Electronic Journal of Statistics
Vol. 12 (2018) 4239–4263
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1500

A notion of stability for k-means clustering*

T. Le Gouic

Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France
& National Research University Higher School of Economics, Moscow, Russia,
e-mail: thibaut.le_gouic@math.cnrs.fr

and

Q. Paris

National Research University Higher School of Economics
& International Laboratory of Stochastic Algorithms and High-Dimensional Inference
Moscow, Russia
e-mail: qparis@hse.ru

Abstract: In this paper, we define and study a new notion of stability for the k-means clustering scheme building upon the field of quantization of a probability measure. We connect this definition of stability to a geometric feature of the underlying distribution of the data, named absolute margin condition, inspired by recent works on the subject.

Keywords and phrases: Clustering, k-means, stability.

Received March 2018.

1. Introduction

Unsupervised classification consists in partitioning a data set into a series of groups (or clusters) each of which may then be regarded as a separate class of observations. This task enables practitioners in many disciplines to get a first intuition about their data by identifying meaningful groups of observations or can be used as an intermediate tool to perform data analysis (see for example Loubes and Pelletier (2017) where clustering is used to estimate conditional distributions). The tools available for unsupervised classification are various. Depending on the nature of the problem, one may rely on a strategy consisting in modeling the unknown distribution of the data as a mixture of known distributions with unknown parameters to be estimated. Another approach, model-free, is embodied by the well known k-means clustering scheme. This paper focuses on the stability of this clustering scheme with respect to the unknown distribution of the data.

1.1. Quantization and the k-means clustering scheme

The k-means clustering scheme prescribes to classify observations according to their distances to chosen representatives. This clustering scheme is strongly

*The study has been funded by the Russian Academic Excellence Project 5-100

connected to the field of quantization of probability measures and this paragraph shortly recalls how these concepts interact. Suppose our data modeled by n i.i.d. random variables $X_1, \dots, X_n$, taking their values in some metric space (E, d), and with same distribution P as (and independent of) a generic random variable X. Let $k \geq 1$ be an integer fixed in advance, representing the prescribed number of clusters, and define a k-points [1] quantizer as any mapping $q : E \to E$ such that[2] $|q(E)| = k$. Denoting $c_1, \dots, c_k$ the values taken by q, the sets $\{x \in E : q(x) = c_j\}$, $1 \leq j \leq k$, partition the space E into k subsets (or cells) and each point c_j (called indifferently a center, a centroid or a code point) stands as a representative of all points in its cell. Given a quantizer q, associated data clusters are defined, for all $1 \leq j \leq k$, by

$$C_j(q) := \{x \in E : q(x) = c_j\} \cap \{X_1, \dots, X_n\}.$$

The performance of this clustering scheme is naturally measured by the average square distance, with respect to P, of a point to its representative. In other words, the risk of q (also referred to as its distortion) is defined by

$$R(q) := \int_E d(x, q(x))^2 \, \mathrm{d}P(x). \tag{1.1}$$

Quantizers of special interest are nearest neighbor (NN) quantizers, *i.e.* quantizers such that, for all $x \in E$,

$$q(x) \in \operatorname*{arg\,min}_{c \in q(E)} d(x, c).$$

The interest for these quantizers relies on the straightforward observation that for any quantizer q, a NN quantizer q' such that $q(E) = q'(E)$ satisfies $R(q') \leq R(q)$. Hence, attention may be restricted to NN quantizers and any optimal quantizer

$$q^\star \in \operatorname*{arg\,min}_q R(q), \tag{1.2}$$

(where q ranges over all quantizers k-points quantizers) is necessarily a NN quantizer. We will denote $\mathfrak{Q}_k$ the set of all k-points NN quantizers and, unless mentioned explicitly, all quantizers involved in the sequel will be considered as members of $\mathfrak{Q}_k$. For $q \in \mathfrak{Q}_k$, the value of its risk is entirely described by its image. Indeed, if $q \in \mathfrak{Q}_k$ takes values $c_1, \dots, c_k$, then

$$R(q) = \int_E \min_{1 \leq j \leq k} d(x, c_j)^2 \, \mathrm{d}P(x). \tag{1.3}$$

Denoting $\mathbf{c} = \{c_1, \dots, c_k\}$, referred to as a codebook, we will often denote by $R(\mathbf{c})$ the right hand side of (1.3) with a slight abuse of notation.

[1] The integer k is supposed fixed throughout the paper and all quantizers considered below are supposed to be k-points quantizers.

[2] For a set A, notation $|A|$ refers to the number of elements in A.

A few additional considerations, relative to NN-quantizers, will be useful in the paper. Given $\mathbf{c} = \{c_1, \dots, c_k\}$, denote $V_j(\mathbf{c})$ the set of points in E closer to c_j than to any other c_ℓ, that is

$$V_j(\mathbf{c}) := \{\, x \in E \,:\, \forall \ell \in \{1, \dots, k\}\,,\, d(x, c_j) \leq d(x, c_\ell)\,\}\,.$$

These sets do not partition the space E since, for $i \neq j$, the set $V_i(\mathbf{c}) \cap V_j(\mathbf{c})$ is not necessarily empty. A Voronoi partition of E relative to $\mathbf{c}$ is any partition $W_1, \dots, W_k$ of E such that, for all $1 \leq j \leq k$, $W_j \subset V_j(\mathbf{c})$ up to relabeling. For instance, given $q \in \mathfrak{Q}_k$ with image $\mathbf{c}$, the sets $W_j = q^{-1}(c_j)$, $1 \leq j \leq k$, form a Voronoi partition relative to $\mathbf{c}$. We call frontier of the Voronoi diagram generated by $\mathbf{c}$ the set

$$\mathcal{F}(\mathbf{c}) := \bigcup_{i \neq j} V_i(\mathbf{c}) \cap V_j(\mathbf{c}). \tag{1.4}$$

Given an optimal quantizer $q^\star$ with image $\mathbf{c}^\star = \{c_1^\star, \dots, c_k^\star\}$, a remarkable property, known as the center condition, states that for all $1 \leq j \leq k$, and provided $|\mathrm{supp}(P)| \geq k$,

$$P(V_j(\mathbf{c}^\star)) > 0 \quad \text{and} \quad c_j^\star \in \operatorname*{arg\,min}_{c \in E} \int_{V_j(\mathbf{c}^\star)} d(x, c)^2 \, \mathrm{d}P(x). \tag{1.5}$$

From now on, the probability measure P will be supposed to have a support of more than k points.

We end this subsection by mentioning that computing an optimal quantizer requires the knowledge of the distribution P. From a statistical point of view, when the only information available about P consists in the sample $X_1, \dots, X_n$, reasonable quantizers are empirically optimal quantizers, *i.e.* NN quantizers associated to any codebook $\hat{\mathbf{c}} = \{\hat{c}_1, \dots, \hat{c}_k\}$ satisfying

$$\hat{\mathbf{c}} \in \operatorname*{arg\,min}_{\mathbf{c} = \{c_1, \dots, c_k\}} \hat{R}(\mathbf{c}) \quad \text{where} \quad \hat{R}(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} d(X_i, c_j)^2. \tag{1.6}$$

In other words, empirically optimal quantizers minimize the risk associated to the empirical measure

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

The computation of empirically optimal centers is known to be a hard problem, due in particular to the non-convexity of $\mathbf{c} \mapsto \widehat{R}(\mathbf{c})$, and is usually performed by Lloyd's algorithm. For works establishing convergence guarantees for Lloyd's algorithm and related questions, we refer the reader to Kumar and Kannan (2010); Tang and Monteleoni (2016); Lu and Zhou (2016) and Levrard (2018).

1.2. Risk bounds

The performance of the k-means clustering scheme, based on the notion of risk, has been widely studied in the literature. Whenever $(E, |.|)$ is a separable Hilbert

space, the existence of an optimal codebook, *i.e.* of $\mathbf{c}^\star = \{c_1^\star, \dots, c_k^\star\}$ such that

$$R(\mathbf{c}^\star) = R^\star = \inf_{\mathbf{c}=\{c_1,\dots,c_k\}} R(\mathbf{c}),$$

is well established (see, e.g, Theorem 4.12 in Graf and Luschgy, 2000) provided $\mathbf{E}|X|^2 < +\infty$. In this same context, works of Pollard (1981, 1982a) and Abaya and Wise (1984) imply that $R(\hat{\mathbf{c}}) \to R^\star$ almost surely as n goes to $+\infty$, where $\hat{\mathbf{c}}$ is as in (1.6). The non-asymptotic performance of the k-means clustering scheme has also received a lot of attention and has been studied, for example, by Chou (1994); Linder, Lugosi and Zeger (1994); Bartlett, Linder and Lugosi (1998); Linder (2000, 2001); Antos (2005); Antos, Györfi and György (2005) and Biau, Devroye and Lugosi (2008). For instance Biau, Devroye and Lugosi (2008) prove that in a separable Hilbert space, and provided $|X| \le L$ almost surely, then

$$\mathbf{E}R(\hat{\mathbf{c}}) - R^\star \le 12kL^2/\sqrt{n},$$

for all $n \ge 1$. A similar result is established in Cadre and Paris (2012) relaxing the hypothesis of bounded support by supposing only the existence of an exponential moment for X. In the context of a separable Hilbert space, Levrard (2015) establishes a stronger result under some conditions involving the quantity $p(t)$ defined as follows.

