Quantization/clustering: when does k-means work?

Clément Levrard

Quantization/Clustering: when and why does k-means work.

Clément Levrard

Abstract: Though mostly used as a clustering algorithm, k-means are originally designed as a quantization algorithm. Namely, it aims at providing a compression of a probability distribution with k points. Building upon [21, 33], we try to investigate how and when these two approaches are compatible. Namely, we show that provided the sample distribution satisfies a margin like condition (in the sense of [27] for supervised learning), both the associated empirical risk minimizer and the output of Lloyd's algorithm provide almost optimal classification in certain cases (in the sense of [6]). Besides, we also show that they achieved fast and optimal convergence rates in terms of sample size and compression risk.

MSC 2010 subject classifications: 62H30, 62E17.
Keywords and phrases: k-means, clustering, quantization, separation rate, distortion.

1. Introduction

Due to its simplicity, k-means is one of the most popular clustering tool. It has been proved fruitful in many applications: as a last step of a spectral clustering algorithm [28], for clustering electricity demand curves [1], clustering DNA microarray data [34, 17] or EEG signals [29] among others. As a clustering procedure, k-means intends to groups data that are relatively similar into several well-separated classes. In other words, for a data set $\{X_1, \dots, X_n\}$ drawn in a Hilbert space $\mathcal{H}$, k-means outputs $\hat{\mathcal{C}} = (C_1, \dots, C_k)$ that is a collection of subsets of $\{1, \dots n\}$. To assess the quality of such a classification, it is often assumed that a target or natural classification $\mathcal{C}^* = (C_1^*, \dots, C_k^*)$ is at hand. Then a classification error may be defined by

$$\hat{R}_{classif}(\hat{\mathcal{C}}, \mathcal{C}^*) = \inf_{\sigma \in \mathcal{S}_k} \frac{1}{n} \sum_{j=1}^{k} \left| \hat{C}_{\sigma(j)} \cap (C_j^*)^c \right|,$$

where σ ranges in the set of k-permutations $\mathcal{S}_k$. Such a target classification $\mathcal{C}^*$ may be provided by a mixture assumption on the data, that is hidden i.i.d latent variables $Z_1, \dots, Z_n \in \{1, \dots, k\}$ are drawn and only i.i.d X_i's such that $X|Z = j \sim \phi_j$ are observed. This mixture assumption on the data is at the core of model-based clustering techniques, that cast the clustering problem into the density estimation framework. In this setting, efficient algorithms may be designed, provided that further assumptions on the ϕ_j's are made. For instance, if the ϕ_j's are supposed to be normal densities, this classification problem may be solved in practice using an EM algorithm [13].

However, by construction, k-means may rather be thought of as a quantization algorithm. Indeed, it is designed to output an empirical codebook $\hat{\mathbf{c}}_n = (\hat{c}_{n,1}, \dots, \hat{c}_{n,k})$, that is a k-vector of codepoints $\hat{c}_{n,j} \in \mathcal{H}$, minimizing

$$\hat{R}_{dist}(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\dots,k} \|X_i - c_j\|^2,$$

over the set of codebooks $\mathbf{c} = (c_1, \dots, c_k)$. Let $V_i(\mathbf{c})$ denote the j-th Voronoi cell associated with $\mathbf{c}$, that is $V_j(\mathbf{c}) = \{x| \quad \forall i \neq j \quad \|x - c_j\| \leq \|x - c_i\|\}$, and $Q_{\mathbf{c}}$ the function that maps every $V_j(\mathbf{c})$ onto $\mathbf{c}_j$, with ties arbitrarily broken. Then $\hat{R}_{dist}(\mathbf{c})$ is $P_n\|x - Q_{\mathbf{c}}(x)\|^2$, where $P_n f$ means integration with respect to the empirical distribution P_n. From this point of view, k-means aims at providing a quantizer $Q_{\hat{\mathbf{c}}_n}$ that realizes a good k-point compression of P, namely that has a low distortion $R_{dist}(\hat{\mathbf{c}}_n) = P\|x - Q_{\hat{\mathbf{c}}_n}(x)\|^2$.

This quantization field was originally developed to answer signal compression issues in the late 40's (see, e.g. [15]), but quantization may also be used as a pre-processing step for more involved statistical procedures, such as modeling meta-models for curve prediction by k 'local' regressions as in [4]. This domain provides most of the theoretical results for k-means (see, e.g., [23, 7]), assessing roughly that it achieves an optimal k-point compression up to $1/\sqrt{n}$ in terms of the distortion $P\|x - Q_{\mathbf{c}}(x)\|^2$, under a bounded support assumption on P. Note that other distortion measures can be considered: L_r distances, $r \geq 1$ (see, e.g., [16]), or replacing the squared Euclidean norm by a Bregman divergence ([14]).

In practice, k-means clustering is often performed using Lloyd's algorithm [24]. This iterative procedure is based on the following: from an initial codebook $\mathbf{c}^{(0)}$, partition the data according to the Voronoi cells of $\mathbf{c}^{(0)}$, then update the code point by computing the empirical mean over each cell. Since this step can only decrease the empirical distortion $\hat{R}_{dist}$, repeat until stabilization and output $\hat{\mathbf{c}}_{KM,n}$. Note that this algorithm is a particular instance of the Classification EM algorithm in the case where the components are assumed to have equal and spherical variance matrices [10]. As for EM's algorithms, the overall quality of the Lloyd's algorithm output mostly depends on the initialization. Most of the effective implementation use several random initializations, as for k-means ++ [3], resulting in an approximation of the true empirical distortion minimizer. This approximation may be build as close as desired (in terms of distortion) to the optimum [19] with high probability, provided that enough random initializations are allowed.

Roughly, these approximation results quantify the probability that a random initialization falls close enough to the empirical distortion minimizer $\hat{\mathbf{c}}_n$. It has been recently proved that, provided such a good random initialization is found, if P_n satisfies some additional clusterability assumption, then some further results on the misclassification error of the Lloyd's algorithm output can be stated. For instance, if $\min_{i\neq j} \|\hat{\mathbf{c}}_{n,i} - \hat{\mathbf{c}}_{n,j}\|/\sqrt{n}$ is large enough, then it is proved that $\hat{\mathbf{c}}_{KM,n}$ provides a close classification to $\hat{\mathbf{c}}_n$ [33]. In other words, if $\mathcal{C}(\hat{\mathbf{c}}_{KM,n})$ and $\mathcal{C}(\hat{\mathbf{c}}_n)$ denote the classifications associated with the Voronoi diagrams of $\hat{\mathbf{c}}_{KM,n}$ and $\hat{\mathbf{c}}_n$, then $\hat{R}_{classif}(\mathcal{C}(\hat{\mathbf{c}}_{KM,n}), \mathcal{C}(\hat{\mathbf{c}}_n))$ is small with high probability, provided

that the empirically optimal cluster centers are separated enough.

