ISSN 1066-5307, Mathematical Methods of Statistics, 2010, Vol. 19, No. 2, pp. 136–150. © Allerton Press, Inc., 2010.

L_1-Quantization and Clustering in Banach Spaces

T. Laloë[1*]

[1]*Univ. de Lyon, Univ. de Lyon 1, ISFA, Laboratoire SAF, France*

Received January 29, 2009; in final form, April 15, 2010

Abstract—Let X be a random variable with distribution μ taking values in a Banach space $\mathcal{H}$. First, we establish the existence of an optimal quantization of μ with respect to the L_1-distance. Second, we propose several estimators of the optimal quantizer in the potentially infinite-dimensional space $\mathcal{H}$, with associated algorithms. Finally, we discuss practical results obtained from real-life data sets.

Key words: quantization, clustering, L_1-distance, Banach space.

2000 Mathematics Subject Classification: 62H30.

DOI: 10.3103/S1066530710020031

1. INTRODUCTION

Clustering consists in partitioning a data set into subsets (or clusters), so that the data in each subset share some common trait. Proximity is determined according to some distance measure. For a thorough introduction to the subject, we refer to the book by Kaufman and Rousseeuw [14]. The origin of clustering goes back to 45 years ago, when some biologists and sociologists began to search for automatic methods to build different groups with their data. Today, clustering is used in many fields. For example, in medical imaging, it can be used to differentiate between different types of tissue and blood in a three-dimensional image. Market researchers use it to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers. There are also many different applications in artificial intelligence, sociology, medical research, or political sciences.

In the present paper, the clustering method we investigate is based on the technique of quantization commonly used in signal compression (Graf and Luschgy [12], Linder [17]). Given a normed space $(\mathcal{H}, \|\cdot\|)$, a codebook (of size k) is defined by a subset $\mathcal{C} \subset \mathcal{H}$ with cardinality k. Then, each $x \in \mathcal{H}$ is represented by a unique $\hat{x} \in \mathcal{C}$ via the function q,

$$\begin{aligned} q\colon \ \mathcal{H} &\to \mathcal{C}, \\ x &\to \hat{x}, \end{aligned}$$

which is called a quantizer. Here we come back to the clustering, as we create clusters in the data by regrouping the observations which have the same image by q.

Denote by d the distance induced by the norm on $\mathcal{H}$:

$$\begin{aligned} d\colon \mathcal{H} \times \mathcal{H} &\to \mathbb{R}^+, \\ (x, y) &\to \|x - y\|. \end{aligned}$$

In this paper, observations are modeled by a random variable X on $\mathcal{H}$ with distribution μ. The quality of the approximation of X by $q(X)$ is then given by the distortion $\mathbb{E}\, d(X, q(X))$. Thus the aim is to minimize $\mathbb{E}\, d(X, q(X))$ among all possible quantizers. However, in practice, the distribution μ of the

*E-mail: thomas.laloe@laposte.net

observations is unknown, and we only have at hand n independent observations $X_1, \ldots, X_n$ with the same distribution as X. The goal is then to minimize the empirical distortion:

$$\frac{1}{n}\sum_{i=1}^{n} d(X_i, q(X_i)).$$

We choose here the distortion measure d for the robustness of the medians. Moreover the properties of the L_1-Wasserstein norm will be primordial in our demonstrations.

Since the early work of Hartigan [13] and Pollard [19], [20], [21], the performance of clustering have been considered by many authors. Convergence properties of the minimizer q_n^* of the empirical distortion have been mostly studied in the case when $\mathcal{H} = \mathbb{R}^d$. Consistency of q_n^* was shown by Pollard [19], [21] and Abaya and Wise [1]. Rates of convergence have been considered by Pollard [20], Linder *et al.* [18], Linder [17].

As a matter of fact, in many practical problems, input data items are in the form of random functions (speech recordings, spectra, images) rather than standard vectors, and this casts the clustering problem into the general class of functional data analysis. Even though in practice such observations are observed at discrete sampling points, the challenge in this context is to infer the data structure by exploiting the infinite-dimensional nature of the observations. The last few years have witnessed important developments in both the theory and practice of functional data analysis, and many traditional data analysis tools have been adapted to handle functional inputs. The book by Ramsay and Silverman [22] provides a comprehensive introduction to the area. Recently, Biau *et al.* [3] gave some consistency results in Hilbert spaces and with an L_2-based distortion.

Thus, the first novelty in this paper is to consider data taking place in a separable and reflexive Banach space, with no restriction on their dimension. The second novelty is that we consider an L_1-based distortion, which leads to more robust estimators. For a discussion of the advantage of the L_1-distance we refer the reader to the paper by Kemperman [15].

This setup calls for substantially different arguments to prove results which are known to be true when considering finite-dimensional spaces and an L_2-based distortion. In particular, specific notions will be required, such as weak topology (Dunford and Schwartz [10]), lower semi-continuity (Ekeland and Temam [10]) and entropy (Van der Vaart and Wellner [23]).

The paper is organized as follows. We first provide the formal context of quantization in Banach space in the first part of Section 2. Then, we focus on the problem of existence of an optimal quantizer. In Sections 3 and 4 we study two consistent estimators of this optimal quantizer, and we confront them to real-life data in Section 5. Proofs are collected in the Appendix.

2. QUANTIZATION IN A BANACH SPACE

2.1. General Framework

The fact that the closed bounded balls are not compact is a major problem when considering infinite-dimensional spaces. To overcome this, the classical solution is to consider reflexive spaces, i.e., spaces in which the closed bounded balls are compact for the weak topology (Dunford and Schwartz [9]). Thus, throughout the paper, $(\mathcal{H}, \|\cdot\|)$ will denote a reflexive and separable Banach space. We let X be an $\mathcal{H}$-valued random variable with distribution μ such that $\mathbb{E}\|X\| < \infty$.

Given a set $\mathcal{C} = \{y_i\}_{i=1}^k$ of points in $\mathcal{H}^k$, any Borel function $q\colon \mathcal{H} \to \mathcal{C}$ is called a quantizer. The set $\mathcal{C}$ is called a codebook, and the y_i, $i = 1, \ldots, k$, are the centers of $\mathcal{C}$. The error made by replacing X by $q(X)$ is measured by the distortion:

$$D(\mu, q) = \mathbb{E}\, d(X, q(X)) = \int_{\mathcal{H}} \|x - q(x)\|\, \mu(dx).$$

Note that $D(\mu, q) < \infty$ since $\mathbb{E}\|X\| < \infty$. For a given k, the aim is to minimize $D(\mu, \cdot)$ among the set $\mathcal{Q}_k$ of all possible k-quantizers. The optimal distortion is then defined by

$$D_k^*(\mu) = \inf_{q \in \mathcal{Q}_k} D(\mu, q).$$

When it exists, a quantizer q^* satisfying $D(\mu, q^*) = D_k^*(\mu)$ is said to be an optimal quantizer.

Any quantizer is characterized by its codebook $\mathcal{C} = \{y_i\}_{i=1}^k$ and a partition of $\mathcal{H}$ in cells $S_i = \{x \in \mathcal{H}\colon q(x) = y_i\}$, $i = 1, \ldots, k$, via the rule

$$q(x) = y_i \iff x \in S_i.$$

Thus, from now on, we will define a quantizer by its codebook and its cells.