Definition 1.1 (Levrard, 2015). *Let $\mathcal{M}$ be the set of all $\mathbf{c}^\star = \{c_1^\star, \dots, c_k^\star\}$ such that $R(\mathbf{c}^\star) = R^\star$. For $t \ge 0$, we define*

$$p(t) := \sup_{\mathbf{c}^\star \in \mathcal{M}} P(\mathcal{F}(\mathbf{c}^\star)^t), \tag{1.7}$$

where, for any set $A \subset E$, the notation A^t stands for the t-neighborhood of A in E defined by $A^t = \{x \in E : d(x, A) \le t\}$ and where $\mathcal{F}(\mathbf{c}^\star)$ is defined in (1.4).

For any codebook $\mathbf{c} = (c_1, \dots, c_k)$, $P(\mathcal{F}(\mathbf{c})^t)$ corresponds to the probability mass of the frontier of the associated Voronoi diagram inflated by t (see Figure 1). Under some slight restrictions and supposing $p(t)$ does not increase too rapidly with t, it appears that the excess risk is of order $\mathcal{O}(1/n)$ as described below.

Theorem 1.2 (Proposition 2.1 and Theorem 3.1 in Levrard, 2015). *Suppose that $(E, |.|)$ is a (separable) Hilbert space. Denote*

$$B = \inf_{\mathbf{c}^\star \in \mathcal{M}, i \ne j} |c_i^\star - c_j^\star| \quad \text{and} \quad p_{\min} = \inf_{\mathbf{c}^\star \in \mathcal{M}, 1 \le j \le k} P(V_j(\mathbf{c}^\star)).$$

(1) *Suppose that $P(x : |x| \le L) = 1$ for some $L > 0$. Then $B > 0$ and $p_{\min} > 0$.*

(2) *Suppose in addition that there exists $r_0 > 0$ such that, for all $0 < t \le r_0$,*

$$p(t) \le \frac{B p_{\min}}{128 L^2} t, \tag{1.8}$$

where $p(t)$ is as in (1.7). Then, for all $x > 0$, and any $\hat{\mathbf{c}}$ minimizing the empirical risk as in (1.6),

$$R(\hat{\mathbf{c}}) - R^\star \le \frac{C(k+x)L^2}{n},$$

with probability at least $1 - e^{-x}$*, where* $C > 0$ *denotes a constant depending on auxiliary (and explicit) characteristics of* P*.*

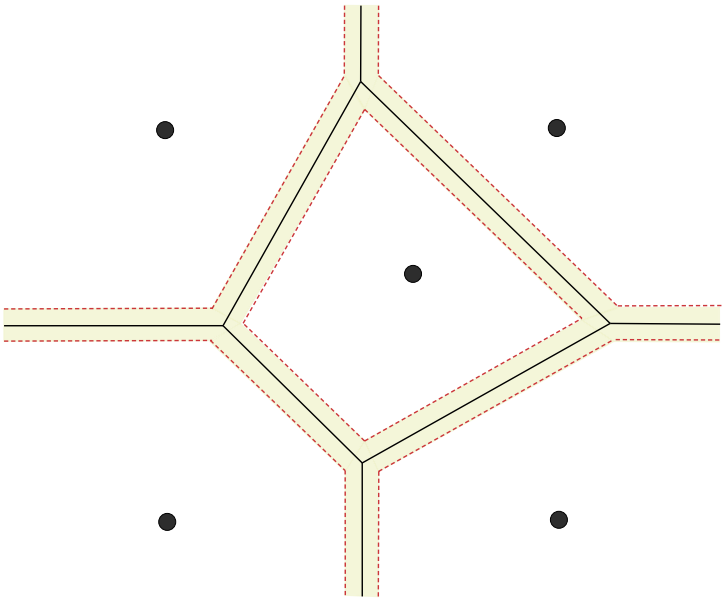

FIG 1. *For* $k = 5$*, the figure represents* k *centers in the Euclidean plane. The black solid lines define the frontier of the associated Voronoi diagram. The light-green area, inside the red dashed lines, corresponds to the* t*-neighborhood of this frontier for some small* t*.*

1.3. Stability

For a quantizer $q \in \mathfrak{Q}_k$, the risk $R(q)$ describes the average square distance of a point $x \in E$ to its representative $q(x)$ whenever x is drawn from P. The risk of q characterizes therefore an important feature of the clustering scheme based on q and defining optimality of q in terms of the value of its risk appears as a reasonable approach. However, an important though simple observation is that the excess risk $R(q) - R(q^\star)$, for an optimal quantizer $q^\star$, isn't well suited to describe the geometric similarity between the clusterings based on q and $q^\star$. For one thing, there might be several optimal codebooks. Also, even in the context where there is a unique optimal codebook, quite different configurations of centers $\mathbf{c}$ may give rise to very similar values of the excess risk $R(\mathbf{c}) - R(\mathbf{c}^\star)$. This observation relates to the difference between *estimating the optimal quantizer* and *learning to perform as well as the optimal quantizer* and is relevant in a more general context as briefly discussed in Appendix B below. The idea of stability we present next consists in identifying situations where having centers $\mathbf{c}$ with small excess risk guarantees that $\mathbf{c}$ isn't far from an optimal center $\mathbf{c}^\star$ geometrically speaking. We formalize this idea below.

Definition 1.3. *Consider functions $F : \mathfrak{Q}_k \times \mathfrak{Q}_k \to \mathbb{R}_+$ and $\phi : \mathbb{R}_+ \to \mathbb{R}_+$. The clustering problem discussed in subsections 1.1 and 1.2 is called (F, ϕ)-stable if, for any optimal quantizer $q^\star$ and any auxiliary quantizer q,*

$$F(q^\star, q) \leq \phi(R(q) - R(q^\star)). \tag{1.9}$$

We say that the clustering problem is strongly *stable for F if ϕ is linear.*

Note first that, for some chosen F, the notion of stability defined above characterizes a property of the underlying distribution P. Here, properties of the function F are deliberately unspecified as, in practice, F can be chosen in order to encode very different properties, of more or less geometric nature. An important property of this notion is that stable clustering problems are such that ε-minimizers of the risk are "close" (in the sense of F) to an optimal quantizer (see Corollary 2.5 below). Note also that if $\phi(0) = 0$ and if F is a metric on $\mathfrak{Q}_k$, then (F, ϕ)-stability imposes uniqueness of the optimal quantizer $q^\star$.

Remark 1.4. *The notion of stability described above differs from the notion of algorithm stability studied in Ben-David, Von Luxburg and Pál (2006) and Ben-David, Pál and Simon (2007). Their notion of stability is defined for a function (called algorithm) $A : \bigcup_n E^n \to \mathfrak{Q}_k$ that maps any data set $\{X_1, \dots, X_n\}$ to a quantizer $A(\{X_1, \dots, X_n\})$. In this context, the stability of A is defined by*

$$\mathrm{Stab}(A, P) = \lim_{n \to \infty} \mathbf{E} D(A(\{X_1, \dots, X_n\}), A(\{Y_1, \dots, Y_n\})),$$

where the X_i's and Y_i's are i.i.d. random variables of common distribution P and D is a (pseudo-) metric on $\mathfrak{Q}_k$. Then, an algorithm is said to be stable for P if $\mathrm{Stab}(A, P) = 0$. According to this definition, any constant algorithm $A = q$ is stable. A notable difference, is that our notion of stability includes a notion of consistency. Indeed, since $q \mapsto R(q)$ is continuous (for a proper choice of the metric on $\mathfrak{Q}_k$), our notion of stability measures the rate at which $q \to q^\star$ whenever $R(q) \to R^\star$. Thus, we focus only on the behaviour of algorithms A such that $R(A(\{X_1, \dots, X_n\})) \to R^\star$.

A first rather obvious choice for F is given by

$$F_1(q^\star, q) := \min_\sigma \max_{1 \leq j \leq k} d(c_j^\star, c_{\sigma(j)}), \tag{1.10}$$

where $q(E) = \{c_1, \dots, c_k\}$, $q^\star(E) = \{c_1^\star, \dots, c_k^\star\}$ and where the minimum is taken over all permutations σ of $\{1, \dots, k\}$ (see Figure 3 for a graphical interpretation).