This *empirical* separation condition has deterministic counterparts that provide classification guarantees for k-means related algorithms, under model-based assumptions. Namely, if the sample is drawn according to a subGaussian mixture, then a separation condition on the true means of the mixture entails guarantees for the classification error $\hat{R}_{classif}(\hat{\mathcal{C}}, \mathcal{C}^*)$, where $\mathcal{C}^*$ is the latent variable classification [25, 9]. As will be detailed in Section 2, it is possible to define a separation condition without assuming that the underlying distribution is a subGaussian mixture (see, e.g., [20, 21]). This so-called margin condition turns out to be satisfied under model-based clustering assumptions such as quasi-Gaussian mixtures. It also holds whenever the distribution is supported on finitely many points.

Section 2 introduces notation and basic structural properties that the margin condition entails for probability distributions. To be more precise, a special attention is paid to the connection between classification and compression such a condition provides. For instance, it is exposed that whenever P satisfies a margin condition, there exist finitely many optimal classifications. Section 3 focuses on the compression performance that an empirical risk minimizer $\hat{\mathbf{c}}_n$ achieves under this margin condition. We state that fast convergence rates for the distortion are attained, that imply some guarantees on the classification error of $\hat{\mathbf{c}}_n$. At last, Section 4 intends to provide similar results, both in compression and classification, for an output $\hat{\mathbf{c}}_{KM,n}$ of the Lloyd's algorithm. We show that our deterministic separation condition ensures that an empirical one in satisfied with high probability, allowing to connect our approach to that of [33]. On the whole, we prove that $\hat{\mathbf{c}}_{KM,n}$ performs almost optimal compression, as well as optimal classification in the framework of [6].

2. Notation and margin condition

Throughout this paper, for $M > 0$ and a in $\mathcal{H}$, $\mathcal{B}(a, M)$ will denote the closed ball with center a and radius M. For a subset A of $\mathcal{H}$, $\bigcup_{a\in A} \mathcal{B}(a, M)$ will be denoted by $\mathcal{B}(A, M)$. With a slight abuse of notation, P is said to be M-bounded if its support is included in $\mathcal{B}(0, M)$. Furthermore, it will also be assumed that the support of P contains more than k points. Recall that we define the closed j-th Voronoi cell associated with $\mathbf{c} = (c_1, \dots, c_k)$ by $V_j(\mathbf{c}) = \{x| \quad \forall i \neq j \quad \|x - c_j\| \leq \|x - c_i\|\}$.

We let $X_1, \dots, X_n$ be i.i.d. random variables drawn from a distribution P, and introduce the following contrast function,

$$\gamma : \begin{cases} (\mathcal{H})^k \times \mathcal{H} & \longrightarrow & \mathbb{R} \\ (\mathbf{c}, x) & \longmapsto & \min_{j=1,\dots,k} \|x - c_j\|^2 \end{cases},$$

so that $R_{dist}(\mathbf{c}) = P\gamma(\mathbf{c}, .)$ and $\hat{R}_{dist}(\mathbf{c}) = P_n\gamma(\mathbf{c}, .)$. We let $\mathcal{M}$ denote the set of minimizers of $P\gamma(\mathbf{c}, .)$ (possibly empty). The most basic property of the set of minimizers is its stability with respect to isometric transformations that are P-compatible. Namely

Lemma 1. *[16, Lemma 4.7]*
Let T be an isometric transformation such that $T\sharp P = P$, where $T\sharp P$ denotes the distribution of $T(X)$, $X \sim P$. Then

$$T(\mathcal{M}) = \mathcal{M}.$$

Other simple properties of $\mathcal{M}$ proceed from the fact that $\mathbf{c} \mapsto \|x - c_j\|^2$ is weakly lower semi-continuous (see, e.g., [8, Proposition 3.13]), as stated below.

Proposition 1. *[14, Corollary 3.1] and [21, Proposition 2.1]*
Assume that P is M-bounded, then

i) $\mathcal{M} \neq \emptyset$.
ii) *If* $B = \inf_{\mathbf{c}^* \in \mathcal{M}, i \neq j} \|c_i^* - c_j^*\|$, *then* $B > 0$.
iii) *If* $p_{min} = \inf_{\mathbf{c}^* \in \mathcal{M}, i} P(V_i(\mathbf{c}^*))$, *then* $p_{min} > 0$.

Proposition 1 ensures that there exist minimizers of the true and empirical distortions R_{dist} and $\hat{R}_{dist}$. In what follows, $\hat{\mathbf{c}}_n$ and $\mathbf{c}^*$ will denote minimizers of $\hat{R}_{dist}$ and R_{dist} respectively. A basic property of distortion minimizers, called the centroid condition, is the following.

Proposition 2. *[16, Theorem 4.1] If $\mathbf{c}^* \in \mathcal{M}$, then, for all $j = 1, \dots, k$,*

$$P(V_j(\mathbf{c}^*))c_j^* = P\left(x 1_{V_j(\mathbf{c}^*)}(x)\right).$$

As a consequence, for every $\mathbf{c} \in \mathcal{H}^k$ and $\mathbf{c}^* \in \mathcal{M}$,

$$R_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c}^*) \leq \|\mathbf{c} - \mathbf{c}^*\|^2.$$

A direct consequence of Proposition 2 is that the boundaries of the Voronoi diagram $V(\mathbf{c})$ has null P-measure. Namely, if

$$N(\mathbf{c}^*) = \bigcup_{i \neq j} \left\{x | \quad \|x - c_i^*\| = \|x - c_j^*\|\right\},$$

then $P(N(\mathbf{c}^*)) = 0$. Hence the quantizer $Q_{\mathbf{c}^*}$ that maps $V_j(\mathbf{c}^*)$ onto c_j^* is well-defined P a.s. For a generic $\mathbf{c}$ in $\mathcal{B}(0, M)$, this is not the case. Thus, we adopt the following convention: $W_1(\mathbf{c}) = V_1(\mathbf{c})$, $W_2(\mathbf{c}) = V_2(\mathbf{c}) \setminus W_1(\mathbf{c})$, ..., $W_k(\mathbf{c}) = V_k(\mathbf{c}) \setminus W_{k-1}(\mathbf{c})$, so that the $W_j(\mathbf{c})$'s form a tessellation of $\mathbb{R}^d$. The quantizer $Q_{\mathbf{c}}$ can now be properly defined as the map that sends each $W_j(\mathbf{c})$ onto c_j. As a remark, if Q is a k-points quantizer, that is a map from $\mathbb{R}^d$ with images $c_1, \dots, c_k$, then it is immediate that $R_{dist}(Q) \geq R_{dist}(Q_{\mathbf{c}})$. This shows that optimal quantizers in terms of distortion are to be found among nearest-neighbor quantizers of the form $Q_{\mathbf{c}}$, $\mathbf{c}$ in $(\mathbb{R}^d)^k$.