Let us consider the particular family of Voronoi partitions constructed by the nearest neighbor rule. That is, for each center of the codebook, a cell is constituted by the elements $x \in \mathcal{H}$ closest to it (Gersho and Gray [11]). A quantizer with such a partition is named a nearest neighbor quantizer, and we denote by $\mathcal{Q}_{knn}$ the set of all k-nearest neighbor quantizers. It can be easily proven (see Lemma 1 in Linder [17]) that

$$\inf_{q \in \mathcal{Q}_k} D(\mu, q) = \inf_{q \in \mathcal{Q}_{knn}} D(\mu, q).$$

More precisely, given two quantizers $q \in \mathcal{Q}_k$ and $q' \in \mathcal{Q}_{knn}$ with the same codebook, we have

$$D(\mu, q') \leq D(\mu, q).$$

Therefore, in the following, we will restrict ourselves to nearest neighbor quantizers.

A complementary result (see Lemma 2 in Linder [17]) is that for a quantizer q with codebook $\mathcal{C}$ and partition S, a quantizer q' with the same partition but with a codebook defined by

$$y_i' \in \arg\min_{y \in \mathcal{H}} \mathbb{E}\big[\|X - y\| \mid X \in S_i\big], \qquad i = 1, \ldots, k,$$

satisfies

$$D(\mu, q') \leq D(\mu, q).$$

Note that since $\mathcal{H}$ is a reflexive Banach space, $\arg\min_{y \in \mathcal{H}} \mathbb{E}[\|X - y\| \mid X \in S_i]$ is non empty thanks to Kemperman [15].

From the two previous optimality results, on the codebook and associated partition, we can derive a simple algorithm in order to find a good quantizer. This algorithm is called the Lloyd algorithm and based on the so-called Lloyd iteration (Gersho and Gray [11], Chapter 6). The outline is as follows:

1. Choose randomly an initial codebook;
2. Given a codebook C_m, build the associated Voronoi partition;
3. Build C_{m+1}, the optimal codebook for the previous partition;
4. Stop when the distortion no longer decreases.

Unfortunately, this algorithm has two drawbacks: it depends on the initial codebook chosen, and it does not necessarily converge to the optimal distortion. In Section 4 we will discuss an alternative to this algorithm leading to an optimal quantizer.

2.2. Existence of an Optimal Quantizer

The aim of this section is to show that the minimization problem of $D(\mu, q)$ has at least one solution. Recall that we consider only nearest neighbor quantizers, which can be entirely characterized by their codebook $(y_1, \ldots, y_k)$, and set $\mathbf{y}_k = (y_1, \ldots, y_k)$.

We denote the associated distortion by

$$D(\mu, q) = D(\mu, \mathbf{y}_k).$$

Therefore our first task is to prove that the function $D(\mu, \cdot)$ has at least one minimum, or, in other words, that there exists at least one optimal codebook.

Theorem 2.1. *Assume that $\mathcal{H}$ is a reflexive and separable Banach space. Then, the function $D(\mu, \cdot)$ admits at least one minimum.*

Theoretically speaking, it is of interest to search for an optimal quantizer. To make the link with clustering, Theorem 2.1 states that there exists at least one optimal repartition of the space $\mathcal{H}$ in different clusters. The next step is to consider the statistical case, in which the distribution of X is unknown.

3. A CONSISTENT ESTIMATOR

3.1. Construction and Consistency

In a statistical context, the distribution μ of X is unknown and we only have at hand n random variables, $X_1, \ldots, X_n$, independent and distributed as X. Let the empirical measure μ_n be defined as

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[X_i \in A]}$$

for any measurable set $A \subset \mathcal{H}$. For any quantizer q, the associated empirical distortion is then given by

$$D(\mu_n, q) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - q(X_i)\|.$$

An empirical quantizer (that is a quantizer depending on the sample set $(X_1, \ldots, X_n)$) $q_n^* = q_n^*(\cdot, X_1, \ldots, X_n)$ satisfying

$$q_n^* \in \arg\min_{q \in \mathcal{Q}_k} \sum_{i=1}^{n} \|X_i - q(X_i)\|$$

is said to be empirically optimal. In particular, if we set (with a slight abuse of notation)

$$D(\mu, q_n^*) = \mathbb{E}\big[\|X - q_n^*(X)\| \mid X_1 \ldots, X_n\big],$$

we have

$$D(\mu_n, q_n^*) = D_k^*(\mu_n).$$

From Theorem 2.1, we know that for every n, an empirically optimal quantizer always exists.

The following theorem, which is an adaptation of Theorem 2 in Linder [17], establishes the asymptotic optimality of the quantizer q_n^* with respect to the distortion.

Theorem 3.1. *Assume that $\mathcal{H}$ is a reflexive and separable Banach space and set $k \geq 1$. Then, any sequence of empirically optimal k-quantizers $(q_n^*)_{n \geq 1}$ satisfies*

$$\lim_{n \to \infty} D(\mu, q_n^*) = D_k^*(\mu) \quad a.s.$$

and

$$\lim_{n \to \infty} D(\mu_n, q_n^*) = D_k^*(\mu) \quad a.s.$$

3.2. Rate of Convergence

Most results in the literature concern the situation when $\mathcal{H} = \mathbb{R}^d$ and the distortion is an L_2-based one (Pollard [20], Linder [17], Linder *et al.* [18]). For example, it is shown in [17] that if there exists $T > 0$ such that $\mathbb{P}[\|X\| \leq T] = 1$, then

$$\mathbb{E}\, D(\mu, q_n^*) - D^*(\mu) \leq C T^2 \sqrt{\frac{k(d+1)\log(k(d+1))}{n}},$$

where $C > 0$ is a universal constant.

Recently, Biau *et al.* [3] proved that when $\mathcal{H}$ is a Hilbert space and the distortion is an L_2-based one, then

$$\mathbb{E}\, D(\mu, q_n^*) - D^*(\mu) \leq C \frac{k}{\sqrt{n}},$$

where $C > 0$ is a universal constant.

In the sequel, our goal is to establish a rate of convergence in a Banach space with an L_1-criterion. This will require some new notions.

Let $\mathcal{P}(\mathcal{H})$ be the set of all probability measures on $\mathcal{H}$.

Definition 3.1. Let $p \in [1, \infty[$.

1. The L_p-Wasserstein distance between $\phi, \xi \in \mathcal{P}(\mathcal{H})$ is defined by:

$$\rho_p(\phi, \xi) = \inf_{X \sim \phi, Y \sim \xi} \left(\mathbb{E}\, d\big(X, Y\big)^p \right)^{\frac{1}{p}}.$$

2. A probability $\phi \in \mathcal{P}(\mathcal{H})$ satisfies a transportation inequality $T_p(\lambda)$ if there exists $\lambda > 0$ such that, for any probability $\xi \in \mathcal{P}(\mathcal{H})$,

$$\rho_p^p(\phi, \xi) \leq \sqrt{\frac{2}{\lambda} H(\xi \mid \phi)},$$

where

$$H(\xi \mid \phi) = \int_{\mathcal{H}} \frac{d\xi}{d\phi} \log \left(\frac{d\xi}{d\phi} \right) d\phi$$

is the Kullback information between ϕ and ξ.

Remarks:

- The L_p-Wasserstein distance, also called L_p-Kantorovich distance, is known to be appropriate for the quantization problem (Graf and Luschgy, Section 3 [12]);
- For this choice of distance, in view of getting rates of convergence, the so-called transportation inequalities, or Talagrand inequalities, are well designed (Ledoux [16]).

Generally speaking, it is a difficult task to determine whether a probability $\mu \in \mathcal{P}(\mathcal{H})$ satisfies a transportation inequality $T_p(\lambda)$. However, the problem is simpler when $p = 1$ as expressed in the theorem below proven in Djellout *et al.* [7] (Theorem 2.3 and Section 1).