Remark 1.5. *Note that $F_1(q^\star, q)$ does not always coincide with the Hausdorff distance $d_H(\mathbf{c}^\star, \mathbf{c})$ between $\mathbf{c} = \{c_1, \dots, c_k\}$ and $\mathbf{c}^\star = \{c_1^\star, \dots, c_k^\star\}$. Indeed, Figure 2 presents a configuration of codebooks $\mathbf{c}$ and $\mathbf{c}^\star$ that have small Hausdorff distance but define NN quantizers q and $q^\star$ with large $F_1(q^\star, q)$. However, it may be seen that inequality*

$$d_H(\mathbf{c}^\star, \mathbf{c}) \leq F_1(q^\star, q)$$

always holds and that, provided

$$d_H(\mathbf{c}^\star, \mathbf{c}) < \frac{1}{2} \min_{i \neq j} |c_i^\star - c_j^\star|,$$

we obtain $d_H(\mathbf{c}^\star, \mathbf{c}) = F_1(q^\star, q)$. *The proof of these statements is reported in Appendix A.1.*

FIG 2. *In this simple case, where k=3, the set of black dots and the set of white dots have small Hausdorff distance but define two NN quantizers, say* q_1 *and* q_2, *for which* $F_1(q_1, q_2)$ *is large.*

Whenever $(E, |.|)$ is Euclidean, it follows from the previous remark and Pollard (1982b) that, provided the optimal codebook $\mathbf{c}^\star$ is unique,

$$F_1(q^\star, \hat{q}) \underset{n \to +\infty}{\longrightarrow} 0, \quad \text{a.s.},$$

when $\hat{q}$ is any quantizer minimizing the empirical risk $\hat{R}$. In Levrard (2015), under the conditions of Theorem 1.2, it is proven that for any optimal quantizer $q^\star$, and any $q \in \mathfrak{Q}_k$ such that $q(E) \subset \{x : |x| \leq L\}$,

$$F_1(q^\star, q)^2 \leq \frac{p_{\min}}{2}(R(q) - R(q^\star)),$$

provided $F_1(q^\star, q) \leq Br_0/4\sqrt{2}L$ which proves in this case (a local version of) the stability of the clustering scheme for F_1 (constants are defined in Theorem 1.2). In the same spirit, when $E = \mathbf{R}^d$ and for a measure P with bounded support, Rakhlin and Caponnetto (2007) show that $F_1(q_n, q_n') \to 0$ as $n \to \infty$ whenever q_n and q_n' are optimal quantizers for empirical measures P_n and P_n' whose supports differ by at most $o(\sqrt{n})$ points. In addition, their Lemma 5.1 shows that, for P with bounded support,

$$d_H(\mathbf{c}^\star, \mathbf{c}) \leq C\, \mathbf{E}[||X - q(X)|^2 - |X - q^\star(X)|^2|]^{\frac{1}{d+2}},$$

for some constant $C > 0$. Note that, since $\mathbf{E}[||X - q(X)|^2 - |X - q^\star(X)|^2|] \geq R(q) - R(q^\star)$, our main result (Theorem 2.3) improves this inequality under suitable conditions discussed below.

While F_1 captures distances between the images of the two quantizers, it is however totally oblivious to the amount of wrongly classified points. From this point of view, a more interesting quantity is described by

$$F_2(q^\star, q) := \min_\sigma P\left[\left(\bigcup_{j=1}^k V_j(\mathbf{c}^\star) \cap V_{\sigma(j)}(\mathbf{c})\right)^c\right], \tag{1.11}$$

where the minimum is taken over all permutations σ of $\{1,\dots,k\}$ (see Figure 3). This quantity measures exactly the amount of points that are misclassified by q compared to $q^\star$, regarding P.

In the present paper, we study a related quantity, of geometric nature, defined simply as the average square distance between a quantizer q and an optimal quantizer $q^\star$, *i.e.*

$$\mathbf{F}(q^\star,q)^2 := \int_E d(q(x),q^\star(x))^2 \,\mathrm{d}P(x). \tag{1.12}$$

As discussed later in the paper (see Subsection 2.2), this quantity may be seen as an intermediate between F_1 and F_2 incorporating both the notion of proximity of the centers and the amount of misclassified points. The general concern of the paper will be to establish conditions under which the clustering scheme is strongly stable for this function $\mathbf{F}^2$.

To conclude this section, it is worth mentioning that if $q^\star$ is an optimal k-points quantizer for P, identity $\mathbf{F}(q^\star,q)^2 = 0$ imposes that $q = q^\star$ P-a.s. and that $q(E) = q^\star(E)$ (according to the center condition (1.5)). In particular, uniqueness of the optimal codebook $\mathbf{c}^\star = q^\star(E)$ for P is necessary for $(\mathbf{F},\phi)$-stability to hold (in the sense of Definition 1.3) when $\phi(0) = 0$. Hence, this uniqueness assumption will be made throughout the rest of paper and does not restrict generality.

2. Stability results

In this section, we present our main results. In the sequel, we restrict ourselves to the case where E is a (separable) Hilbert space with scalar product $\langle .,. \rangle$ and associated norm $|.|$. For any E-valued random variable Z, we'll denote

$$\|Z\|^2 := \mathbf{E}|Z|^2,$$

for brevity.

2.1. Absolute margin condition

We first address the issue of characterizing the stability of the clustering scheme in terms of the function $\mathbf{F}$ defined in (1.12). The next definition plays a central role in our main result. Recall that X denotes a generic random variable with distribution P.

Definition 2.1 (Absolute margin condition). *Suppose that $\int |x|^2 \mathrm{d}P(x) < +\infty$ and that P has a unique (P-a.s.) optimal k-points quantizer $q^\star$. For $\lambda \geq 0$, define*

$$A(\lambda) = \{x \in E : q^\star(x + \lambda(x - q^\star(x))) = q^\star(x)\}.$$

Then, P is said to satisfy the absolute margin condition with parameter $\lambda_0 > 0$, *if both the following conditions hold:*

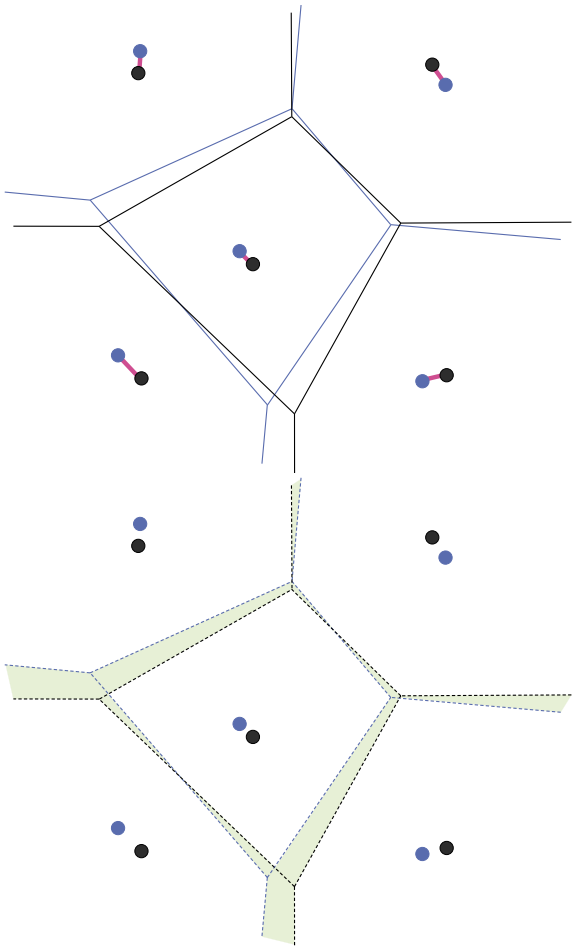

FIG 3. *The image of $q^\star$ (resp. q) is represented by the black (resp. blue dots). The quantity $F_1(q^\star, q)$ corresponds to the length of the longest pink segment in the first (left) figure. The quantity $F_2(q^\star, q)$ is the P measure of the light green area in the second (right) figure.*

1. $P(A(\lambda_0)) = 1$.
2. *For any $0 \leq \lambda < \lambda_0$, the law of*

$$X_\lambda := X + \lambda(X - q^\star(X))$$

has a unique optimal k-quantizer $q_\lambda \in \mathfrak{Q}_k$.