An other key parameter for quantization purpose is the separation factor, that seizes the difference between local and global minimizers in terms of distortion.

Definition 1. *Denote by $\bar{\mathcal{M}}$ the set of codebooks that satisfy*

$$P(W_i(\mathbf{c}))\, c_i = P\left(x 1_{W_i(\mathbf{c})}(x)\right),$$

for any $i = 1, \dots, k$. *Let* $\varepsilon > 0$, *then* P *is said to be* ε*-separated if*

$$\inf_{\mathbf{c} \in \bar{\mathcal{M}} \setminus \mathcal{M}} R_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c}^*) \geq \varepsilon,$$

where $\mathbf{c}^* \in \mathcal{M}$.

The separation factor ε quantifies how difficult the identification of global minimizer might be. Its empirical counterpart in terms of $\hat{R}_{dist}$ can be thought of as the minimal price one has to pay when the Lloyd's algorithm ends up at a stationary point that is not an optimal codebook.

Note that local minimizers of the distortion, such as $k-1$-optimal quantizers, satisfy the centroid condition. Whenever $\mathcal{H} = \mathbb{R}^d$, P has a density and $P\|x\|^2 < \infty$, it can be proved that the set of minimizers of R_{dist} coincides with the set of codebooks satisfying the centroid condition, also called stationary points (see, e.g., Lemma A of [31]). However, this result cannot be extended to non-continuous distributions, as proved in Example 4.11 of [16].

Up to now, we only know that the set of minimizers of the distortion $\mathcal{M}$ is non-empty. From the compression point of view, this is no big deal if $\mathcal{M}$ is allowed to contain an infinite number of optimal codebooks. From the classification viewpoint, such a case may be interpreted as a case where P carries no natural classification of $\mathcal{H}$. For instance, if $\mathcal{H} = \mathbb{R}^2$ and $P \sim \mathcal{N}(0, I_2)$, then easy calculation and Lemma 1 show that $\mathcal{M} = \left\{(c_1, c_2) | \quad c_2 = -c_1, \quad \|c_1\| = 2/\sqrt{2\pi}\right\}$, hence $|\mathcal{M}| = +\infty$. In this case, it seems quite hard to define a natural classification of the underlying space, even if the $\mathbf{c}^*$'s are clearly identified. The following margin condition is intended to depict situations where a natural classification related with P exists.

Definition 2 (Margin condition)**.** *A distribution* P *satisfies a margin condition with radius* $r_0 > 0$ *if and only if*

i) P *is* M*-bounded,*
ii) for all $0 \leq t \leq r_0$,

$$\sup_{\mathbf{c}^* \in \mathcal{M}} P\left(\mathcal{B}(N(\mathbf{c}^*), t)\right) := p(t) \leq \frac{B p_{min}}{128 M^2} t. \tag{1}$$

Since $p(2M) = 1$, such a r_0 must satisfy $r_0 < 2M$. The margin condition introduced above asks that every classification associated with an optimal codebook $\mathbf{c}^*$ is a somehow natural classification. In other words P has to be concentrated enough around each c_j^*. This margin condition may also be thought of as a counterpart of the usual margin conditions for supervised learning stated in [27], where the weight of the neighborhood of the critical area $\{x| \quad P(Y = 1|X = x) = 1/2\}$ is controlled.

The scope of the margin condition allows to deal with several very different situations in the same way, as illustrated below.

2.1. Some instances of 'natural classifications'

Finitely supported distributions: If P is supported on finitely many points, say $x_1, \dots, x_r$. Then, $\mathcal{M}$ is obviously finite. Since, for all $\mathbf{c}^*$ in $\mathcal{M}$, $P(N(\mathbf{c}^*)) = 0$, we may deduce that $\inf_{\mathbf{c}^*,j} d(x_j, N(\mathbf{c}^*)) = r_0 > 0$. Thus, $p(t) = 0$ for $t \leq r_0$, and P satisfies a margin condition with radius r_0.

Truncated Gaussian mixtures: A standard assumption assessing the existence of a natural classification is the Gaussian mixture assumption on the underlying distribution, that allows to cast the classification issue into the density estimation framework. Namely, for $\mathcal{H} = \mathbb{R}^d$, $\tilde{P}$ is a Gaussian mixture if it has density

$$\tilde{f}(x) = \sum_{i=1}^{k} \frac{\theta_i}{(2\pi)^{d/2}\sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(x-m_i)^t \Sigma_i^{-1}(x-m_i)},$$

where the θ_i's denote the weights of the mixture, the m_i's the means and the Σ_i's are the $d \times d$ covariance matrices of the components.

Also denote by $\tilde{B} = \min_{i \neq j} \|m_i - m_j\|$ the minimum distance between two components, and by σ^2 and σ_-^2 the largest and smallest eigenvalues of the Σ_i's. It seems natural that the larger $\tilde{B}$ is compared to σ, the easier the classification problem would be. To this aim, we may define, for $\mathcal{C}$ and $\mathcal{C}^*$ two classifications the classification risk as the probability that a random point is misclassified, that is

$$R_{classif}(\mathcal{C}, \mathcal{C}^*) = \inf_{\sigma \in \mathcal{S}_k} P\left(\bigcup_{j=1}^{k} C_{\sigma(j)} \cap (C_j^*)^c \right).$$

In the case $k = 2$, $\theta_i = 1/2$ and $\Sigma_i = \sigma^2 I_d$, [6, Theorem 1 and 2] show that

$$\inf_{\hat{C}} \sup_{\sigma/\tilde{B} \leq \kappa} \mathbb{E} R_{classif}(\hat{C}, \mathcal{C}^*) \asymp \kappa^2 \sqrt{\frac{d}{n}},$$

up to log factors, where $\mathcal{C}^*$ denote the Bayes classification. Note that in this case, the Bayes classification is given by $C_j^* = V_j(\mathbf{m})$, that is the Voronoi diagram associated with the vector of means. Similarly we will show that for $\sigma/\tilde{B}$ small enough, a margin condition is satisfied.

Since Gaussian mixture have unbounded distributions, we may define a truncated Gaussian Mixture distribution by its density of the form

$$\tilde{f}(x) = \sum_{i=1}^{k} \frac{\theta_i}{(2\pi)^{d/2} N_i \sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(x-m_i)^t \Sigma_i^{-1}(x-m_i)} 1_{\mathcal{B}(0,M)}(x),$$

where N_i denotes a normalization constant for each truncated Gaussian variable. To avoid boundary issues, we will assume that M is large enough so that $M \geq 2 \sup_j \|m_j\|$. On the other hand, we also assume that M scales with σ, that is $M \leq c\sigma$, for some constant c. In such a setting, the following hold.