Theorem 3.2. *A probability $\phi \in \mathcal{P}(\mathcal{H})$ satisfies a transportation inequality $T_1(\lambda)$ if and only if, for all $\alpha < \lambda/2$,*

$$\int_{\mathcal{H}} e^{\alpha \|x - y\|^2} d\mu(x) < \infty$$

for one (and therefore for all) y in $\mathcal{H}$.

In the sequel, we will only consider the case $p = 1$, and we set $\rho = \rho_1$. For any set $\Lambda \subset \mathcal{H}$, let $\mathcal{P}(\Lambda)$ be the set of all probability measures on Λ. Let also $\mathcal{N}(r, \Lambda)$ be the smallest number of balls of radius r (for the metric ρ) required to cover $\mathcal{P}(\Lambda)$, that is

$$\mathcal{N}(r, \Lambda) = \inf \left\{ n \in \mathbb{N} \text{ s.t. } \exists x_1, \ldots, x_n \in \mathcal{P}(\Lambda) \colon \bigcup_{i=1}^{n} B_{\mathcal{P}(\Lambda)}(x_i, r) \supset \mathcal{P}(\Lambda) \right\},$$

where $B_{\mathcal{P}(\Lambda)}(x_i, r)$ is the ball in $\mathcal{P}(\Lambda)$ centered at x_i and with radius r (for the metric ρ). The quantity $\log(\mathcal{N}(r, \Lambda))$ is the entropy of $\mathcal{P}(\Lambda)$ (Van der Vaart and Wellner [23]).

In the same way, let $N(r, \Lambda)$ be the smallest number of balls of radius r required to cover Λ, with respect to the metric of $\mathcal{H}$.

In order to state a rate of convergence for $D^*(\mu_n)$, we introduce the following assumptions:

A1: There exists $\lambda > 0$ such that μ satisfies a transportation inequality $T_1(\lambda)$;

A2: There exists a Banach space $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ and a compact embedding

$$I\colon (\mathcal{G}, \|\cdot\|_{\mathcal{G}}) \hookrightarrow (\mathcal{H}, \|\cdot\|)$$

such that the closed bounded balls B in $\mathcal{G}$ are totally bounded if we see them as subsets of $\mathcal{H}$. That is, for all $r > 0$, $N(r, \overline{I(B)})$ is finite. Moreover, we suppose that $\mu \in \mathcal{P}(I(\mathcal{G}))$.

Note that **A1** is satisfied for paths of stochastic differential equations

$$dX_t = b(X_t)\,dt + s(X_t)\,dW_t,$$

where $t \in [0, T]$, $T < \infty$, and $b(\cdot)$, $s(\cdot)$ satisfy suitable properties (Djellout *et al.* [7], Corollary 4.1). **A2** is satisfied, for example, if $\mathcal{G}$ is a Sobolev space on a compact domain of $\mathbb{R}^d$ (Cucker and Smale [6], Example 3).

From now on, B_R stands for the ball of center 0 and radius R in $\mathcal{G}$. According to assumption **A2** and Theorem A.1 in Bolley *et al.* [4] there exists a positive constant C such that for all $r, R > 0$,

$$\mathcal{N}(r, \overline{I(B_R)}) \le \left(\frac{CR}{r}\right)^{N(r/2, \overline{I(B_R)})}. \tag{3.1}$$

In this context, the crux is to identify the entropy of the balls in $\mathcal{G}$ with respect to the norm $\|\cdot\|$. For some examples, we refer the reader to [2].

Theorem 3.3. *Assume that $\mathcal{H}$ is a reflexive and separable Banach space and* **A1**, **A2** *are satisfied. Then, for all $\lambda' < \lambda$ and $\varepsilon > 0$, there exist three positive constants K, γ, and R_1 such that if $R = R_1 \max(1, \varepsilon^2, \log(1/\varepsilon^2))^{1/2}$ and $n \ge K \log(\mathcal{N}(\gamma\varepsilon, B_R))/\varepsilon^2$, we have:*

$$\mathbb{P}[\rho(\mu, \mu_n) \ge \varepsilon] \le e^{-(\lambda'/2)n\varepsilon^2}.$$

Using the inequality

$$D(\mu, q_n^*) - D^*(\mu) \le 2\rho(\mu, \mu_n),$$

we deduce the following corollary.

Corollary 3.1. *Assume that $\mathcal{H}$ is a reflexive and separable Banach space and* **A1**, **A2** *are satisfied. Then, for all $\lambda' < \lambda$ and $\varepsilon > 0$, there exist three positive constants K, γ, and R_1 such that if $R = R_1 \max(1, \varepsilon^2, \log(1/\varepsilon^2))^{1/2}$ and $n \ge K \log(\mathcal{N}(\gamma\varepsilon, B_R))/\varepsilon^2$, we have:*

$$\mathbb{P}[D(\mu, q_n^*) - D(\mu, q^*) \ge \varepsilon] \le e^{-(\lambda'/8)n\varepsilon^2}.$$

Let $\mathcal{R}$ be the function from $\mathbb{R}_+^*$ to $\mathbb{R}_+^*$ defined by

$$\mathcal{R}(x) = R_1 \max\left(1, x^2, \log(1/x^2)\right)^{1/2},$$

and denote $\mathcal{M}$ the function from $\mathbb{R}_+^*$ to $\mathbb{R}_+^*$ defined by

$$\mathcal{M}(x) = K \log(\mathcal{N}(\gamma x, B_{\mathcal{R}(x)}))/x^2. \tag{3.2}$$

Theorem 3.4 below gives us the desired rate of convergence.

Theorem 3.4. *Assume that $\mathcal{H}$ is a reflexive and separable Banach space,* **A1**, **A2** *are satisfied, and $\mathcal{M}$ is invertible on some interval $]0, a]$. Then, there exists $C_0 > 0$ such that*

$$\mathbb{E}\, D(\mu, q_n^*) - D(\mu, q^*) \le C_0 \max(\mathcal{M}^{-1}(n), n^{-1/2}).$$

Note there is no restriction on the support of μ. In particular, we do not require that the support of μ is bounded. This is an important point, since such an assumption is not verified, for example, by the distributions of classical diffusion processes, yet widely used in stochastic modeling.

Besides, the interval $]0, a]$ in Theorem 3.4 is not empty, for instance, in all the examples cited in the paper of Biau *et al.* [2].

Example: Suppose that Assumption **A1** is satisfied. Consider the Example 3 in Cucker and Smale [6], in which $\mathcal{G}$ is a Sobolev space on a compact domain set of $\mathbb{R}^d$. Using the entropy of $\overline{I(B_R)} \subset \mathcal{H}$ (Cucker and Smale [6]) and Theorem 3.4, we have

$$\mathbb{E}\, D(\mu, q_n^*) - D(\mu, q^*) \leq \frac{C}{(\log n)^{s/d}},$$

where C is a positive constant.

3.3. Algorithm

Calculating q_n^* appears to be an NP-complete problem. In order to approximate q_n^* one can adapt the Lloyd algorithm, which has been presented in Section 2, to the statistical context, in which we use μ_n instead of μ. Moreover, rather to calculate empirical medians in each cell, a possible solution is to consider medoids, i.e., centers taken within the sample $\{X_1, \ldots, X_n\}$. For more details about the Lloyd algorithm and medoids, we refer the reader to the book by Kaufman and Rousseeuw [14].