The second condition means that every probability measure, in a neighborhood of P (including of course P), has a unique optimal k-quantizer. Note that $A(0) = E$ and that $A(\lambda) \subset A(\lambda')$ for $\lambda' \leq \lambda$. Letting $\mathbf{c}^\star = q^\star(E)$, the first point of this definition states that the neighborhood $E \setminus A(\lambda_0)$ of the frontier $\mathcal{F}(\mathbf{c}^\star)$ is of probability zero (see Figure 4). The next remark discusses the geometry

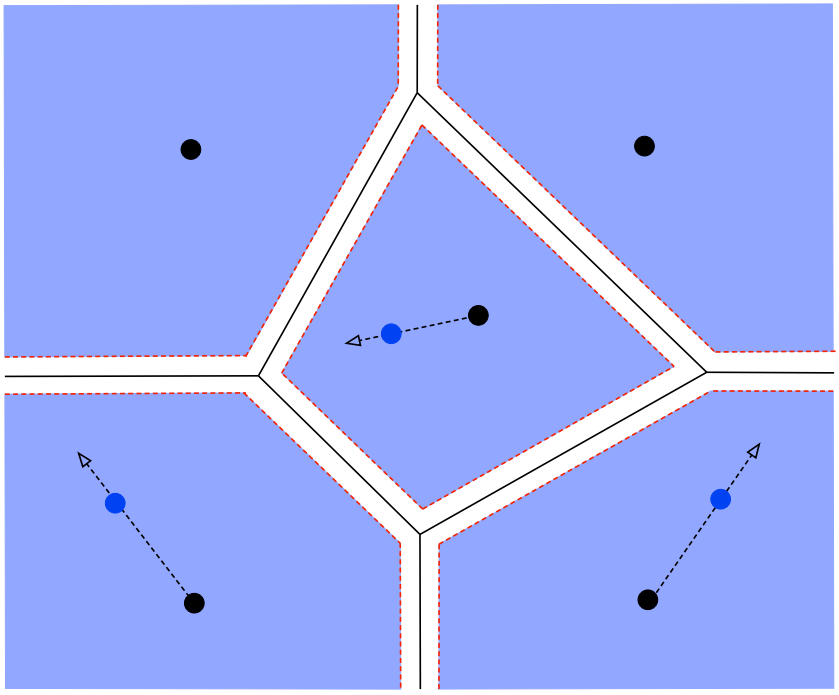

FIG 4. *The figure represents an optimal codebook (the black dots) for a distribution P, the frontier of the associated Voronoi diagram (the solid black line) and, for a fixed value of $\lambda > 0$, the set $A(\lambda)$ (the light blue area). For points x in $A(\lambda)$ (blue dots), the figure represents the associated point $x + \lambda(x - q^\star(x))$ (tip of the arrows) which, by definition of $A(\lambda)$, belongs to $A(\lambda)$.*

of the set $A(\lambda)$, involved in the previous definition, in comparison with the sets $\mathcal{F}(\mathbf{c}^\star)^t$ used in Definition 1.1. In particular, it follows from the next remark that, for appropriate $0 < t_1 < t_2$, the set $E \setminus A(\lambda)$ satisfies

$$\mathcal{F}(\mathbf{c}^\star)^{t_1} \subset (E \setminus A(\lambda_0)) \subset \mathcal{F}(\mathbf{c}^\star)^{t_2}.$$

Remark 2.2. *Let $\mathbf{c} = \{c_1, \ldots, c_k\} \subset E$ and let $q \in \mathfrak{Q}_k$ be the NN quantizer with codebook $\mathbf{c}$. Denote*

$$m(\mathbf{c}) = \min_{i \neq j} |c_i - c_j| \quad \text{and} \quad M(\mathbf{c}) = \max_{i \neq j} |c_i - c_j|.$$

For all $\lambda \geq 0$ and $t > 0$, let

$$A(\lambda) := \{x \in E : q(x + \lambda(x - q(x))) = q(x)\} \quad \text{and} \quad B(t) := E \setminus \mathcal{F}(\mathbf{c})^t.$$

Then the following statements hold.

1. *For all $0 < t < M(\mathbf{c})/2$,*

$$B(t) \subset A\left(\frac{2t}{M(\mathbf{c}) - 2t}\right).$$

2. *For all $\lambda > 0$*

$$A(\lambda) \subset B\left(\frac{m(\mathbf{c})\lambda}{2(1+\lambda)}\right).$$

It follows from this observation that, when P has bounded support, the absolute margin condition introduced in Definition 2.1 is essentially more restrictive than the margin condition (1.8). Note indeed that the absolute margin condition could also be stated in terms of the function $p(t)$, as defined in Levrard (2015), in the form of "$p(t) = 0$ for $t \leq t_0$ and some $t_0 > 0$" (this would involve the smallest and largest distance between optimal centroids).

We are now in position to state the main result of this paper.

Theorem 2.3. *Suppose that $\int |x|^2 \mathrm{d}P(x) < +\infty$ and that P has a unique (P-a.s.) optimal k-points quantizer $q^\star$. Suppose that P satisfies the absolute margin condition 2.1 with parameter $\lambda_0 > 0$. Then, for any $q \in \mathfrak{Q}_k$, it holds that*

$$\mathbf{F}(q^\star, q)^2 \leq \frac{1 + \lambda_0}{\lambda_0}(R(q) - R(q^\star)).$$

Remark 2.4. *The above theorem states that the clustering scheme is strongly stable for $\mathbf{F}^2$ provided the absolute margin condition holds. Here, we briefly argue that this result is optimal in the sense that strong stability requires that both hypotheses of the absolute margin condition 2.1 hold in general.*

1. *The following example shows that the first point of the absolute margin condition cannot be dropped. Take P uniform on $[-1, 1] \times [-1/2, 1/2]$ and*

fix $k = 2$. Then the first point of the absolute margin condition is clearly not satisfied. The codebook

$$\mathbf{c}^\star = \{(-1/2, 0), (1/2, 0)\}$$

defines the unique optimal quantizer. For $\varepsilon > 0$, consider now

$$\mathbf{c}_\varepsilon = \{(-1/2, \varepsilon), (1/2, -\varepsilon)\}.$$

Then it can be checked through straightforward computations that $\mathbf{F}(q^\star, q_\varepsilon) = \varepsilon$ and that $R(q_\varepsilon) - R(q^\star) \leq \varepsilon^2$, so that there exists no $\lambda > 0$ for which inequality

$$\mathbf{F}(q^\star, q_\varepsilon)^2 \leq \frac{1+\lambda}{\lambda}(R(q_\varepsilon) - R(q^\star))$$

holds for all $\varepsilon > 0$.

2. *If there is no unique optimal quantizer for P, then the result clearly cannot hold as mentioned at the end of the previous section. However, this uniqueness property is not enough. To illustrate this statement, suppose $P = (\mu_1 + \mu_2)/2$ where μ_1 is uniform on $[-1; 1] \times \{1\}$ and μ_2 is uniform on $[-1; 1] \times \{-1\}$. For $k = 2$, the codebook*

$$\mathbf{c}^\star = \{(0, 1), (0, -1)\}$$

defines the unique optimal quantizer for P. The distribution P satisfies the first point of the absolute margin condition for any $\lambda > 0$, but fails to satisfy the second point for large λ. It can be understood from details in the proof of Theorem 2.3 that the desired inequality cannot hold for large λ. Therefore, the second point of the absolute margin condition can not be simply dropped either.

An interesting consequence of Theorem 2.3 holds in the context of empirical measures for which the absolute margin condition always holds. Consider a sample $X_1, \ldots, X_n$ composed of i.i.d. variables with distribution P and let

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

The next result ensures that an ε-empirical risk minimizer (i.e. a quantizer q_ε such that $R_n(q_\varepsilon) \leq \inf_q R_n(q) + \varepsilon$) is at a distance (in terms of $\mathbf{F}$) at most $\varepsilon(1+\lambda)/\lambda$ to an empirical risk minimizer for some λ depending only on P_n.

Corollary 2.5. *Let P_n be the empirical measure of a measure P, associated with sample $\{X_1, \ldots, X_n\}$. Then, P_n has a* unique *optimal quantizer $\hat{q}$ if and only if there exists some $\lambda_n > 0$ such that*

$$\mathbf{F}(\hat{q}, q)^2 \leq \frac{1+\lambda_n}{\lambda_n}\left(\hat{R}(q) - \hat{R}(\hat{q})\right).$$

Remark that $\mathbf{F}$ here, relates to the measure P_n and not P. And the result actually holds for any finitely supported measure P_n.

The proof is quite straightforward from Theorem 2.3, and is postponed in Section 3. The interpretation of this corollary is that any algorithm producing a quantizer q with small empirical risk $\hat{R}(q)$ will be, automatically, such that $\mathbf{F}(\hat{q}, q)$ is small if λ_n is large. The parameter λ_n of the absolute margin condition thus provides a key feature for stability of the k-means algorithm. A nice property of the previous result is that λ_n is of course independent of the ε-minimizer q_ε. However, an important remaining problem, of large practical value, would be to provide a lower bound for λ_n (valid with high probability) to assess the size of the coefficient $(1 + \lambda_n)/\lambda_n$. This is left for future research.

