 2, \ldots, N-1 \quad \text{with} \quad \widetilde{\theta}_r = (\check{\Phi}_r)^{-1} \circ \Phi_r.$$

Now let us focus for a while on Φ_r itself on $\mathbb{R}_+$. As $\Psi_r(z) \geq z^{r+1}$ on $[0, 1]$ (see (2.8)), one has for every $x > 0$,

$$\begin{aligned}
\Phi_r(x) &\geq x^r e^{-x} \int_0^1 z^{r+1} e^{xz} dz = x^{-2} e^{-x} \int_0^x z^{r+1} e^z dz \\
&= x^{r-1} - (r+1) x^{-2} e^{-x} \int_0^x z^r e^z dz \\
&\geq x^{r-1} - (r+1) x^{r-2} (1 - e^{-x}) \geq \tfrac{1}{2} x^{r-1} \quad \text{for} \quad x \geq 2(r+1).
\end{aligned}$$

Hence $\lim_{x \to +\infty} \Phi_r(x) = +\infty$ which in turn implies that $\lim_{x \to +\infty} \widetilde{\theta}_r(x) = +\infty$. Consequently the continuous function $x \mapsto \sum_{k=1}^{N-1} \widetilde{\theta}_r^{\circ(N-k)}(x)$ is null at 0 and goes to infinity at infinity so that the equation

$$\sum_{k=0}^{N-2} \widetilde{\theta}_r^{\circ k}(x) = b$$

always has a solution $x^{r,b}$ and we may set $\Delta x_i = \widetilde{\theta}_r^{\circ(N-i)}(x^{r,b})$, $i = 2, \ldots, N$. We know that the solution is unique by Theorem 2.1.

Unfortunately, we have no semi-closed form for $\widetilde{\theta}_r$ like in the quadratic case for θ_2 since we could not find an *ODE* satisfied by $\widetilde{\theta}_r$ in full generality.

When r is an integer, then Ψ_r is a polynomial function with degree $r-1$ if r is even and $r+1$ if r is odd, whose coefficients of degrees 0 and r are always 0. Then, having in mind that

$$\forall\, n\in\mathbb{N},\quad e^{-x}\int_0^x z^n e^z dz = (-1)^n n!\Big(\sum_{k=0}^{n}(-1)^k\frac{x^k}{k!}-e^{-x}\Big)$$

it follows that Φ_r reads

$$\Phi_r(x)=\operatorname{sign}(x)^r\left(P_r(x)-e^{-x}\Big(\frac{c_r}{x^2}+Q_r(x)\Big)\right)$$

where P_r and Q_r are polynomial functions with degree $r-1$ and $r-2$ respectively that can be computed explicitly and $c_r=0$ if r is even.

– Case $r=1$. When $r=1$, $\theta_1(z)=2z(1-z)$ so that $\Psi_1(z)=2z^2$. Hence

$$\begin{aligned}\Phi_1(x)&=\frac{2\operatorname{sign}(x)}{x^2}e^{-x}\int_0^x z^2e^z dz=\frac{4\operatorname{sign}(x)}{x^2}\Big(\frac{x^2}{2}-x+1-e^{-x}\Big),\\ \check\Phi_1(x)&=\frac{4\operatorname{sign}(x)}{x^2}\Big(e^x-1-x-\frac{x^2}{2}\Big).\end{aligned}\tag{4.39}$$

In particular, Φ_1 is increasing on $\mathbb{R}_+$, C^1, with $\lim_{x\to+\infty}\Phi_1(x)=2$ so that Φ_1 is a C^1-diffeomorphism from $(0,+\infty)$ to $(0,2)$. Moreover, it is clear that $\check\Phi_1$ is a C^1-diffeomorphism of $(0,+\infty)$.

In that case we can again write the equation in a forward way

$$\Delta x_{i+1}=\theta_1(\Delta x_i),\ i=2,\dots,N-1\qquad\text{where}\qquad \theta_1=\Phi_1^{-1}\circ\check\Phi_1$$

is defined, C^1, increasing on $\big[0,(\check\Phi_1)^{-1}(2)\big)$ non-negative, satisfies $\theta_1(0)=0$ and $\lim_{x\to\check\Phi_1^{-1}(2)}\theta_1(x)=+\infty$. Consequently the mapping $x\mapsto\sum_{k=1}^{N-1}\theta_1^{\circ k}(x)$ is defined on an open bounded interval with left endpoint 0, null at 0 and goes to infinity at its right endpoint so that the equation

$$\sum_{k=0}^{N-2}\theta_1^{\circ k}(x)=b$$

always has a solution $x^{1,b}$ and we may set $\Delta x_i=\theta_1^{\circ(i-2)}(x^{1,b})$, $i=2,\dots,N$.