However, this Lloyd algorithm with medoids has the same drawbacks as the Lloyd algorithm presented in Section 2: non-optimality and dependence on the initial codebook. Thus, in the next section, we will present a new estimator in order to overcome these drawbacks.

4. MINIMIZATION ON DATA

4.1. Construction and Consistency

The basic idea of the estimator presented in this section consists in searching the minimum of the empirical distortion $D(\mu_n, \cdot)$ within the sample $\{X_1, \ldots, X_n\}$. It is a generalization of a method of Cadre [5] who considered the case $k = 1$ only. Formally, our estimator $\mathbf{y}_{k,n}^* = (y_{1,n}^*, \ldots, y_{k,n}^*)$ is defined by

$$\mathbf{y}_{k,n}^* \in \underset{\mathbf{z} \in \{X_1, \ldots, X_n\}^k}{\arg\min} D(\mu_n, \mathbf{z}).$$

Let $\|\cdot\|_k$ be a norm on $\mathcal{H}^k$ (as an example, for $\mathbf{z} = (z_1, \ldots, z_k) \in \mathcal{H}^k$, $\|\mathbf{z}\|_k = \max_{i=1,\ldots,k} \|z_i\|$) and $B_{\mathcal{H}^k}(\mathbf{z}, r)$ the associated closed ball in $\mathcal{H}^k$ centered at $\mathbf{z}$ and with radius r.

Theorem 4.1. *Assume that $\mathcal{H}$ is a reflexive and separable Banach space and there exists $\mathbf{y}_k^*$ an optimal codebook for μ, which satisfies*

$$\forall \varepsilon > 0, \quad \mathbb{P}\big[(X_1, \ldots, X_k) \in B_{\mathcal{H}^k}(\mathbf{y}_k^*, \varepsilon)\big] > 0. \tag{4.3}$$

Then,

$$\lim_{n \to \infty} D(\mu, \mathbf{y}_{k,n}^*) = D_k^*(\mu) \quad a.s.$$

Remark: Condition (4.3) in Theorem 4.1 simply requires that the probability that k observations fall in the neighborhood of $\mathbf{y}_k^*$ is not zero. The necessity of this condition is easy to understand. Indeed, suppose there exists $\varepsilon > 0$ such that for optimal codebook $\mathbf{y}_k^*$ for μ, $(X_1, \ldots, X_k) \notin B_{\mathcal{H}^k}(\mathbf{y}_k^*, \varepsilon)$ with probability 1. Then, by construction, $D(\mu, \mathbf{y}_{k,n}^*)$ can not converge to $D_k^*(\mu)$.

Theorem 4.2. *Assume that $\mathcal{H}$ is a reflexive and separable Banach space and* (4.3)*,* **A1***, and* **A2** *hold. Then, we have*

$$\lim_{n \to \infty} \mathbb{E} D(\mu, \mathbf{y}_{k,n}^*) = D_k^*(\mu).$$

4.2. Rate of Convergence

The next theorem states that $D(\mu, \mathbf{y}_{n,k}^*)$ converges to $D_k^*(\mu)$ at the same rate as $D(\mu, q_n^*)$. Remember that the function $\mathcal{M}$ is defined in (3.2), and let $\mathbf{y}_k^*$ be an optimal codebook for μ. For $\varepsilon > 0$ we set

$$f(\mathbf{y}_k^*, \varepsilon) = \mathbb{P}\big[(X_1, \ldots, X_k) \in B_{\mathcal{H}^k}(\mathbf{y}_k^*, \varepsilon)\big].$$

We also introduce the assumption:

A3: There exist a decreasing function $V : \mathbb{N}^* \to \mathbb{R}_+^*$ and positive constants u, v, C such that

$$\max\left(\int_0^u \big(1 - f(\mathbf{y}_k^*, \varepsilon)\big)^{\lfloor n/k \rfloor}\, d\varepsilon, \int_v^{+\infty} \big(1 - f(\mathbf{y}_k^*, \varepsilon)\big)^{\lfloor n/k \rfloor}\, d\varepsilon\right) \le V(n).$$

Theorem 4.3. *Assume that $\mathcal{H}$ is a reflexive and separable Banach space and* **A1** *and* **A2** *are satisfied. Let $\mathbf{y}_k^*$ be an optimal codebook for μ satisfying* **A3**. *Then, if $\mathcal{M}$ is invertible on some interval $]0, b]$, there exists a positive constant C_0 such that, for n large enough,*

$$\mathbb{E}D(\mu, \mathbf{y}_{k,n}^*) - D_k^*(\mu) \le C_0 \max\big(\mathcal{M}^{-1}(n), V(n), \lfloor n/k \rfloor^{-1/2}\big).$$

Remarks:

- Assumption **A3** requires that the probability that data are present in a neighborhood of an optimal quantizer increases fast enough with n. It is an essential assumption in the proof of Theorem 4.3.
- Assumption **A3** is satisfied if the following assumptions hold:

  A4: There exists $c_1 > 0$ such that $f(\mathbf{y}_k^*, \varepsilon) \ge 1 - \exp(-\varepsilon^2)$ for $\varepsilon \in]0, c_1]$;

  A5: There exists $c_2 > 0$ such that $f(\mathbf{y}_k^*, \varepsilon) \ge 1 - \exp(-\varepsilon^2)$ for $\varepsilon \in [c_2, +\infty[$.

- Assume that **A4** and **A5** are satisfied. Then we have

  $$\mathbb{E}D(\mu, \mathbf{y}_{k,n}^*) - D_k^*(\mu) \le C_0 \max\big(\mathcal{M}^{-1}(n), \lfloor n/k \rfloor^{-1/2}\big).$$

  That is, $D(\mu_n, \mathbf{y}_{n,k}^*)$ converges to $D_k^*(\mu)$ at the same rate as $D_k^*(\mu_n)$.

- Assumption **A5** is satisfied if μ has a bounded support.

5. CONCLUSION

This paper thus provides an answer to the problem of functional L_1-clustering: first, we prove that for any measure $\mu \in \mathcal{P}(\mathcal{H})$ with finite moment an optimal quantization always exists (Theorem 2.1). Then we propose a consistent estimator of q^* (Theorem 3.1) and state its rate of convergence (Theorem 3.4). In order to offset the main drawbacks of the Lloyd algorithm, we derive another method minimizing directly the distortion on the data.

One of the most interesting points in our results is that the assumptions we make are as light as possible. For example, we made no restriction on the support of μ, and the assumptions **A1**, **A2** are satisfied in classical stochastic modeling.

APPENDIX: PROOFS

A.1. Proof of Theorem 2.1

Before we prove Theorem 2.1, we need the following definition.

Definition A.1. A function $\phi\colon \mathcal{H} \to \bar{\mathbb{R}}$ is called lower semi-continuous for the weak topology (abbreviated weakly l.s.c.) if it satisfies one of the following equivalent conditions:

(i) $\forall t \in \mathbb{R}$, $\{u \in \mathcal{H}\colon \phi(u) \leq t\}$ is closed for the weak topology.

(ii) $\forall \bar{u} \in \mathcal{H}$, $\liminf_{u \overset{w}{\to} \bar{u}} \phi(u) \geq \phi(\bar{u})$ (where $\overset{w}{\to}$ denotes the weak convergence in $\mathcal{H}$).

For a proof of this equivalence and of the following proposition, we refer the reader to the book by Ekeland and Temam [10].