Proposition 3. *Denote by* $\eta = \min_i 1 - N_i$. *Then there exists constants* $c_1(k, \eta, d, \theta_{min})$ *and* $c_2(k, \eta, d, \theta_{min}, c_-, c)$ *such that*

- *If* $\sigma/\tilde{B} \leq \frac{1}{16c_1\sqrt{d}}$, *then for all* j *and* $\mathbf{c}^*$ *in* $\mathcal{M}$, $\|c_j^* - m_j\| \leq c_1\sigma\sqrt{d}$.
- *Assume that* $\sigma_- \geq c_-\sigma$, *for some constant* c_-. *If* $\sigma/\tilde{B} \leq c_2$, *then* $\mathbf{c}^*$ *is unique and* P *satisfies a margin condition with radius* $\tilde{B}/8$.

A possible choice of c_1 *is* $\sqrt{\frac{k2^{d+2}}{(1-\eta)\theta_{min}}}$.

A short proof is given in Section 6.1. Proposition 3 entails that (truncated) Gaussian mixtures are in the scope of the margin condition, provided that the components are well-separated. As will be detailed in Section 4, this implies that under the conditions of Proposition 3 the classification error of the outputs of the k-means algorithm is of order $\kappa^2\sqrt{d/n}$ as in [6].

2.2. An almost necessary condition

As described above, if the distribution P is known to carry a natural classification, then it is likely that it satisfies a margin condition. It is proved below that conversely an optimal codebook $\mathbf{c}^*$ provides a not so bad classification, in the sense that the mass around $N(\mathbf{c}^*)$ must be small. To this aim, we introduce, for $\mathbf{c}$ in $\mathcal{B}(0, M)^k$, and $i \neq j$, the following mass

$$p_{ij}(\mathbf{c}, t) = P\left(\left\{x| \quad 0 \leq \left\langle x - \frac{c_i + c_j}{2}, \frac{c_j - c_i}{r_{i,j}(\mathbf{c})}\right\rangle \leq t\right\} \cap V_j(\mathbf{c})\right),$$

where $r_{i,j}(\mathbf{c}) = \|c_i - c_j\|$. It is straightforward that $P(\mathcal{B}(N(\mathbf{c}), t)) \leq \sum_{i\neq j} p_{i,j}(\mathbf{c}, t)$. The necessary condition for optimality in terms of distortion is the following.

Proposition 4. *Suppose that* $\mathbf{c}^* \in \mathcal{M}$. *Then, for all* $i \neq j$ *and* $t < 1/2$,

$$\int_0^{tr_{i,j}(\mathbf{c}^*)} p_{i,j}(\mathbf{c}^*, s)ds \leq 2t^2 r_{i,j}(\mathbf{c}^*)\left[\frac{p_i(\mathbf{c}^*)}{1-2t} \wedge \frac{p_j(\mathbf{c}^*)}{1+2t}\right],$$
$$\int_0^{tr_{i,j}(\mathbf{c}^*)} p_{i,j}(\mathbf{c}^*, s)ds \leq t^2 r_{i,j}(\mathbf{c}^*)\frac{p_i(\mathbf{c}^*) + p_j(\mathbf{c}^*)}{2},$$

where $p_j(\mathbf{c}^*)$ *denotes* $P(V_j(\mathbf{c}^*))$.

A proof of Proposition 4 is given in Section 6.2. Whenever $p_{i,j}(\mathbf{c}^*, .)$ is continuous, Proposition 4 can provide a local upper bound on the mass around $N(\mathbf{c}^*)$.

Corollary 1. *Assume that* $\mathbf{c}^* \in \mathcal{M}$ *and, for all* $i \neq j$ *and* $t \leq t_0$ $p_{i,j}$ *is continuous on* $[0, t_0]$. *Then there exists* $r_0 > 0$ *such that, for all* $r \leq r_0$,

$$P(\mathcal{B}(N(\mathbf{c}^*), r)) \leq \frac{8k}{B}r.$$

Note that whenever $\mathcal{H} = \mathbb{R}^d$ and P has a density, the assumptions of Corollary 1 are satisfied. In this case, Corollary 1 states that all optimal codebooks satisfy a condition that looks like Definition 2, though with a clearly worse constant than the required one. Up to a thorough work on the constants involved in those results, this suggests that margin conditions (or at least weaker but sufficient versions) might be quite generally satisfied. As exposed below, satisfying such a condition provides interesting structural results.

2.3. Structural properties under margin condition

The existence of a natural classification, stated in terms of a margin condition in Definition 2, gives some guarantees on the set of optimal codebooks $\mathcal{M}$. Moreover, it also allows local convexity of the distortion R_{dist}. These properties are summarized in the following fundamental Proposition.

Proposition 5. *[21, Proposition 2.2] Assume that P satisfies a margin condition with radius r_0, then the following properties hold.*

i) For every $\mathbf{c}^*$ in $\mathcal{M}$ and $\mathbf{c}$ in $\mathcal{B}(0,M)^k$, if $\|\mathbf{c}-\mathbf{c}^*\| \leq \frac{Br_0}{4\sqrt{2}M}$, then

$$R_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c}^*) \geq \frac{p_{min}}{2}\|\mathbf{c}-\mathbf{c}^*\|^2. \tag{2}$$

ii) $\mathcal{M}$ is finite.

iii) There exists $\varepsilon > 0$ such that P is ε-separated.

iv) For all $\mathbf{c}$ in $\mathcal{B}(0,M)^k$,

$$\frac{1}{16M^2}\operatorname{Var}(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*(\mathbf{c}),.)) \leq \|\mathbf{c}-\mathbf{c}^*(\mathbf{c})\|^2 \leq \kappa_0\left(R_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c}^*)\right), \tag{3}$$

where $\kappa_0 = 4kM^2\left(\frac{1}{\varepsilon} \vee \frac{64M^2}{p_{min}B^2r_0^2}\right)$, and $\mathbf{c}^*(\mathbf{c}) \in \underset{\mathbf{c}^*\in\mathcal{M}}{\arg\min}\|\mathbf{c}-\mathbf{c}^*\|$.

Properties *ii)* and *iii)* guarantee that whenever a margin condition is satisfied, there exist finitely many optimal codebooks that are clearly separated in terms of distortion. When $P \sim \mathcal{N}(0, I_d)$, since $|\mathcal{M}| = +\infty$, P does not satisfy a margin condition. This finite set property also allows to give some structural results about the optimal codebooks. Namely, we can easily deduce the following.