2.2. Comparing notions of stability

This subsection describes some relationships existing between the function $\mathbf{F}$ involved in our main result, with the two functions F_1 and F_2 mentioned earlier in section 1.3. Below, we restrict attention to the case where there is a unique optimal quantizer $q^\star$. Comparing $\mathbf{F}$ and F_2 can be done straightforwardly. Let

$$m = \inf_{i \neq j} |c_i^\star - c_j^\star| \quad \text{and} \quad M = \sup_{i \neq j} |c_i^\star - c_j^\star|.$$

Observe that, for $F_1(q^\star, q)$ small enough, the permutation reaching the minimum in the definitions of F_1 and F_2 is the same and can be assumed to be the identity without loss of generality. Then, it follows that, for $F_1(q^\star, q)$ small enough,

$$\begin{aligned}
\mathbf{F}(q^\star, q)^2 &= \sum_{i,j=1}^{k} P(V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c}))|c_i^\star - c_j|^2 \\
&\leq \sum_{i=1}^{k} P(V_i(\mathbf{c}^\star) \cap V_i(\mathbf{c}))|c_i^\star - c_i|^2 \\
&\quad + \sum_{i \neq j} P(V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c}))(|c_j^\star - c_j| + M)^2 \\
&\leq F_1(q^\star, q)^2 + F_2(q^\star, q)(F_1(q^\star, q) + M)^2,
\end{aligned}$$

and similarly, when $m \geq F_1(q^\star, q)$,

$$\begin{aligned}
\mathbf{F}(q^\star, q)^2 &\geq \sum_{i=1}^{k} P(V_i(\mathbf{c}^\star) \cap V_i(\mathbf{c}))|c_i^\star - c_i|^2 \\
&\quad + \sum_{i \neq j=1}^{k} P(V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c}))(m - |c_j^\star - c_j|)^2 \\
&\geq F_2(q^\star, q)(m - F_1(q^\star, q))^2.
\end{aligned}$$

These two inequalities imply that $\mathbf{F}^2$ and F_2 are comparable whenever F_1 is small enough.

Comparing F_1 and $\mathbf{F}$ requires more effort, although one inequality is also quite straightforward. Recall the notation $p_{\min} = \inf_i P(V_i(\mathbf{c}^\star))$. Suppose again that the optimal permutation in the definition of F_1 is the identity. Then, remark that $F_1(q^\star, q) \leq m/2$, implies $|c_i^\star - c_i| \leq |c_i^\star - c_j|$, for all i, j. Thus, in this case,

$$\begin{aligned}\mathbf{F}(q^\star, q)^2 &= \mathbf{E}|q^\star(X) - q(X)|^2 \\ &= \sum_{i,j=1}^{k} P(V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c}))|c_i^\star - c_j|^2 \\ &\geq \sum_{i=1}^{k}\sum_{j=1}^{k} P(V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c}))|c_i^\star - c_i|^2 \\ &\geq p_{\min} F_1(q^\star, q)^2.\end{aligned}$$

In view of providing a more detailed result, we define the function $p^\star$, similar in nature to the function p introduced by Levrard (2015) and defined in 1.1.

Definition 2.6. *For a metric space (E, d) and a probability measure P on E, let X be a random variable of distribution P. Denote $q^\star$ an optimal quantizer of P with image $\mathbf{c}^\star = \{c_1^\star, \ldots, c_k^\star\}$ and $\partial V_i(\mathbf{c}^\star)$ the frontier of the Voronoi cell associated to $c_i^\star$. Then, for all $t > 0$, we let*

$$p^\star(t) := \mathbf{P}\left(\bigcup_{i=1}^{k} \left\{md(X, \partial V_i(\mathbf{c}^\star)) \leq 2d(X, q^\star(X))t + 2t^2\right\}\right),$$

where $m = \inf_{i \neq j} |c_i^\star - c_j^\star|$.

While $p(t)$ corresponds to the probability of the t-inflated frontier of the Voronoi cells (defined in Definition 1.1), $p^\star(t)$ corresponds to a similar object in which the inflation of the frontier gets larger as the points go further from their representative in the codebook $\mathbf{c}^\star$. These two functions can thus differ significantly, in general. However, since $m/4 \leq d(X, q^\star(X))$ for X such that $d(X, \partial V_i(\mathbf{c}^\star)) < m/4$, it follows that

$$p(t) \leq p^\star(2t),$$

whenever $0 < t < m/4$. And when the probability measure P has its support in a ball of diameter $R > 0$, it can be readily seen that for all $t > 0$

$$p^\star(t) \leq p\left(m^{-1}\left[2Rt + 2t^2\right]\right).$$

If the support of P is not contained in a ball, the comparison is not as straightforward.

We can now state the last comparison inequality.

Proposition 2.7. *In the setting of Definition 2.6,*

$$\mathbf{F}(q^\star, q)^2 \leq F_1(q^\star, q)^2 + p^\star(F_1(q^\star, q))(M + F_1(q^\star, q))^2.$$

A consequence of this proposition, and Theorem 1.2, is the following.

Corollary 2.8. *Under the conditions of Theorem 1.2, with high probability,*

$$\mathbf{F}(q^\star, \hat{q})^2 = \mathcal{O}\left(\frac{1}{n}\right) + p^\star\left(\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right),$$

for any empirical risk minimizer $\hat{q}$.

Note that, if P has a bounded support, under the absolute margin condition, the term $p^\star\left(\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right)$ is 0, for n large enough and the result becomes

$$\mathbf{F}(q^\star, \hat{q})^2 = \mathcal{O}\left(\frac{1}{n}\right).$$

3. Proofs

This section gathers the proofs of the main results of the paper. Additional proofs are postponed to the appendices.

3.1. Proof of Theorem 2.3

Recall that E is a Hilbert space with scalar product $\langle ., . \rangle$, norm $|.|$ and that, for an E-valued random variable Z with square integrable norm, we denote $\|Z\|^2 = \mathbf{E}|Z|^2$ for brevity. For $\lambda > 0$, set

$$x_\lambda = x + \lambda(x - q^\star(x)).$$

As E is a Hilbert space, we have for all $y, z \in E$ and all $t \in [0, 1]$,

$$|ty + (1-t)z|^2 = t|y|^2 + (1-t)|z|^2 + t(1-t)|y-z|^2.$$

Now for all $x \in E$, any quantizer $q \in \mathfrak{Q}_k$ and any $\lambda > 0$, using the previous inequality with $y = x_\lambda - q(x)$, $z = q^\star(x) - q(x)$ and $t = (1+\lambda)^{-1}$, it follows that

$$\begin{aligned}
|q^\star(x) - q(x)|^2 &= \frac{1+\lambda}{\lambda}(|x - q(x)|^2 - |x - q^\star(x)|^2) \\
&\quad + \frac{|x_\lambda - q^\star(x)|^2 - |x_\lambda - q(x)|^2}{\lambda} \\
&\leq \frac{1+\lambda}{\lambda}(|x - q(x)|^2 - |x - q^\star(x)|^2) \\
&\quad + \frac{|x_\lambda - q^\star(x)|^2 - |x_\lambda - q(x_\lambda)|^2}{\lambda},
\end{aligned}$$

where the last inequality follows from the fact that q is a nearest neighbor quantizer. Integrating this inequality with respect to P, we obtain

$$\mathbf{F}(q^\star, q)^2 \leq \frac{1+\lambda}{\lambda}(R(q) - R(q^\star)) + \frac{1}{\lambda}c_q(\lambda), \tag{3.1}$$

where we have denoted

$$c_q(\lambda) := \|X_\lambda - q^\star(X)\|^2 - \|X_\lambda - q(X_\lambda)\|^2.$$

Observe that $\lambda \mapsto c_q(\lambda)$ is continuous. Now, define

$$c_\infty(\lambda) := \sup_q c_q(\lambda),$$

where the supremum is taken over all k-points quantizers $q \in \mathfrak{Q}_k$. The function $\lambda \mapsto c_\infty(\lambda)$ satisfies obviously $c_\infty(\lambda) \geq c_{q^\star}(\lambda) \geq 0$, for all $\lambda > 0$. To prove the theorem, we will show that $c_\infty(\lambda_0) \leq 0$, whenever P satisfies the absolute margin condition with parameter $\lambda_0 > 0$. To that aim, we provide two auxiliary results.

Lemma 3.1. *Suppose there exists $R > 0$ such that $P(B(0,R)) = 1$. For all $\lambda > 0$, denote q_λ any quantizer such that $c_{q_\lambda}(\lambda) = c_\infty(\lambda)$ and denote $q^\star$ an optimal quantizer of the law of X. Suppose the first point of absolute margin condition holds for $\lambda_0 > 0$. Then, for all $0 < \lambda_1 < \lambda_0$, there exists $\varepsilon > 0$ such that for all $0 < \lambda < \lambda_1$, if $F_1(q_\lambda, q^\star) < \varepsilon$, then*

$$q^\star = q_\lambda.$$

Proof of lemma 3.1. The main idea of the proof is that since the Voronoi cells are well separated (inflated borders have probability 0), when a quantizer is close enough to the optimal one, it shares its Voronoi cell (on the support of P) and thus, centroid condition requires that quantizer have to be centroid of its cell to be optimal.