Proposition A.1. *With the notation of Definition* A.1*, the two following properties hold*:

(i) *If ϕ is continuous and convex, then it is weakly l.s.c.*

(ii) *If ϕ is weakly l.s.c. on a set Λ which is compact for the weak topology, then ϕ has a minimum on Λ.*

Lemma A.1 is a straightforward adaptation of the results proven in the first part of the proof of Theorem 1 in Linder [17].

Lemma A.1. *There exists $A > 0$ and $\ell \leq k$ such that*

$$\inf_{\mathbf{y}_k \in \mathcal{H}^k} D(\mu, \mathbf{y}_k) = \inf_{\mathbf{y}_\ell \in B_A^\ell} D(\mu, \mathbf{y}_\ell).$$

For all x in $\mathcal{H}$, we define the functions $g_{i,x}\colon \mathcal{H}^k \to \mathbb{R}$ and $g_x\colon \mathcal{H}^k \to \mathbb{R}$ by:

$$g_{i,x}(\mathbf{y}_k) = \|x - y_i\|,$$

and

$$g_x(\mathbf{y}_k) = \min_{i=1,\ldots,k} g_{i,x}(\mathbf{y}_k).$$

Lemma A.2. *For any x in $\mathcal{H}$, the function g_x is weakly l.s.c. on $\mathcal{H}^k$.*

Proof. For each x in $\mathcal{H}$, the functions $g_{i,x}$ are continuous and convex, thus they are weakly l.s.c. according to Proposition A.1. For all t in $\mathbb{R}$, the sets

$$\{\mathbf{y}_k \in \mathcal{H}^k\colon g_{i,x}(\mathbf{y}_k) \leq t\}$$

are then weakly closed. We deduce that

$$\{\mathbf{y}_k \in \mathcal{H}^k\colon g_x(\mathbf{y}_k) \leq t\} = \bigcup_{i=1}^{k} \{\mathbf{y}_k \in \mathcal{H}^k\colon g_{i,x}(\mathbf{y}_k) \leq t\}$$

is weakly closed. Lemma A.2 follows by using statement (i) in Definition A.1. □

Lemma A.3. *The function $D(\mu, \cdot)$ is weakly l.s.c. on $\mathcal{H}^k$.*

Proof. For each $\mathbf{y}_k^* \in \mathcal{H}^k$, we can write:

$$\begin{aligned}
\liminf_{\mathbf{y}_k \xrightarrow{w} \mathbf{y}_k^*} D(\mu, \mathbf{y}_k) &= \liminf_{\mathbf{y}_k \xrightarrow{w} \mathbf{y}_k^*} \int_{\mathcal{H}} g_x(\mathbf{y}_k)\, \mu(dx) \\
&\geq \int_{\mathcal{H}} \liminf_{\mathbf{y}_k \xrightarrow{w} \mathbf{y}_k^*} g_x(\mathbf{y}_k)\, \mu(dx) \qquad \text{(by Fatou's Lemma)} \\
&\geq \int_{\mathcal{H}} g_x(\mathbf{y}_k^*)\, \mu(dx) \qquad \text{(by Lemma A.2 and Definition A.1 (ii))} \\
&= D(\mu, \mathbf{y}_k^*),
\end{aligned}$$

which proves that $D(\mu, \cdot)$ satisfies Definition A.1 (ii). □

We are now in a position to prove Theorem 2.1.

Proof of Theorem 2.1. According to Lemma A.1, there exists $R > 0$ such that the infimum of $D(\mu, \cdot)$ on $\mathcal{H}^k$ is also the infimum of $D(\mu, \cdot)$ on B_R^k. Moreover, on the one hand B_R^k is compact for the weak topology, and on the other hand $D(\mu, \cdot)$ is weakly l.s.c. according to Lemma A.3. Thus, according to Proposition A.1, the function $D(\mu, \cdot)$ reaches its infimum on B_R^k. □

A.2. Proof of Theorem 3.3

The proof is adapted from the proof of Theorem 1 by Bolley *et al.* [4]. It can be decomposed in three steps:

1. First, we show we can consider truncated versions of the probability measures μ and μ_n on the ball B_R of $\mathcal{G}$;
2. Then we cover the space $\mathcal{P}(I(B_R))$ by small balls of radius r;
3. Finally, we optimize the various parameters introduced in the proof.

Each of the next three lemmas matches a step.

Remember that under Assumption **A2**, $\mu \in \mathcal{P}(\mathcal{G})$. Let $R > 0$. We consider μ^R defined, for any Borel set $A \subset \mathcal{H}$, by:

$$\mu^R[A] = \frac{\mu[A \cap I(B_R)]}{\mu[I(B_R)]} = \mu[A \mid I(B_R)].$$

Consider now the independent random variables $\{X_i\}_{i=1}^n$ with distribution μ and $\{Y_i\}_{i=1}^n$ with distribution μ^R. We define, for $i \leq n$,

$$X_i^R = \begin{cases} X_i & \text{if} \quad \|X_i\|_{\mathcal{G}} \leq R, \\ Y_i & \text{if} \quad \|X_i\|_{\mathcal{G}} > R. \end{cases}$$

Let δ_x be the Dirac measure at point x. The empirical measures μ_n and μ_n^R are defined by

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i} \qquad \text{and} \qquad \mu_n^R = \frac{1}{n} \sum_{i=1}^n \delta_{X_i^R}.$$

Denote $E_\alpha = \int_{\mathcal{H}} \exp(\alpha \|x\|^2)\, \mu(dx)$. Since we suppose that μ satisfies a $T_1(\lambda)$-inequality, we have $E_\alpha < \infty$ for $\alpha < \lambda/2$.

Lemma A.4. *Let $\eta \in]0,1[$, $\varepsilon, \theta > 0$, $\alpha_1 \in]0, \lambda/2[$, and $\alpha \in]\alpha_1, \lambda/2[$. Then, for all*

$$R > \max(\sqrt{1/2\alpha}, 2\theta/\alpha_1),$$

we have

$$\mathbb{P}[\rho(\mu, \mu_n) > \varepsilon] \le \mathbb{P}\big[\rho(\mu^R, \mu_n^R) > \eta\varepsilon - 2E_\alpha R e^{-\alpha R^2}\big] + \exp\big(-n[\theta(1-\eta)\varepsilon - E_\alpha e^{(\alpha_1-\alpha)R^2}]\big).$$

Proof. For a fixed $\varepsilon > 0$, we bound $\mathbb{P}[\rho(\mu, \mu_n) > \varepsilon]$ by a function of μ^R and μ_n^R. First, following the arguments of the proof of Theorem 1.1 by Bolley *et al.* (step 1) [4], it can be proven that for all $\alpha < \lambda/2$ and $R \ge \sqrt{1/2\alpha}$,

$$\rho(\mu, \mu^R) \le 2E_\alpha R e^{-\alpha R^2}. \tag{A.1}$$

Second, the probability measures μ_n and μ_n^R satisfy

$$\rho(\mu, \mu_n^R) \le \frac{1}{n}\sum_{i=1}^n \|X_i^R - X_i\| \le \frac{1}{n}\sum_{i=1}^n Z_i,$$

where $Z_i = 2\|X_i\|\mathbf{1}_{\|X_i\|_{\mathcal{G}} > R}$ $(i = 1, \dots, n)$. Using a similar argument as in the proof of Theorem 1.1 by Bolley *et al.* (step 1) [4], we deduce that if ε, θ are positive and $\alpha < \lambda/2$,

$$\mathbb{P}[\rho(\mu, \mu_n^R) > \varepsilon] \le \exp\big(-n[\theta\varepsilon - E_\alpha e^{(\alpha_1-\alpha)R^2}]\big). \tag{A.2}$$