Corollary 2. *Let $\mathcal{T}$ be the isometry group of P, and let $\mathbf{c}^* \in \mathcal{M}$. If P satisfies a margin condition, then $|\mathcal{T}(\mathbf{c}^*)| < +\infty$.*

An easy instance of application of Corollary 2 can be stated in the truncated Gaussian Mixture model exposed in Section 2.1. Let $S(\mathbf{m})$ denote the subset of $\{1,\dots,d\}$ such that, for all j and $r \notin S(\mathbf{m})$ $m_j^{(r)} = 0$, where $m_j^{(r)}$ denotes the r-th coordinate of m_j. Under the conditions of Proposition 3, if we further require that for all j and r,s in $S(\mathbf{m}) \times S(\mathbf{m})^c$, $\Sigma_{j,rs} = 0$, then it is immediate that $S(\mathbf{c}^*) \subset S(\mathbf{m})$. Such a property might be of particular interest when variable selection is performed as in [22].

Properties $i)$ and $iv)$ of Proposition 5 allow to make connections between the margin condition defined in Definition 2 and earlier results on improved convergence rates for the distortion. To be more precise, it is proved in [11] that if P has a continuous density, unique optimal codebook $\mathbf{c}^*$, and if the distortion function R_{dist} has a positive Hessian matrix at $\mathbf{c}^*$, then $R_{dist}(\hat{\mathbf{c}}_n) - R_{dist}(\mathbf{c}^*) = O_{\mathbb{P}}(1/n)$. It is straightforward that in the case where P has a continuous density and a unique optimal codebook, (2) yields that the Hessian matrix of the distortion is positive, hence the margin condition gives the convergence rate in $O_{\mathbb{P}}(1/n)$ for the distortion in this case.

On the other hand, it is proved in [2, Theorem 2] that, if $\mathrm{Var}(\gamma(\mathbf{c},.) - \gamma(\mathbf{c}^*(\mathbf{c}),.) \leq A(R_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c}^*))$, for some constant A, then the convergence rate $\mathbb{E}(R_{dist}(\hat{\mathbf{c}}_n) - R_{dist}(\mathbf{c}^*)) \leq C/n$ can be attained for the expected distortion of an empirical distortion minimizer. Thus, if P satisfies a margin condition, then (3) shows that P is in the scope of this result. In the following section, more precise bounds are derived for this excess distortion when P satisfies a margin condition.

At last, Properties $i)$ and $iv)$ allow to relate excess distortion and excess classification risk, when appropriate. For a codebook $\mathbf{c}$ in $\mathcal{H}^k$, we denote by $\mathcal{C}(\mathbf{c})$ its associated Voronoi partition (with ties arbitrarily broken).

Corollary 3. *Assume that P satisfies a margin condition (Definition 2) with radius r_0. Let δ denote the quantity $\frac{p_{min}B^2r_0^2}{64M^2} \wedge \varepsilon$. For every $\mathbf{c} \in \mathcal{H}^k$ such that $R_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c}^*) \leq \delta$, we have*

$$R_{classif}(\mathcal{C}(\mathbf{c}), \mathcal{C}(\mathbf{c}^*(\mathbf{c}))) \leq \frac{\sqrt{p_{min}}}{16M}\sqrt{R_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c}^*)},$$

where $\mathbf{c}^*(\mathbf{c})$ is a closest optimal codebook to $\mathbf{c}$.

A short proof of Corollary 3 is given in Section 6.3. Corollary 3 summarizes the connection between classification and distortion carried by the margin condition: if a natural classification exists, that is if P is separated into k spherical components, then this classification can be inferred from quantizers that are designed to achieve a low distortion. As exposed in the following section, an other interest in satisfying a margin condition is achieving an improved convergence rate in terms of distortion for the empirical distortion minimizer.

3. Convergence of the ERM

If P is M-bounded, then the excess distortion of an empirical distortion minimizer can be bounded by

$$\mathbb{E}(R_{dist}(\hat{\mathbf{c}}_n) - R_{dist}(\mathbf{c}^*)) \leq \frac{C(k)M^2}{\sqrt{n}}.$$

Such a result can be found in [23] for the case $\mathcal{H} = \mathbb{R}^d$, and in [7] for the general case where $\mathcal{H}$ is a separable Hilbert space. When P satisfies a margin condition,

faster rates can be achieved. The following Theorem is a refined version of [21, Theorem 3.1].

Theorem 1. *We assume that P satisfies a margin condition (Definition 2) with radius r_0, and we let δ denote the quantity $\frac{p_{min}B^2r_0^2}{64M^2} \wedge \varepsilon$. Then,*

$$\mathbb{E}\left(R_{dist}(\hat{\mathbf{c}}_n) - R_{dist}(\mathbf{c}^*)\right) \leq \frac{C\left(k + \log\left(|\bar{\mathcal{M}}|\right)\right)M^2}{np_{min}} + \left[\frac{20kM^2}{\sqrt{n}} - \delta\right] 1_{\delta < \frac{12kM^2}{\sqrt{n}}} + \left[e^{-\frac{n}{32M^4}\left((\delta - \frac{12kM^2}{\sqrt{n}})\right)^2}\frac{16M^2}{\sqrt{n}}\right] 1_{\delta \geq \frac{12kM^2}{\sqrt{n}}},$$

where C denotes a (known) constant and $|\bar{\mathcal{M}}|$ denotes the number of optimal codebooks up to relabeling.

A short proof is given in Section 6.4. Theorem 1 confirms that the fast $1/n$ rate for the distortion may be achieved as in [31] or [2], under slightly more general conditions. It also emphasizes that the convergence rate of the distortion is 'dimension-free', in the sense that it only depends on the dimension through the radius of the support M. For instance, quantization of probability distributions over the unit L_2-ball of $L_2([0,1])$ (squared integrable functions) is in the scope of Theorem 1. Note that a deviation bound is also available for $R_{dist}(\hat{\mathbf{c}}_n) - R_{dist}(\mathbf{c}^*)$, stated as (9).