Set $q_\lambda(E) = \{c_1, ..., c_k\}$ and $\{c_1^\star, \dots, c_k^\star\} = q^\star(E)$. Suppose without loss of generality that the optimal permutation in the definition of F_1 is the identity. The absolute margin condition implies that, with probability one, for each $1 \leq i \leq k$, on the event $q^\star(X) = c_i^\star$, the inequality $|X_{\lambda_0} - c_i^\star|^2 \leq |X_{\lambda_0} - c_j^\star|^2$ holds, or equivalently

$$2(1+\lambda_0)\langle X - c_i^\star, c_j^\star - c_i^\star\rangle \leq |c_i^\star - c_j^\star|^2. \tag{3.2}$$

However,

$$\begin{aligned}
|X_{\lambda_0} - c_i|^2 =& (1+\lambda_0)^2|X - c_i^\star|^2 + |c_i^\star - c_i|^2 + 2(1+\lambda_0)\langle X - c_i^\star, c_i^\star - c_i\rangle \\
|X_{\lambda_0} - c_j|^2 =& (1+\lambda_0)^2|X - c_i^\star|^2 + |c_i^\star - c_j|^2 + 2(1+\lambda_0)\langle X - c_i^\star, c_i^\star - c_j\rangle
\end{aligned}$$

so that $|X_{\lambda_0} - c_i|^2 \leq |X_{\lambda_0} - c_j|^2$ if

$$2(1+\lambda_0)\langle X - c_i^\star, c_j - c_i\rangle \leq |c_i^\star - c_j|^2 - |c_i^\star - c_i|^2.$$

Since (3.2) holds, for all $\lambda_1 < \lambda_0$, there exists therefore $\varepsilon = \varepsilon(\lambda_0, \lambda_1, R, \max\{|c_i^\star - c_j^\star| : i \neq j\})$ such that, if $F_1(q^\star, q) < \varepsilon$, then for all $\lambda \leq \lambda_1$,

$$|X_\lambda - c_i|^2 < |X_\lambda - c_j|^2,$$

on the event $q^\star(X) = c_i^\star$. As a result,

$$\mathbf{P}\left(\bigcup_{i=1}^{k}\{q^\star(X) = c_i^\star\} \cap \{q_\lambda(X_\lambda) = c_i\}\right) = 1.$$

This means that $q^\star$ and q_λ share the same cells on the support of P. Thus,

$$\begin{aligned}\|X_\lambda - q_\lambda(X_\lambda)\|^2 &= (1+\lambda)^2 \sum_{i=1}^{k} \mathbf{E1}_{\{q^\star(X)=c_i^\star\}} \left|X - \frac{\lambda c_i^\star + c_i}{1+\lambda}\right|^2 \\ &\geq (1+\lambda)^2 \sum_{i=1}^{k} \mathbf{E1}_{\{q^\star(X)=c_i^\star\}} |X - c_i^\star|^2 \qquad (3.3) \\ &= \|X_\lambda - q^\star(X)\|^2,\end{aligned}$$

where inequality (3.3) follows from the center condition (1.5). Therefore, since q_λ minimizes $\|X_\lambda - q(X_\lambda)\|^2$ amongst NN quantizers, (3.3) is an equality. Therefore, $c_i = c_i^\star$ i.e. $q^\star = q_\lambda$, since optimal centroids $(c_i^\star)_i$ and $(c_i)_i$ are minimizers of $a \mapsto \mathbf{E1}_{q^\star(X)=c_i^\star}|X-a|^2$ - it is the centroid condition (Theorem 4.1 of Graf and Luschgy, 2000). □

Lemma 3.2. *Suppose X satisfies the conditions of Lemma 3.1. Denote*

$$\lambda^- = \min\{\lambda : c_\infty(\lambda) > 0\}.$$

Then $\lambda^- \geq \lambda_0$.

Proof of lemma 3.2. The idea of the proof of this lemma is that uniqueness condition of the margin condition implies continuity of the optimal quantizer with respect to λ and previous lemma states that the only optimal quantizers for X_λ that is close to $q^\star$ is $q^\star$.

Suppose $\lambda^- < \lambda_0$ in order to prove a contradiction. Then, by second point of the margin condition (Definition 2.1), there exists only one quantizer q_{λ^-} such that $c_\infty(\lambda^-) = c_{q_{\lambda^-}}(\lambda^-)$ - it is the optimal quantizer of the law of X_{λ^-}. By definition of λ^-, $c_{q_{\lambda^-}}(\lambda) \leq 0$ for $\lambda < \lambda^-$. Therefore, by continuity of $\lambda \mapsto c_{q_{\lambda^-}}(\lambda)$,

$$c_{q_{\lambda^-}}(\lambda^-) = 0,$$

and thus; by the second point of absolute margin condition, $q_{\lambda^-} = q^\star$, since $c_{q^\star}(\lambda^-) = 0$. Now, for all $\lambda > \lambda^-$, denote by q_λ any quantizer such that $c_{q_\lambda}(\lambda) = c_\infty(\lambda)$, which exists by Lemma A.1. Then by Lemma A.1 again, $F_1(q_\lambda, q_{\lambda^-}) \to 0$ as $\lambda \to \lambda^-$, so that for all $\lambda - \lambda^- > 0$ small enough, Lemma 3.1 applies and states $q_\lambda = q_{\lambda^-} = q^\star$; which contradicts the definition of λ^-. □

Therefore, $c_\infty(\lambda_0) = 0$, and thus (3.1) gives

$$\mathbf{F}(q^\star, q)^2 \leq \frac{1+\lambda_0}{\lambda_0}(R(q) - R(q^\star)).$$

Finally, by a continuity argument, the result still holds without the assumption of boundedness $\mathbf{P}(|X| \leq R) = 1$.

Indeed, it is straightforward to check that, given a random variable Y of distribution Q and optimal k-quantizer q_Q, the distribution $Q^{-\lambda}$ of

$$Y^{-\lambda} := Y - \frac{\lambda}{1+\lambda}(Y - q_Q(Y)),$$

satisfies the absolute margin condition with parameter λ. Then, for large $r > 0$, denote P_r^λ the distribution of

$$X_\lambda = X + \lambda(X - q^\star(X)),$$

conditioned to $X_\lambda \in B(0, r)$. Thus, $\left(P_r^\lambda\right)^{-\lambda}$ satisfies the absolute margin condition with parameter λ. Denote $q_{r,\lambda}$ the optimal quantizer of $\left(P_r^\lambda\right)^{-\lambda}$ – which is the same as the optimal quantizer of P_r^λ. Hence, we have proved

$$\mathbf{F}(q_{r,\lambda}, q)^2 \leq \frac{1+\lambda}{\lambda}(R_r(q) - R_r(q_{r,\lambda})), \tag{3.4}$$

where R_r denotes the risk with respect to $\left(P_r^\lambda\right)^{-\lambda}$. Clearly, for $\lambda < \lambda_0$, $P_r^\lambda \to P^\lambda$, in the topology of the Wasserstein distance, as $r \to \infty$. Therefore, for $\lambda < \lambda_0$, Theorem 9 of Pollard (1982b) ensures that $q_{r,\lambda} \to q_\star$ as $r \to \infty$. Hence, in Wasserstein topology, for $\lambda < \lambda_0$, $\left(P_r^\lambda\right)^{-\lambda} \to P$, as $r \to \infty$. Thus, for all $\lambda < \lambda_0$, as $r \to \infty$, (3.4) becomes

$$\mathbf{F}(q^\star, q)^2 \leq \frac{1+\lambda}{\lambda}(R(q) - R(q^\star)).$$

The proof ends letting $\lambda \to \lambda_0$.

3.2. *Proof of Corollary 2.5*

First, remark that it is straightforward that inequality

$$\mathbf{F}(\hat{q}, q)^2 \leq \frac{1+\lambda_n}{\lambda_n}\left(\hat{R}(q) - \hat{R}(\hat{q})\right),$$

for any quantizer q, imply that minimizer $\hat{q}$ of $\hat{R}$ is unique.

Then, using Theorem 2.3, it just remains to show that P_n satisfies the absolute margin condition. The first point of the absolute margin condition follows easily from Theorem 4.2 in Graf and Luschgy (2000) (stating that $P_n(\mathcal{F}(\hat{\mathbf{c}})) = 0$, for $\hat{\mathbf{c}} = \hat{q}(E)$, and thus $P_n(A(\lambda)) = 1$ for some $\lambda > 0$). For a measure Q denote Q_λ the distribution of

$$Y_\lambda = Y + \lambda(Y - q_Q(Y)),$$

where Y has distribution Q and q_Q stands for an optimal quantizer of Q. Then, denoting q_λ an optimal quantizer of P_n^λ, letting $\lambda \to 0$ implies that $q_\lambda \to q^\star$ (by Theorem 9 of Pollard, 1982b), so that Lemma 3.1 ensures that $q_\lambda = q^\star$ for λ small enough. Hence q_λ is unique for λ small enough (otherwise, one would have to extract a subsequence of q_λ not converging to $q^\star$, which contradicts Theorem 9 of Pollard, 1982b).