Moreover, Φ_1 satisfies the following *ODE* on $\mathbb{R}\setminus\{0\}$,

$$\Phi_1'(x)=-\Big(\frac{2}{x}+1\Big)\Phi_1(x)+2\operatorname{sign}(x)$$

from which we derive that, for $x>0$ small enough,

$$\theta_1'(x)=\frac{(\check\Phi_1)'}{\Phi_1'\circ\theta_1}(x)=\frac{\check\Phi_1'(x)}{2-\big(\frac{2}{\theta_1(x)}+1\big)\Phi_1(\theta_1(x))}=\frac{\check\Phi_1'(x)}{2-\big(\frac{2}{\theta_1(x)}+1\big)\check\Phi_1(x)}$$

which can be rewritten (in a neighbourhood of 0 on $\mathbb{R}_+$) as the non-linear *ODE*

$$2\check\Phi_1(x)\theta_1'(x)-(\theta_1^2)'(x)\big(1-\tfrac12\check\Phi_1(x)\big)+\theta_1(x)\check\Phi_1'(x)=0,\quad\theta_1(0)=0.$$

This *ODE* can be solved as a power series with positive convergence radius. First note that

$$\forall\, x \geq 0, \quad \check{\Phi}_1(x) = \sum_{k\geq 1} b_k x^k \quad \text{and} \quad b_k = \frac{4}{(k+2)!}$$

(so that $b_1 = \frac{2}{3}$, $b_2 = \frac{1}{6}$, etc) owing to (4.39). Assume *a priori* that θ_1 can be expanded as

$$\theta_1(x) = \sum_{k\geq 1} a_k x^k, \quad x \geq 0.$$

Then, if we set

$$\widetilde{a}_k^{(2)} = (k+1)\sum_{\ell=1}^{k} a_\ell a_{k+1-\ell},\ k \geq 1, \text{ so that } (\theta_1^2)'(x) = \sum_{k\geq 1} a_k^{(2)} x^k,$$

elementary though tedious computations show that the sequence $(a_k)_{k\geq 1}$ satisfies the following induction formula (with the convention $\sum_{\varnothing} = 0$)

$$a_1 = 1, \quad a_k = \frac{3}{2(k+2)}\left(\sum_{\ell=1}^{k-1}\left(\widetilde{a}_\ell^{(2)}\frac{b_{k-\ell}}{2} + (k-\ell+1)a_\ell b_{k-\ell+1}\right) - (k+1)\sum_{\ell=2}^{k-1} a_\ell a_{k+1-\ell}\right),\ k \geq 2,$$

Remark. Another (less tractable) inductive formula can be derived by dealing directly with the identity $\Phi_1 \circ \theta_1 = \check{\Phi}_1$.

5 Concluding and provisional remarks

In this paper we established for dual (or Delaunay) quantization uniqueness of the critical points (hence of the optimal quantizer) of its L^r-quantization errors at level N under a log-concave assumption on the (density of) the distribution under consideration. This is the exact counterpart of Trushkin's Theorem established for primal (or Voronoi) quantization, except for the bounded support intrinsic restriction inherent to Delaunay quantization. We also devised an avatar of the celebrated Lloyd algorithm (also known as k-means or Forgy's algorithm in its batch version) for Delaunay quantization in the quadratic setting. This avatar also converges at an exponential rate still under this strong unimodal assumption. We finally propose a way to exploit the "master equations " for specific distributions, possibly in non-quadratic settings.