In fact, this result shows that the key parameters that drive the convergence rate are rather the minimal distance between optimal codepoints B, the margin condition radius r_0 and the separation factor ε. These three parameters provide a local scale δ such that, if n is large enough to distinguish codebooks at scale δ in terms of slow-rated distortion, i.e. $\sqrt{n}\delta \geq 12kM^2$, then the distortion minimization boils down to k well separated mean estimation problems, leading to an improved convergence rate in $kM^2/(np_{min})$. Indeed, Theorem 1 straightforwardly entails that, for n large enough,

$$\mathbb{E}\left(R_{dist}(\hat{\mathbf{c}}_n) - R_{dist}(\mathbf{c}^*)\right) \leq \frac{C'(k + \log(|\bar{\mathcal{M}}|)M^2}{np_{min}}.$$

Thus, up to the $\log(|\bar{\mathcal{M}}|)$ factor, the right-hand side corresponds to $\sum_{j=1}^k \mathbb{E}\left(\|X - c_j^*\|^2 | X \in V_j(\mathbf{c}^*)\right)$. Combining Theorem 1 and Corollary 3 leads to the following classification error bound for the empirical risk minimizer $\hat{\mathbf{c}}_n$. Namely, for n large enough, it holds

$$\mathbb{E}\left[R_{classif}(\mathcal{C}(\hat{\mathbf{c}}_n), \mathcal{C}(\mathbf{c}^*(\hat{\mathbf{c}}_n)))\right] \leq C'\frac{\sqrt{k + \log(|\bar{\mathcal{M}}|}}{\sqrt{n}}.$$

This might be compared with the $1/\sqrt{n}$ rate obtained in [6, Theorem 1] for the classification error under Gaussian mixture with well-separated means assumption. Note however that in such a framework $\mathcal{C}(\mathbf{c}^*)$ might not be the optimal classification. However, under the assumptions of Proposition 3, $\mathcal{C}(\mathbf{c}^*)$ and $\mathcal{C}(\mathbf{m})$ can be proved close, and even the same in some particular cases as exposed in Corollary 4.

Next we intend to assess the optimality of the convergence rate exposed in Theorem 1, by investigating lower bounds for the excess distortion over class of distributions that satisfy a margin condition. We let $\mathcal{D}(B_-, r_{0,-}, p_-, \varepsilon_-)$ denote the set of distributions satisfying a margin condition with parameters $B \geq B_-$, $r_0 \geq r_{0,-}$, $p_{min} \geq p_-$ and $\varepsilon \geq \varepsilon_-$. Some lower bound on the excess distortion over these sets are stated below.

Proposition 6. *[21, Proposition 3.1] If $\mathcal{H} = \mathbb{R}^d$, $k \geq 3$ and $n \geq 3k/2$, then*

$$\inf_{\hat{\mathbf{c}}} \sup_{P \in \mathcal{D}(c_1 M k^{-1/d}, c_2 M k^{-1/d}, c_3/k, c_4 M^2 k^{-2/d}/\sqrt{n})} \mathbb{E}\left[R_{dist}(\hat{\mathbf{c}}) - R_{dist}(\mathbf{c}^*)\right] \geq c_0 \frac{M^2 k^{\frac{1}{2}-\frac{1}{d}}}{\sqrt{n}},$$

where c_0, c_1, c_2, c_3 and c_4 are absolute constants.

Thus, for a fixed choice of r_0, B and p_{min}, the upper bound given by Theorem 1 turns out to be optimal if the separation factor ε is allowed to be arbitrarily small (at least $\delta \lesssim kM^2/\sqrt{n}$). When all these parameters are fixed, the following Proposition 7 ensures that the $1/n$ rate is optimal.

Proposition 7. *Let $d = dim(\mathcal{H})$. Assume that $n \geq k$, then there exist constants c_1, c_2, c_3 and c_0 such that*

$$\inf_{\hat{\mathbf{c}}} \sup_{P \in \mathcal{D}(c_1 M k^{-1/d}, c_2 M k^{-1/d}, 1/k, c_3 M^2 k^{-(1+2/d)})} \mathbb{E}\left[R_{dist}(\hat{\mathbf{c}}) - R_{dist}(\mathbf{c}^*)\right] \geq c_0 \frac{M^2 k^{1-\frac{2}{d}}}{n}.$$

A proof of Proposition 7 can be found in Section 6.5. Proposition 7 ensures that the $1/n$-rate is optimal on the class of distributions satisfying a margin condition with fixed parameters. Concerning the dependency in k, note that Proposition 7 allows for $d = +\infty$, leading to a lower bound in k. In this case the lower bound differs from the upper bound given in Theorem 1 up to a $1/p_{min} \sim k$ factor. A question raised by the comparison of Proposition 6 and Proposition 7 is the following: can we retrieve the $1/\sqrt{n}$ rate when allowing other parameters such as B_- or $r_{0,-}$ to be small enough and ε_- fixed? A partial answer is provided by the following structural result, that connects the different quantities involved in the margin condition.

Proposition 8. *Assume that P satisfies a margin condition with radius r_0. Then the following properties hold.*

i) $\varepsilon \leq \frac{B^2}{4}$.
ii) $r_0 \leq 2B$.

A proof of Proposition 8 is given in Section 6.7. Such a result suggests that finding distributions that have B small enough whereas ε or r_0 remains fixed is difficult. As well, it also indicates that the separation rate in terms of B should be of order $Mk^{-1/d}n^{-1/4}$. Slightly anticipating, this can be compared with the $n^{-1/4}$ rate for the minimal separation distance between two means of a Gaussian mixture to ensure a consistent classification, as exposed in [6, Theorem 2].

4. Convergence of the k-means algorithm

Up to now some results have been stated on the performance of an empirical risk minimizer $\hat{\mathbf{c}}_n$, in terms of distortion or classification. Finding such a minimizer is in practice intractable (even in the plane this problem has been proved NP-hard, [26]). Thus, most of k-means algorithms provide an approximation of such a minimizer. For instance, Lloyd's algorithm outputs a codebook $\hat{\mathbf{c}}_{KM,n}$ that is provably only a stationary point of the empirical distortion $\hat{R}_{dist}$. Similarly to the EM algorithm, such a procedure is based on a succession of iterations that can only decrease the considered empirical risk $\hat{R}_{dist}$. Thus many random initializations are required to ensure that at least one one them falls into the basin of attraction of an empirical risk minimizer.

Interestingly, when such a good initialization has been found, some recent results ensure that the output $\hat{\mathbf{c}}_{KM,n}$ of Lloyd's algorithm achieves good classification performance, provided that the sample is in some sense well-clusterable. For instance, under the model-based assumption that X is a mixture of sub-Gaussian variables with means $\mathbf{m}$ and maximal variances σ^2, [25, Theorem 3.2] states that, provided $\tilde{B}/\sigma$ is large enough, after more that $4\log(n)$ iterations from a good initialization Lloyd's algorithm outputs a codebook with classification error less that $e^{-\tilde{B}^2/(16\sigma^2)}$. Note that the same kind of results hold for EM-algorithm in the Gaussian mixture model, under the assumption that $\tilde{B}/\sigma$ is large enough and starting from a good initialization (see, e.g., [12]).

In the case where P is not assumed to have a mixture distribution, several results on the classification risk $\hat{R}_{classif}(\hat{\mathbf{c}}_{KM,n}, \hat{\mathbf{c}}_n)$ are available, under clusterability assumptions. Note that this risk accounts for the misclassifications encountered by the output of Lloyd's algorithm compared to the empirical risk minimizer, in opposition to a latent variable classification as above.