3.3. Proof of Proposition 2.7

The following proof borrows some arguments from the proof of Lemma 4.2 of Levrard (2015). Recall that $m = \inf_{i\neq j} |c_i^\star - c_j^\star|$. Take $1 \leq i, j \leq k$ and consider the hyperplane

$$h_{i,j}^\star := \{x \in E : |x - c_i^\star| = |x - c_j^\star|\}.$$

Then, for all $x \in V_i(\mathbf{c}^\star)$,

$$\begin{aligned} d(x, h_{i,j}^\star) &= \frac{|\langle c_i^\star + c_j^\star - 2x, c_i^\star - c_j^\star\rangle|}{2|c_i^\star - c_j^\star|} \\ &\leq \frac{|\langle c_i^\star + c_j^\star - 2x, c_i^\star - c_j^\star\rangle|}{2m} \\ &= \frac{|x - c_j^\star|^2 - |x - c_i^\star|^2}{2m}. \end{aligned} \tag{3.5}$$

Without loss of generality, suppose now for simplicity that the permutation σ achieving the minimum in the definition of $F_1(q^\star, q)$ is the identity, $\sigma(j) = j$, so that

$$F_1(q^\star, q) = \max_i |c_i^\star - c_i|.$$

Then, it follows that for $x \in V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c})$,

$$\begin{aligned} |x - c_j^\star|^2 - |x - c_i^\star|^2 &\leq (|x - c_j| + |c_j - c_j^\star|)^2 - |x - c_i^\star|^2 \\ &\leq (|x - c_i| + |c_j - c_j^\star|)^2 - |x - c_i^\star|^2 \\ &\leq (|x - c_i^\star| + |c_i - c_i^\star| + |c_j - c_j^\star|)^2 - |x - c_i^\star|^2 \\ &= 2|x - c_i^\star|(|c_i - c_i^\star| + |c_j - c_j^\star|) + (|c_i - c_i^\star| + |c_j - c_j^\star|)^2 \\ &\leq 4|x - c_i^\star|F_1(q^\star, q) + 4F_1(q^\star, q)^2. \end{aligned} \tag{3.6}$$

Thus, using the fact that, for all $x \in V_i(\mathbf{c}^\star)$, we have

$$d(x, \partial V_i(\mathbf{c}^\star)) = \min_{i\neq j} d(x, h_{i,j}^\star),$$

we deduce from the previous observations that, for all $i \neq j$,

$$V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c}) \subset \left\{x \in E : md(x, \partial V_i(\mathbf{c}^\star)) \leq 2|x - q^\star(x)|F_1(q^\star, q) + 2F_1(q^\star, q)^2\right\}.$$

The right hand side being independent of j, we obtain in particular,

$$\begin{aligned} &\bigcup_{j\neq i} V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c}) \\ &\quad \subset \left\{x \in E : md(x, \partial V_i(\mathbf{c}^\star)) \leq 2|x - q^\star(x)|F_1(q^\star, q) + 2F_1(q^\star, q)^2\right\}. \end{aligned}$$

Therefore,

$$\mathbf{E}|q^\star(X) - q(X)|^2 = \sum_{i,j=1}^{k} P(V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c}))|c_i^\star - c_j|^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{k} P(V_i(\mathbf{c}^\star) \cap V_i(\mathbf{c}))|c_i^\star - c_i|^2 \\
&\quad + \sum_{i\neq j, i=1, j=1}^{k} P(V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c}))|c_i^\star - c_j|^2 \\
&\leq \sum_{i=1}^{k} P(V_i(\mathbf{c}^\star) \cap V_i(\mathbf{c}))|c_i^\star - c_i|^2 \\
&\quad + \sum_{i\neq j, i=1, j=1}^{k} P(V_i(\mathbf{c}^\star) \cap V_j(\mathbf{c}))(|c_j^\star - c_j| + M)^2 \\
&\leq F_1(q^\star, q)^2 + p^\star(F_1(q^\star, q))(F_1(q^\star, q) + M)^2
\end{aligned}
$$

which shows the desired result.

Appendix A: Technical results

A.1. Proofs for Remark 1.5

Recall that, for any two sets A, $B \subset E$, their Hausdorff distance is defined by

$$d_H(A, B) := \inf\{\, \varepsilon > 0 : \, A \subset B^\varepsilon \text{ and } B \subset A^\varepsilon \,\},$$

where $A^\varepsilon = \{x \in E : d(x, A) \leq \varepsilon\}$. The fact that $d_H(\mathbf{c}^\star, \mathbf{c}) \leq F_1(q^\star, q)$ then follows easily from definitions. Now, to prove the second statement, observe that, in the context of the finite sets $\mathbf{c}$ and $\mathbf{c}^\star$, the infimum in the definition of $\delta := d_H(\mathbf{c}^\star, \mathbf{c})$ is attained so that, for any $i \in \{1, \dots, k\}$, there exists some $j \in \{1, \dots, k\}$ such that $c_j^\star \in B(c_i, \delta) = \{x \in E : |x - c_i| \leq \delta\}$. Now suppose that

$$\delta < \frac{1}{2} \min_{i\neq j} |c_i^\star - c_j^\star|.$$

Then, the balls $B(c_i, \delta)$ are necessarily disjoint and therefore contain one and only one element of $\mathbf{c}^\star$, denoted $c_{\sigma(i)}^\star$. As a result,

$$F_1(q^\star, q) \leq \max_i |c_i - c_{\sigma(i)}^\star| = \delta,$$

where the last equality follows by construction. This implies the desired result.

A.2. Proof for Remark 2.2

Let $x \in E$ and denote $c_i = q(x)$. First, it may be checked that assumption $d(x, \mathcal{F}(\mathbf{c})) > \varepsilon$ holds if, and only if,

$$\forall\, j \neq i : \quad \frac{\langle x - c_i, c_j - c_i\rangle}{|c_j - c_i|} < \frac{|c_j - c_i|}{2} - \varepsilon. \tag{A.1}$$

Similarly, observe that $q(x_\lambda) = q(x)$ if and only if, for $j \neq i$, we have $|x_\lambda - c_i| < |x_\lambda - c_j|$. Using the definition of x_λ, this last condition may be equivalently written, for all $j \neq i$, as

$$\begin{aligned}(1+\lambda)^2|x-c_i|^2 &< |x-c_j|^2 + 2\lambda\langle x-c_i, x-c_j\rangle + \lambda^2|x-c_i|^2 \\ &= (1+\lambda^2)|x-c_i|^2 + 2\langle x-c_i, \lambda(x-c_j)+c_i-c_j\rangle + |c_i-c_j|^2.\end{aligned} \tag{A.2}$$

After simplification in (A.2), we therefore obtain that $q(x_\lambda) = q(x)$ if, and only if,

$$\forall\, j \neq i: \quad 0 < |c_i - c_j| - 2(1+\lambda)\frac{\langle x - c_i, c_j - c_i\rangle}{|c_j - c_i|}. \tag{A.3}$$

The result now easily follows from combining (A.1) and (A.3).

A.3. A consistency result

The next result is adapted from Theorems 4.12 and 4.21 in Graf and Luschgy, 2000.

Lemma A.1. *Suppose $X \in L^2(\mathbf{P})$. Then, letting $q^\star$ be an optimal k-points quantizer for the distribution of X and denoting $X_\lambda = X + \lambda(X - q^\star(X))$, the following statements hold.*

1. *For any $\lambda \geq 0$, there exists a k-points NN quantizer q_λ such that*

$$\|X_\lambda - q_\lambda(X_\lambda)\|^2 = \min_q \|X_\lambda - q(X_\lambda)\|^2,$$

   where the minimum is taken over all k-points quantizers.
2. *For all $\lambda_0 \geq 0$, if q_{λ_0} is unique,*

$$\lim_{\lambda \to \lambda_0} F_1(q_\lambda, q_{\lambda_0}) = 0.$$

Proof of lemma A.1. We state the result for a measure with bounded support and refer to Graf and Luschgy (2000) for unbounded case.