Future investigations should focus on a tractable L^r-extension of this "Delaunay" Lloyd-like algorithm. The main asset of Delaunay quantization compared its Voronoi counterpart is the fact that any dual quantization is unbiased, regardless of its optimality, which is for instance an important advantage when using quantization for unbiased information transmission and more recently for federated computation (see e.g. [12] among many others). But this still requires algorithms to design the underlying (hyper-)triangulation (see [20]) which is quite demanding as the dimension grows. On the other hand, for numerical purposes, the optimality remains crucial in medium dimension. Therefore elucidating the structure of optimal dual quantizers (rather than looking for a hopeless uniqueness result) as well as proposing efficient algorithms to compute them remain major challenges. Note that these issues have not been satisfactorily solved so far for Voronoi quantization either.

References

[1] Alfonsi, A. Corbetta J. and Jourdain, B. (2020). Sampling of probability measures in the convex order by Wasserstein projection, *Annales de l'Institut Henri Poincaré B, Probabilités et Statistiques*, **56**(3):1706-1729.

[2] Cohort, P. (2000). *Sur quelques problèmes de quantification*, PhD thesis, www.theses.fr/2000PA066112, Univ. Pierre & Parie Curie, Paris, 187p.

[3] 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*, **22**(3).

[4] De March, H. (2018). Entropic approximation for multi-dimensional martingale optimal transport, *arXiv 1812.11104*.

[5] Gersho A. and Gray R.M. eds (1982). Special issue on quantization. *IEEE Trans. Inform. Theory*, **28** (2), Vol I & II.

[6] Graf, S. and Luschgy, H. (2000). *Foundations of quantization for probability distributions*, LNM 1730, Springer, Berlin, 230p.

[7] Guo, G. and Oblòj, J. (2017). Computational Methods for Martingale Optimal Transport problems, *Ann. Appl. Probab.*, **29** (6):3311–3347.

[8] Henry-Labordère, P. (2019). (Martingale) optimal transport and anomaly detection with neural networks : a primal-dual algorithm, *arXiv:1904.04546*.

[9] Jourdain, B. and Pagès, G. (2020). Quantization and martingale couplings, *arXiv:2012.10370*.

[10] Jourdain, B. and Pagès, G. (2019). Convex order, quantization and monotone approximations of ARCH models, *arXiv:1910.00799*.

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

[12] Konečný, J. McMahan, H. B. Yu, F. Richtçrik, P. Suresh, A.T. Bacon, D. (2017). Federated Learning: Strategies for Improving Communication Efficiency, arXiv1610.05492, arxiv.org/abs/1610.05492.

[13] Lloyd, S.P. (1957). Least squares quantization in PCM, *IEEE Transactions on Information Theory*, **28**(2):129–137 (reprinted from a Bell Telephone Memorandum labs, Murray Hill, NJ, 1957).

[14] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pp. 281-297.

[15] Lamberton, D. and Pagès, G. (1996). On the critical points of the 1-dimensional Competitive Learning Vector Quantization Algorithm. Proceedings of the ESANN'96, (ed., M. Verleysen), Editions D Facto, Bruxelles, 97–106.

[16] Pagès, G. (2018). *Numerical Probability: an introduction with applications to Finance*, Springer-Verlag, xvi +579p.

[17] Pagès, G. (2015). Introduction to optimal quantization for numerics, *ESAIM Proc. & Surveys*, **48**:29–79.

[18] Pagès, G. (2016). Convex order for path-dependent derivatives: a dynamic programming approach. *Séminaire de Probabilités XLVIII*, C. Donati, A. Lejay, A. Rouault eds, LNM 2168, Springer, Cham, 33–96.

[19] Pagès, G. and Wilbertz, B. (2012). Dual Quantization for random walks with application to credit derivatives, *J. Comp. Finance*, **16**(2):33–60.

[20] Pagès, G. and Wilbertz, B. (2012). Intrinsic stationarity for vector quantization: foundation of dual quantization. *SIAM J. Numer. Anal.* **50**(2):747–780.

[21] Pagès, G. and Wilbertz, B. (2012). Optimal Delaunay and Voronoi quantization schemes for pricing American style options. *Numerical methods in Finance*, 171–213, Springer Proc. Math., **12**, Springer, Heidelberg.

[22] Pagès, G. and Wilbertz, B. (2018). Sharp rate for the dual quantization problem, *Séminaire de Probabilités XLV*, C. Donati, A. Lejay, A. Rouault eds, LNM 2215, Springer, Cham, 119–164.

[23] 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.

[24] Struwe M. (1990). *Variational Methods (Application to non linear p.d.e & Hamiltonian Systems)*, Springer, 244p.