Definition 3. *[33, Definition 1] A sample $X_1, \dots, X_n$ is f-clusterable if there exists a minimizer $\hat{\mathbf{c}}_n$ of $\hat{R}_{dist}$ such that, for $j \neq i$,*

$$\|\hat{\mathbf{c}}_{n,i} - \hat{\mathbf{c}}_{n,j}\| \geq f\sqrt{\hat{R}_{dist}(\hat{\mathbf{c}}_n)}\left(\frac{1}{\sqrt{n_i}} + \frac{1}{\sqrt{n_j}}\right),$$

where n_ℓ denotes $|\{i| \quad X_i \in V_\ell(\hat{\mathbf{c}}_n)\}|$.

It is important to mention that other definitions of clusterability might be found, for instance in [18, 5], each of them requiring that the optimal empirical codepoints are well-separated enough. Under such a clusterability assumption, the classification error of $\hat{\mathbf{c}}_{KM,n}$ can be proved small provided that a good initialization is chosen.

Theorem 2. *[33, Theorem 2] Assume that $X_1, \dots, X_n$ is f-clusterable, with $f > 32$ and let $\hat{\mathbf{c}}_n$ denote the corresponding minimizer of $\hat{R}_{dist}$. Suppose that the initialization codebook $\mathbf{c}^{(0)}$ satisfies*

$$\hat{R}_{dist}(\mathbf{c}^{(0)}) \leq g\hat{R}_{dist}(\hat{\mathbf{c}}_n),$$

with $g < \frac{f^2}{128} - 1$. *Then the outputs of Lloyds algorithm satisfies*

$$\hat{R}_{classif}(\hat{\mathbf{c}}_{KM,n}, \hat{\mathbf{c}}_n) \leq \frac{81}{8f^2}.$$

The requirement on the initialization codebook $\mathbf{c}^{(0)}$ is stated in terms of g-approximation of an empirical risk minimizer. Finding such approximations can be carried out using some approximated k-means techniques, such as k-means ++ ([3]), single Linkage ([33]), spectral clustering ([25]), or even more involved procedures as in [30] coming with complexity guarantees. All of them entail that a g-approximation of an empirical risk minimizer can be found with high probability (depending on g), that can be used as an initialization for the Lloyd's algorithm.

Interestingly, the following Proposition allows to think of Definition 3 as a margin condition (Definition 2) for the empirical distribution.

Proposition 9. *Let* $\hat{p}(t)$, $\hat{B}$ *and* $\hat{p}_{min}$ *denote the empirical counterparts of* $p(t)$, B *and* p_{min}. *If*

$$\hat{p}\left(\frac{16M^2 f}{\sqrt{n\hat{p}_{min}}\hat{B}}\right) \leq \hat{p}_{min},$$

then $X_1, \ldots, X_n$ *is* f*-clusterable.*

A proof of Proposition 9 can be found in Section 6.8. Intuitively, it seems likely that if $X_1, \ldots, X_n$ is drawn from a distribution P that satisfies a margin condition, then $X_1, \ldots, X_n$ is clusterable in the sense of Definition 3. This is formalized by the following Theorem.

Theorem 3. *Assume that* P *satisfies a margin condition. Let* $p > 0$. *Then, for* n *large enough, with probability larger than* $1 - 3n^{-p} - e^{-\frac{n}{32M^4}\left((\delta - \frac{12kM^2}{\sqrt{n}})\right)^2}$, $X_1, \ldots, X_n$ *is* $8\sqrt{p_{min}n}$*-clusterable. Moreover, on the same event, we have*

$$\|\hat{\mathbf{c}}_n - \hat{\mathbf{c}}_{KM,n}\| \leq \frac{3M}{np_{min}^2}.$$

A proof of Theorem 3 can be found in Section 6.9. Combining Theorem 3 and Theorem 2 ensures that whenever P satisfies a margin condition, then with high probability the classification error of the k-means codebook starting from a good initialization, $\hat{R}_{classif}(\hat{\mathbf{c}}_{KM,n}, \hat{\mathbf{c}}_n)$, is of order $1/(np_{min})$. Thus, according to Corollary 3, the classification error $\hat{R}_{classif}(\hat{\mathbf{c}}_{KM,n}, \mathbf{c}^*(\hat{\mathbf{c}}_{KM,n}))$ should be of order $\sqrt{(k + \log(|\bar{\mathcal{M}}|)/n}$, for n large enough. This suggests that the misclassifications of $\hat{\mathbf{c}}_{KM,n}$ are mostly due to the misclassifications of $\hat{\mathbf{c}}_n$, rather than the possible difference between $\hat{\mathbf{c}}_n$ and $\hat{\mathbf{c}}_{KM,n}$.

Combining the bound on $\|\hat{\mathbf{c}}_n - \hat{\mathbf{c}}_{KM,n}\|$ with a bound on $\|\hat{\mathbf{c}}_n - \mathbf{c}^*(\hat{\mathbf{c}}_n)\|$ that may be deduced from Theorem 1 and Proposition 5 may lead to guarantees on the distortion and classification risk $R_{dist}(\hat{\mathbf{c}}_{KM,n})$ and $R_{classif}(\mathcal{C}(\hat{\mathbf{c}}_{KM,n}), \mathcal{C}(\mathbf{c}^*(\hat{\mathbf{c}}_{KM,n}))$. An illustration of this point is given in Corollary 4.

Note also that the condition on the initialization in Theorem 2, that is $g \leq f^2/128 - 1$, can be written as $g \leq np_{min}/2 - 1$ in the framework of Theorem 3. Thus, for n large enough, provided that $R_{dist}(\mathbf{c}^*) > 0$, every initialization $\mathbf{c}^{(0)}$ turns out to be a good initialization.

Corollary 4. *Under the assumptions of Proposition 3, for $k = 2$, $\Sigma_i = \sigma I_d$, and $p_{min} = 1/2$, if n is large enough then*

$$\mathbb{E}R_{classif}\left(\mathcal{C}(\hat{\mathbf{c}}_{KM,n}), \mathcal{C}(\mathbf{m})\right) \leq C\sigma\sqrt{\frac{\log(n)}{n}},$$

where $\hat{\mathbf{c}}_{KM,n}$ denotes the output of the Lloyd's algorithm.