1. Let q_n be a sequence of quantizers such that

$$\|X_\lambda - q_n(X)\|^2 \to \inf_q \|X_\lambda - q(X)\|^2,$$

   as $n \to \infty$. Since balls in E are weakly compact, the centers $q_n(E) = \{c_1^n, \ldots, c_k^n\}$ weakly converge to some limit $\{c_1, \ldots, c_k\}$ up to a subsequence. Denote $q_0(X_\lambda)$ a limit of a weakly converging subsequence of $q_n(X_\lambda)$, realizing the limit $\liminf \|X - q_n(X_\lambda)\|^2$ then, by Fatou Lemma,

$$\begin{aligned}&\liminf \|X_\lambda - q_n(X_\lambda)\|^2 \\ &\quad \geq \mathbf{E} \liminf |X_\lambda - q_n(X_\lambda)|^2\end{aligned}$$

$$= \|X_\lambda - q_0(X_\lambda)\|^2 + \mathbf{E}\liminf |q_0(X_\lambda) - q_n(X_\lambda)|^2$$
$$\geq \inf_q \|X_\lambda - q(X_\lambda)\|^2 + \mathbf{E}\liminf |q_0(X_\lambda) - q_n(X_\lambda)|^2,$$

which shows that q_0 realizes the minimum of $\inf_q \|X_\lambda - q(X)\|^2$.

2. Similarly, for any sequence $\lambda_n \to \lambda_0$ as $n \to \infty$, $q_{\lambda_n}(X_{\lambda_n})$ has a weak limit $q_0(X_{\lambda_0})$ (up to subsequence). Then,

$$\begin{aligned}
&\|X_{\lambda_0} - q_{\lambda_0}(X_{\lambda_0})\|^2 \\
&\quad = \liminf_{n\to\infty} \|X_{\lambda_n} - q_{\lambda_0}(X_{\lambda_n})\|^2 \\
&\quad \geq \liminf_{n\to\infty} \|X_{\lambda_n} - q_{\lambda_n}(X_{\lambda_n})\|^2 \\
&\quad \geq \mathbf{E}\liminf |X_{\lambda_n} - q_{\lambda_n}(X_{\lambda_n})|^2 \\
&\quad \geq \|X_{\lambda_0} - q_0(X_{\lambda_0})\|^2 + \mathbf{E}\liminf |q_0(X_{\lambda_0}) - q_{\lambda_n}(X_{\lambda_n})|^2 \\
&\quad \geq \|X_{\lambda_0} - q_{\lambda_0}(X_{\lambda_0})\|^2 + \mathbf{E}\liminf |q_0(X_{\lambda_0}) - q_{\lambda_n}(X_{\lambda_n})|^2.
\end{aligned}$$

The last inequality holds because q_{λ_0} is optimal. This shows $\mathbf{E}\liminf |q_0(X) - q_n(X)|^2 = 0$, and since q_{λ_0} is supposed to be unique $q_0 = q_{\lambda_0}$ and therefore, every subsequence converges to the same limit q_{λ_0}; so $F_1(q_\lambda, q_{\lambda_0}) \to 0$. □

Appendix B: Stability of a learning problem

In this section, we briefly argue that the problem considered in the paper, while of special interest in the context of unsupervised learning, finds a natural extension in a more general framework of learning theory, namely the context of contrast minimization. Let $\mathcal{Z}$ be a measurable space equipped with a probability distribution P and let T be a given set of parameters. Suppose given a sample $Z_1, \ldots, Z_n$ of i.i.d. variables with common distribution P. Given a contrast function

$$\mathcal{C} : \mathcal{Z} \times T \to \mathbb{R}_+,$$

consider the problem of designing a data driven t, based on the sample $Z_1, \ldots, Z_n$, achieving a small value of the risk function

$$R(t) := \int \mathcal{C}(z, t)\, \mathrm{d}P(z).$$

This general problem, known as contrast minimization, is a classical way to unify the supervised and unsupervised learning approaches as illustrated in the next example.

Example B.1. *Classical examples include the following.*

- **Supervised learning.** *The supervised learning problem corresponds to the contrast minimization problem where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where T is a class of candidate functions $t : \mathcal{X} \to \mathcal{Y}$ and where, for a given loss function $\ell : \mathcal{Y}^2 \to \mathbb{R}_+$, the contrast is*

$$\mathcal{C}((x, y), t) = \ell(y, t(x)).$$

- **Unsupervised learning.** *The unsupervised learning problem discussed earlier in the present paper corresponds to the contrast minimization problem where $\mathcal{Z}$ is a metric space (E,d), where T is the set $\mathfrak{Q}$ of all k-points quantizers, for a given integer k, and where the contrast function is*

$$\mathcal{C}(x,q) = d(x,q(x))^2. \tag{B.1}$$

Given the general problem of contrast minimization, formulated above, one may naturally extend the question discussed in the present paper by considering the following notion of stability.

Definition B.2. *Consider a function $F : T^2 \to \mathbb{R}_+$ and an increasing function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$. Then, the contrast minimization problem is called (F,ϕ,ε)-stable if, for any $t^\star$ minimizing the risk on T,*

$$F(t^\star,t) \le \varepsilon \quad \Rightarrow \quad F(t^\star,t) \le \phi(R(t) - R(t^\star)).$$

Our main result, Theorem 2.3, proves the stability of the contrast minimization problem for the contrast function defined in (B.1). The following result proves the stability of the supervised learning problem for a strongly convex loss function.

Example B.3. *Consider the supervised learning problem described in the example above. Suppose there exists $\alpha > 0$ such that, for all $y \in \mathcal{Y}$, the function $u \in \mathcal{Y} \mapsto \ell(y,u)$ is α-strongly convex. Then, for any convex class T of functions $t : \mathcal{X} \to \mathcal{Y}$ and any $t^\star$ minimizing the risk on T, we have*

$$\int (t - t^\star)^2 \, d\mu \le \frac{4}{\alpha}(R(t) - R(t^\star)),$$

for any $t \in T$, where μ is the marginal of P on $\mathcal{X}$. In particular, for all $\varepsilon > 0$, this learning problem is (ε,ϕ)-stable for the $L^2(\mu)$ metric with $\phi(u) = 2\sqrt{u}/\sqrt{\alpha}$.

Acknowledgments

The authors are obliged to the anonymous referee for helpful comments and suggestions.

References

Abaya, E. A. and Wise, G. L. (1984). Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal of Applied Mathematics* **44** 183-189. MR0730008

Antos, A. (2005). Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Transactions on Information Theory* **51** 4022-4032. MR2239018

Antos, A., Györfi, L. and György, A. (2005). Improved convergence rates in empirical vector quantizer design. *IEEE Transactions on Information Theory* 4013-4022. MR2239017

Bartlett, P. L., Linder, T. and Lugosi, G. (1998). The minimax distorsion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory* **44** 1802-1813. MR1664098

Ben-David, S., Pál, D. and Simon, H. U. (2007). Stability of k-means clustering. In *International Conference on Computational Learning Theory* 20–34. Springer. MR1476916

Ben-David, S., Von Luxburg, U. and Pál, D. (2006). A sober look at clustering stability. In *International Conference on Computational Learning Theory* 5–19. Springer.

Biau, G., Devroye, L. and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory* **54** 781-790. MR2444554

Cadre, B. and Paris, Q. (2012). On Hölder fields clustering. *Test* **21** 301-316. MR2935361

Chou, P. A. (1994). The distorsion of vector quantizers trained on n vectors decreases to the optimum at $O_P(1/n)$. *IEEE Transactions on Information Theory* 457-457.

Graf, S. and Luschgy, H. (2000). *Foundations of quantization for probability distributions*. Springer-Verlag, New-York. MR1764176

Kumar, A. and Kannan, R. (2010). Clustering with spectral norm and the k-means algorithm. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science* 299-308. IEEE Computer Society. MR3025203

Levrard, C. (2015). Nonasymptotic bounds for vector quantization in Hilbert spaces. *The Annals of Statistics* **43** 592-619. MR3316191

Levrard, C. (2018). Quantization/clustering: when and why does k-means work? *Arxiv e-prints*. MR3803122

Linder, T. (2000). On the training distortion of vector quantizers. *IEEE Transactions on Information Theory* 1617-1623.

Linder, T. (2001). *Learning-theoretic methods in vector quantization*. Lecture Notes for the Advanced School on the Principle of Nonparametric Learning, Udine, Italy, July 9-13.

Linder, T., Lugosi, G. and Zeger, K. (1994). Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory* **40** 1728-1740. MR1322387

Loubes, J. M. and Pelletier, B. (2017). Prediction by quantization of a conditional distribution. *Electronic journal of statistics* **11** 2679–2706. MR3679906

Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of Lloyd's algorithm and its variants. *arXiv:1612.02099*.

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics* **9** 135-140. MR0600539

Pollard, D. (1982a). A central limit theorem for k-means clustering. *The Annals of Probability* **10** 199-205. MR0672292

Pollard, D. (1982b). Quantization and the method of k-means. *IEEE Transactions on Information Theory* **28** 1728-1740. MR0651814

Rakhlin, A. and Caponnetto, A. (2007). Stability of k-means clustering. In *Advances in neural information processing systems* 1121–1128.

Tang, C. and Monteleoni, C. (2016). On Lloyd's Algorithm: New Theoretical Insights for Clustering in Practice. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (A. Gretton and C. C. Robert, eds.). *Proceedings of Machine Learning Research* **51** 1280-1289. PMLR, Cadiz, Spain.