The conclusion follows from (A.1), (A.2), and the triangle inequality for ρ. □

Lemma A.5. *Given $\theta, \alpha, \alpha_1, \lambda_1 > 0$ such that $\lambda_1 < \lambda$, $\alpha \in]\alpha_1, \lambda/2[$, and $\zeta > 1$, there exist positive constants $\delta_1, \lambda_2 < \lambda_1$, K_1, and K_2 such that, for all $R > \zeta \max\left(\sqrt{1/2\alpha}, 2\theta/\alpha_1\right)$ and $\varepsilon > 0$,*

$$\begin{aligned}\mathbb{P}[\rho(\mu, \mu_n) > \varepsilon] \le \mathcal{N}(\delta_1\varepsilon/2, B_R)\exp\Big(-n\Big[\frac{\lambda_2}{2}\varepsilon^2 - K_1R^2e^{-\alpha R^2}\Big]\Big)\\ + \exp\big(-n[K_2\zeta\varepsilon - K_3e^{(\alpha_1-\alpha)R^2}]\big),\end{aligned}$$

where K_3 is a positive constant depending only on θ and α_1.

Proof. We start by proving that μ^R satisfies a modified $T_1(\lambda)$-inequality. Let Λ be a Borel set of $\mathcal{P}(\overline{I(B_R)})$. Following the arguments of the proof of Theorem 1.1 of Bolley *et al.* (step 2) [4], one may write

$$\mathbb{P}[\mu_n^R \in \Lambda] \le \exp\big(-n \inf_{\nu\in\Lambda} H(\nu \mid \mu^R)\big). \tag{A.3}$$

From now on, we assume that $\mathcal{P}(\overline{I(B_R)})$ is equipped with the distance ρ. Consider $\delta > 0$ and A a measurable subset of $\mathcal{P}(\overline{I(B_R)})$. We set $\mathcal{N}^A = \mathcal{N}(\delta/2, A)$. Then there exist $\mathcal{N}^A$ balls B_i, $i = 1, \dots, \mathcal{N}^A$, covering A. Each of this balls is convex and included in the δ-neighborhood A_δ of A. Moreover, by assumption **A2**, the balls B_i are totally bounded.

It is easily inferred from equation (A.3) that

$$\mathbb{P}[\mu_n^R \in A] \le \mathcal{N}^A \exp\big(-n \inf_{\nu\in A_\delta} H(\nu \mid \mu^R)\big). \tag{A.4}$$

Define now

$$A = \big\{\nu \in \mathcal{P}(\overline{I(B_R)})\colon \rho(\nu, \mu^R) \ge \eta\varepsilon - 2E_\alpha R e^{-\alpha R^2}\big\}.$$

According to the basic inequality

$$\forall a \in]0,1[, \exists C > 0 \text{ such that } \forall x, y \in \mathbb{R}, (x-y)^2 \ge (1-a)x^2 - Cy^2, \tag{A.5}$$

we have, for any $\nu \in A_\delta$,

$$\forall \lambda_1 < \lambda, \exists K > 0 \text{ such that } H(\nu \mid \mu^R) \ge \frac{\lambda_1}{2}\rho^2(\mu^R, \nu) - KR^2e^{-\alpha R^2}.$$

Thus we can write

$$\forall \nu \in A_\delta, \quad H(\nu \mid \mu^R) \geq \frac{\lambda_1}{2}\rho^2(\mu^R, \nu) - KR^2e^{-\alpha R^2} \geq \frac{\lambda_1}{2}m^2 - KR^2e^{-\alpha R^2},$$

where

$$m = \max\big(\eta\varepsilon - 2E_\alpha Re^{-\alpha R^2} - \delta, 0\big).$$

From this and equation (A.4) we conclude that

$$\mathbb{P}\big[\rho(\mu^R, \mu_n^R) \geq \eta\varepsilon - 2E_\alpha Re^{-\alpha R^2}\big] \leq \mathcal{N}^{\mathcal{A}} \exp\Big(- n\Big[\frac{\lambda_1}{2}m^2 - KR^2e^{-\alpha R^2}\Big]\Big). \tag{A.6}$$

Now, given $\lambda_2 < \lambda_1$, it follows from (A.5) that there exist three positive constants δ_1, η_1, and K_1 depending only on α, λ_1, and λ_2 such that

$$\frac{\lambda_1}{2}m^2 - KR^2e^{-\alpha R^2} \geq \frac{\lambda_2}{2}\varepsilon^2 - K_1R^2e^{-\alpha R^2},$$

where $\delta = \delta_1\varepsilon$. This leads, together with (A.6), to

$$\mathbb{P}\big[\rho(\mu^R, \mu_n^R) \geq \eta\varepsilon - 2E_\alpha Re^{-\alpha R^2}\big] \leq \mathcal{N}^{\mathcal{A}} \exp\Big(- n\Big[\frac{\lambda_2}{2}\varepsilon^2 - K_1R^2e^{-\alpha R^2}\Big]\Big). \tag{A.7}$$

To bound $\mathcal{N}^A$, we observe that since $A \subset \mathcal{P}(\overline{I(B_R)})$,

$$\mathcal{N}^A \leq \mathcal{N}(\delta/2, \overline{I(B_R)}) = \mathcal{N}(\delta_1\varepsilon/2, \overline{I(B_R)}).$$

The conclusion follows by Lemma A.4 and inequality (A.7). □

The following lemma simplifies the results of the previous one.

Lemma A.6. *Let $\lambda' < \lambda$, $\alpha < \lambda/2$, and $\alpha' < \alpha$. There exists $\delta_1 > 0$ such that, for all $\varepsilon > 0$,*

$$\mathbb{P}\big[\rho(\mu, \mu_n) > \varepsilon\big] \leq \exp\Big(- \frac{\lambda'}{2}n\varepsilon^2\Big) + \exp(-\alpha' n\varepsilon^2)$$

as soon as

$$R^2 \geq R_2 \max\Big(1, \varepsilon^2, \log\Big(\frac{1}{\varepsilon^2}\Big)\Big) \qquad \text{and} \qquad n \geq K_4\frac{\log\big(\mathcal{N}(\delta_1\varepsilon/2, \overline{I(B_R)})\big)}{\varepsilon^2},$$

where R_2 and K_4 are some positive constants depending on μ through λ and α.

Proof. On the one hand, under the assumptions and notation of Lemma A.5, we have, for all $\lambda' < \lambda_2$,

$$\log\Big(\mathcal{N}(\delta_1\varepsilon/2, \overline{I(B_R)}) \exp\Big(- n\Big[\frac{\lambda_2}{2}\varepsilon^2 - K_1R^2e^{-\alpha R^2}\Big]\Big)\Big) \leq \frac{-n\lambda'\varepsilon^2}{2} \tag{A.8}$$

as soon as R, $R/\log(1/\varepsilon^2)$, and $n\varepsilon^2/\log(\mathcal{N}(\delta_1\varepsilon/2, \overline{I(B_R)}))$ are large enough (see the third step of the proof of Theorem 1.1 by Bolley *et al.* [4]).

On the other hand, let $\alpha' < \alpha_2 < \alpha_1$. We can choose ζ such that $K_2\zeta = \alpha_2\varepsilon$. With this choice we obtain

$$\exp\big(- n[K_2\zeta\varepsilon - K_3e^{(\alpha_1-\alpha)R^2}]\big) = \exp\big(- n[\alpha_2\varepsilon^2 - K_3e^{(\alpha_1-\alpha)R^2}]\big),$$

which can be bounded by $\exp(-\alpha' n\varepsilon^2)$, for R and $R^2/\log(1/\varepsilon^2)$ large enough. This, together with (A.8), leads to the conclusion. □

Theorem 3.3 is then a straightforward consequence of Lemma A.6, noticing that, for any $K < \min((\lambda'/2), \alpha')$ and n large enough, we have

$$\exp\Big(- \frac{\lambda'}{2}n\varepsilon^2\Big) + \exp\big(- \alpha' n\varepsilon^2\big) \leq \exp\big(- Kn\varepsilon^2\big).$$

A.3. Proof of Theorem 3.4

Let $\varepsilon > 0$ be small enough. According to Corollary 3.1 we have

$$\mathbb{P}\big[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\big] \leq e^{-(\lambda'/8)n\varepsilon^2}$$

as soon as $n \geq \mathcal{M}(\varepsilon)$. Therefore we can write:

$$\begin{aligned}
\mathbb{E}D(\mu, q_n^*) - D(\mu, q^*) &= \int_0^{+\infty} \mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] d\varepsilon \\
&= \int_0^{\mathcal{M}^{-1}(n)} \mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] d\varepsilon + \int_{\mathcal{M}^{-1}(n)}^{+\infty} \mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] d\varepsilon \\
&\leq \mathcal{M}^{-1}(n) + \int_0^{+\infty} e^{-(\lambda'/8)n\varepsilon^2} d\varepsilon \leq C_0 \max(\mathcal{M}^{-1}(n), n^{-1/2}),
\end{aligned}$$

as desired. □

A.4. Proof of Theorem 4.1

One can easily show that

$$D(\mu_n, \mathbf{y}_{k,n}^*) - D(\mu, \mathbf{y}_{k,n}^*) \leq \rho(\mu, \mu_n). \tag{A.9}$$

Thus, by Lemma 4 in Linder [17] and Varadarajan's Theorem [8], we deduce that:

$$D(\mu_n, \mathbf{y}_{k,n}^*) - D(\mu, \mathbf{y}_{k,n}^*) \to 0 \quad \text{a.s.} \quad \text{as } n \to \infty. \tag{A.10}$$

Let $p \leq n$ and $\mathbf{z} \in \{X_1, \ldots, X_p\}^k$. Since $D(\mu_n, \mathbf{y}_{k,n}^*) \leq D(\mu_n, \mathbf{z})$ and, by the law of large numbers, $D(\mu_n, \mathbf{z}) \to D(\mu, \mathbf{z})$ a.s., we have

$$\limsup_n D(\mu_n, \mathbf{y}_{k,n}^*) \leq D(\mu, \mathbf{z}) \quad \text{a.s.}$$

From (A.10), we deduce that, for all $p \geq 1$,

$$\limsup_n D(\mu, \mathbf{y}_{k,n}^*) \leq \min_{\mathbf{z} \in \{X_1, \ldots, X_p\}^k} D(\mu, \mathbf{z}). \tag{A.11}$$

Let us now evaluate the limit of the right-hand term in (A.11) as $p \to \infty$. Note, for $\varepsilon > 0$ and $p \geq 1$,

$$N(p, \varepsilon) = \Big[\exists\, \mathbf{z}^* \in \operatorname*{arg\,min}_{\mathbf{z} \in \{X_1, \ldots, X_p\}^k} D(\mu, \mathbf{z}) \cap B_{\mathcal{H}^k}(\mathbf{y}_k^*, \varepsilon),\ D(\mu, \mathbf{z}^*) \geq D(\mu, \mathbf{y}_k^*) + 2\varepsilon\Big].$$

Since, $\forall\, \mathbf{y}_k, \mathbf{y}_k' \in \mathcal{H}^k,\ |D(\mu, \mathbf{y}_k) - D(\mu, \mathbf{y}_k')| \leq \|\mathbf{y}_k - \mathbf{y}_k'\|_k$, we obtain

$$N(p, \varepsilon) \subset \big[D(\mu, \mathbf{y}_k^*) \geq D(\mu, \mathbf{y}_k^*) + \varepsilon\big] = \emptyset.$$

Therefore as soon as $p \geq k$,

$$\begin{aligned}
\mathbb{P}\Big[\min_{\mathbf{z} \in \{X_1, \ldots, X_p\}^k} D(\mu, \mathbf{z}) - D(\mu, \mathbf{y}_k^*) > 2\varepsilon\Big] &\leq \mathbb{P}[N(p, \varepsilon)] + \mathbb{P}\big[\forall\, \mathbf{z} \in \{X_1, \ldots, X_p\}^k,\ \mathbf{z} \notin B_{\mathcal{H}^k}(\mathbf{y}_k^*, \varepsilon)\big] \\
&\leq \mathbb{P}\big[(X_1, \ldots, X_k) \notin B_{\mathcal{H}^k}(\mathbf{y}_k^*, \varepsilon)\big]^{\lfloor p/k \rfloor} = \big(1 - \mathbb{P}[(X_1, \ldots, X_k) \in B_{\mathcal{H}^k}(\mathbf{y}_k^*, \varepsilon)]\big)^{\lfloor p/k \rfloor},
\end{aligned} \tag{A.12}$$

where $\lfloor \cdot \rfloor$ stands for the integer part function. Then, by the Borel–Cantelli lemma,

$$\lim_{p \to \infty} \min_{\mathbf{z} \in \{X_1, \ldots, X_p\}^k} D(\mu, \mathbf{z}) = D(\mu, \mathbf{y}_k^*) \quad \text{a.s.}$$

This result, together with (A.11), leads to the conclusion. □

A.5. Proof of Theorem 4.2

On the one hand we can write:

$$\begin{aligned} D(\mu, \mathbf{y}^*_{k,n}) - D^*(\mu) &= D(\mu, \mathbf{y}^*_{k,n}) - D(\mu_n, \mathbf{y}^*_{k,n}) + D(\mu_n, \mathbf{y}^*_{k,n}) - D^*(\mu) \\ &\le |D(\mu, \mathbf{y}^*_{k,n}) - D(\mu_n, \mathbf{y}^*_{k,n})| + |D(\mu_n, \mathbf{y}^*_{k,n}) - D^*(\mu)| \\ &\le \rho(\mu, \mu_n) + |D(\mu_n, \mathbf{y}^*_{k,n}) - D^*(\mu)|, \end{aligned}$$

according to (A.9).

On the other hand,

$$\lim_{n\to\infty} D(\mu_n, \mathbf{y}^*_{k,n}) = D^*(\mu) \quad \text{a.s.}$$

Moreover,

$$\begin{aligned} D(\mu_n, \mathbf{y}^*_{k,n}) &= \min_{\mathbf{z}\in\{X_1,\dots,X_n\}^k} \frac{1}{n}\sum_{i=1}^n \min_{j=1,\dots,k} \|X_i - \mathbf{z}_j\| \\ &\le \frac{1}{n}\sum_{i=1}^n \|X_i - X_1\| \le \frac{1}{n}\sum_{i=1}^n \|X_i\| + \|X_1\|. \end{aligned}$$

Hence $D(\mu_n, \mathbf{y}^*_{k,n})$ is uniformly integrable, which proves that it converges in L_1.

Finally, $\mathbb{E}\rho(\mu, \mu_n) \to 0$ by Theorem 3.3, and we deduce the proof of Theorem 4.2. □

A.6. Proof of Theorem 4.3

First we can write

$$\begin{aligned} D(\mu, \mathbf{y}^*_{k,n}) - D^*_k(\mu) = D(\mu, \mathbf{y}^*_{k,n}) - D(\mu_n, \mathbf{y}^*_{k,n}) + D(\mu_n, \mathbf{y}^*_{k,n}) \\ - \min_{\mathbf{z}\in\{X_1,\dots,X_n\}^k} D(\mu, \mathbf{z}) + \min_{\mathbf{z}\in\{X_1,\dots,X_n\}^k} D(\mu, \mathbf{z}) - D^*_k(\mu). \end{aligned}$$

Then, according to Lemma 3 in Linder [17], we have

$$D(\mu, \mathbf{y}^*_{k,n}) - D(\mu_n, \mathbf{y}^*_{k,n}) \le \rho(\mu, \mu_n)$$

and

$$D(\mu_n, \mathbf{y}^*_{k,n}) - \min_{\mathbf{z}\in\{X_1,\dots,X_n\}^k} D(\mu, \mathbf{z}) \le \rho(\mu, \mu_n).$$

Thus,

$$D(\mu, \mathbf{y}^*_{k,n}) - D^*_k(\mu) \le 2\rho(\mu, \mu_n) + \min_{\mathbf{z}\in\{X_1,\dots,X_n\}^k} D(\mu, \mathbf{z}) - D^*_k(\mu). \tag{A.13}$$

Moreover, according to inequality (A.12) we have for $n \ge k$:

$$\mathbb{P}\Big[\min_{\mathbf{z}\in\{X_1,\dots,X_n\}^k} D(\mu, \mathbf{z}) - D^*_k(\mu) \ge 2\varepsilon\Big] \le \big[1 - f(\mathbf{y}^*_k, \varepsilon)\big]^{\lfloor n/k \rfloor}.$$

We deduce

$$\begin{aligned} \mathbb{E}\min_{\mathbf{z}\in\{X_1,\dots,X_n\}^k} D(\mu, \mathbf{z}) - D^*_k(\mu) &= \int_0^{+\infty} \mathbb{P}\Big[\min_{\mathbf{z}\in\{X_1,\dots,X_n\}^k} D(\mu, \mathbf{z}) - D^*_k(\mu) \ge \varepsilon\Big]\, d\varepsilon \\ &\le 2\int_0^{+\infty} \big(1 - f(\mathbf{y}^*_k, \varepsilon)\big)^{\lfloor n/k \rfloor} d\varepsilon \\ &\le 2\Bigg(\int_{[0,u]\cup[v,\infty[} \big(1 - f(\mathbf{y}^*_k, \varepsilon)\big)^{\lfloor n/k \rfloor}\, d\varepsilon + \int_u^v \big(1 - f(\mathbf{y}^*_k, \varepsilon)\big)^{\lfloor n/k \rfloor}\, d\varepsilon\Bigg) \end{aligned}$$

$$\leq 2\left(2V(n) + \int_u^v \left[1 - f(\mathbf{y}_k^*, \varepsilon)\right]^{\lfloor n/k \rfloor} d\varepsilon\right) \quad \text{(according to assumption } \mathbf{A3}\text{)}$$

$$\leq 2\big(2V(n) + (v-u)\Gamma^{\lfloor n/k \rfloor}\big) \leq C \max\left(\lfloor n/k \rfloor^{-1/2}, V(n)\right) \quad \text{for } n \text{ large enough,}$$

where $\Gamma < 1$ and C are some positive constants. Theorem 4.3 follows from (A.13), Theorem 3.3, and Theorem 3.4. □

REFERENCES

1. E. A. Abaya and G. L. Wise "Convergence of Vector Quantizers with Application to Optimal Quantization", SIAM J. Appl. Math. **44**, 183–189 (1984).
2. G. Biau, F. Cérou, and A. Guyader, "Rate of Convergence of the Functional k-Nearest Neighbor Estimate", IEEE Trans. Inform. Theory (2010) (in press).
3. G. Biau, L. Devroye, and G. Lugosi, "On the Performance of Clustering in Hilbert Spaces", IEEE Trans. Inform. Theory, **54**, 781–790 (2007).
4. F. Bolley, A. Guillin, and C. Villani, "Quantitative Concentration Inequalities for Empirical Measures on Noncompact Spaces", Probab. Theory and Rel. Fields **137** (3–4), 541–593 (2007).
5. B. Cadre, "Convergent Estimators for the L_1-Median of a Banach Valued Random Variable", Statistics **35** (4), 509–521 (2001).
6. F. Cucker and S. Smale, "On the Mathematical Foundations of Learning", Amer. Math. Soc. Bulletin. New Series **39** (1), 1–49 (2002) (electronic).
7. H. Djellout, A. Guillin, and L. Wu, "Transportation Cost-Information Inequalities and Applications to Random Dynamical Systems and Diffusions", Ann. Probab. **32** (3B), 2702–2732 (2004).
8. R. M. Dudley, *Real Analysis and Probability* in *Cambridge Studies in Advanced Mathematics*, Revised reprint of the 1989 original (Cambridge Univ. Press, Cambridge, 2002) Vol. 74.
9. N. Dunford and J. T. Schwartz, *Linear Operators*, Pt. I: *General Theory*, With the assistance of W. G. Bade and R. G. Bartle, Reprint of the 1958 original, in *Wiley Classics Library* (Wiley, New York, 1988), A Wiley-Interscience Publication.
10. I. Ekeland and R. Temam, *Analyse Convexe et Problèmes Variationnels* (Dunod, 1974).
11. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression* (Kluwer, Norwell, MA, USA, 1991).
12. S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, in *Lecture Notes in Mathematics* (Springer, Berlin, 2000), Vol. 1730.
13. J. A. Hartigan, *Clustering Algorithms*, in *Wiley Series in Probab. and Math. Statist.* (Wiley, New York–London–Sydney, 1975).
14. L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data, An Introduction to Cluster Analysis*, in *Wiley Series in Probab. and Math. Statist.: Appl. Probab. and Statist.*, (Wiley, New York–London–Sydney, 1990), A Wiley-Interscience Publication.
15. J. H. B. Kemperman, "The Median of a Finite Measure on a Banach Space," in *Statistical Data Analysis Based on the L_1-Norm and Related Methods (Neuchâtel, 1987)*, (North-Holland, Amsterdam, 1987), pp. 217–230.
16. M. Ledoux, *The Concentration of Measure Phenomenon*, in *Mathematical Surveys and Monographs* (AMS, 2001), Vol. 89.
17. T. Linder, "Learning-Theoretic Methods in Vector Quantization", in *Principles of Nonparametric Learning (Udine, 2001), CISM Courses and Lectures* (Springer, Vienna, 2002), Vol. 434, pp. 163–210.
18. T. Linder, G. Lugosi, and K. Zeger, "Rates of Convergence in the Source Coding Theorem, in Empirical Quantizer Design, and in Universal Lossy Source Coding", IEEE Trans. Inform. Theory **40**, 1728–1740 (1994).
19. D. Pollard, "Strong Consistency of k-Means Clustering", Ann. Statist. **9**, 135–140 (1981).
20. D. Pollard, "A Central Limit Theorem for k-Means Clustering", Ann. Probab. **10**, 919–926 (1982).
21. D. Pollard, "Quantization and the Method of k-Means", IEEE Trans. Inform. Theory **28**, 199–205 (1982).
22. J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, 2nd ed. in *Springer Series in Statistics* (Springer, New York, 2005).
23. A. W. Van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, in *Springer Series in Statistics* (Springer, New York, 1996).