Note that in this case $\mathcal{C}(\mathbf{m})$ corresponds to the Bayes classification $\mathcal{C}^*$. Thus, in the 'easy' classification case $\frac{\sigma}{B}$ small enough, the output of the Lloyd's algorithm achieves the optimal classification error. It may be also worth remarking that this case is peculiar in the sense that $\mathcal{C}(\mathbf{c}^*) = \mathcal{C}(\mathbf{m})$, that is the classification targeted by k-means is actually the optimal one. In full generality, since $\mathbf{c}^* \neq \mathbf{m}$, a bias term accounting for $R_{classif}\left(\mathcal{C}(\mathbf{c}^*), \mathcal{C}(\mathbf{m})\right)$ is likely to be incurred.

5. Conclusion

As emphasized by the last part of the paper, the margin condition we introduced seems a relevant assumption when k-means based procedures are used as a classification tool. Indeed, such an assumption in some sense postulates that there exists a natural classification that can be reached through the minimization of a least-square criterion. Besides, it also guarantees that both a true empirical distortion minimizer and the output of the Lloyd's algorithm approximate well this underlying classification.

From a technical point a view, this condition was shown to connect a risk in distortion and a risk in classification. As mentioned above, this assesses the relevance of trying to find a good classifier via minimizing a distortion, but this also entails that the distortion risk achieves a fast convergence rate of $1/n$. Though this rate seems optimal on the class of distributions satisfying a margin condition, a natural question is whether fast rates of convergence for the distortion can occur more generally.

One negative clue is given by Proposition 4, where it is suggested that the margin condition is in fact quite mild. However, since the margin condition entails quite strong guarantees on the structure of optimal codebooks $\mathcal{M}$, it is possible to figure out situations when a margin condition cannot be satisfied but the convergence rate of the distortion of an empirical minimizer has to be fast. Indeed, consider P_0 a 2 components truncated Gaussian mixtures on $\mathbb{R}$ satisfying the requirements of Proposition 3. Then set P has a distribution over $\mathbb{R}^2$, invariant through rotations, and that has marginal distribution P_0 on the first coordinate. According to Corollary 2, P cannot satisfy a margin condition. However, by decomposing the distortion of codebooks into a radial and an orthogonal component, it can be shown that such a distribution gives

a fast convergence rate for the expected distortion of the empirical distortion minimizer.

The immediate questions issued by Proposition 4 and the above example are about the possible structure of the set of optimal codebooks: can we find distributions with infinite set of optimal codebooks that have finite isometry group? If not, through quotient-like operations can we always reach a fast convergence rate for the empirical risk minimizer? Beyond the raised interrogations, this short example allows to conclude that our margin condition cannot be necessary for the distortion of the ERM to converge fast.

6. Proofs

6.1. Proof of Proposition 3

The proof of Proposition 3 is based on the following Lemma.

Lemma 2. *[22, Lemma 4.2] Denote by* $\eta = \sup_{j=1,\dots,k} 1 - N_i$. *Then the risk* $R(\mathbf{m})$ *may be bounded as follows.*

$$R(\mathbf{m}) \leq \frac{\sigma^2 k \theta_{max} d}{(1-\eta)}, \tag{4}$$

where $\theta_{max} = \max_{j=1,\dots,k} \theta_j$. *For any* $0 < \tau < 1/2$, *let* $\mathbf{c}$ *be a codebook with a code point* c_i *such that* $\|c_i - m_j\| > \tau \tilde{B}$, *for every* j *in* $\{1,\dots,k\}$. *Then we have*

$$R(\mathbf{c}) > \frac{\tau^2 \tilde{B}^2 \theta_{min}}{4} \left(1 - \frac{2\sigma\sqrt{d}}{\sqrt{2\pi}\tau\tilde{B}} e^{-\frac{\tau^2\tilde{B}^2}{4d\sigma^2}}\right)^d, \tag{5}$$

where $\theta_{min} = \min_{j=1,\dots,k} \theta_j$. *At last, if* $\sigma^- \geq c_- \sigma$, *for any* τ' *such that* $2\tau + \tau' < 1/2$, *we have*

$$\forall t \leq \tau' \tilde{B} \quad p(t) \leq t \frac{2k^2 \theta_{max} M^{d-1} S_{d-1}}{(2\pi)^{d/2}(1-\eta) c_-^d \sigma^d} e^{-\frac{\left[\frac{1}{2} - (2\tau+\tau')\right]^2 \tilde{B}^2}{2\sigma^2}}, \tag{6}$$

where S_{d-1} *denotes the Lebesgue measure of the unit ball in* $\mathbb{R}^{d-1}$.

Proof. Proof of Proposition 3 We let $\tau = \frac{c_1\sqrt{d}\sigma}{\tilde{B}}$, with $c_1 = \sqrt{\frac{k2^{d+2}}{(1-\eta)\theta_{min}}}$. Note that $\frac{\sigma}{\tilde{B}} \leq \frac{1}{16\sqrt{d}c_1}$ entails $\tau \leq \frac{1}{16}$. Let $\mathbf{c}$ be a codebook with a code point c_i such that $\|c_i - m_j\| > \tau\tilde{B}$, for every j in $\{1,\dots,k\}$. Then (5) gives

$$\begin{aligned} R(\mathbf{c}) &> \frac{c_1^2 \sigma^2 \theta_- d 2^{-d}}{4} \\ &> \frac{k\sigma^2 d}{(1-\eta)} \\ &> R(\mathbf{m}), \end{aligned}$$

according to (4). Thus, an optimal codebook $\mathbf{c}^*$ satisfies, for all $j = 1, \dots, k$, $\|c_j^* - m_j\| \leq c_1\sqrt{d}\sigma$, up to relabeling. Under the condition $\frac{\sigma}{\tilde{B}} \leq \frac{1}{16\sqrt{d}c_1}$ and $\tau = \frac{c_1\sqrt{d}\sigma}{\tilde{B}}$, we have, since $\tau \leq \frac{1}{16}$, for every $\mathbf{c}^* \in \mathcal{M}$ and $j = 1, \dots, k$,

$$\tilde{B} \geq \frac{B}{2}, \quad \text{and} \quad \mathcal{B}(m_j, \frac{\tilde{B}}{4}) \subset V_j(\mathbf{c}^*).$$

We thus deduce that

$$\begin{aligned} p_{min} &\geq \frac{\theta_{min}}{(2\pi)^{\frac{d}{2}}} \int_{\mathcal{B}(0, \frac{\tilde{B}}{4})} e^{-\frac{\|u\|^2}{2}} du \\ &\geq \frac{\theta_{min}}{\frac{d}{2}} \left(1 - \frac{4\sigma\sqrt{d}}{\sqrt{2\pi}\tilde{B}} e^{-\frac{\tilde{B}^2}{16d\sigma^2}}\right)^d \\ &\geq \frac{\theta_{min}}{2^d(2\pi)^{\frac{d}{2}}}. \end{aligned}$$

Recall that we have $M \leq c\sigma$ for some constant $c > 0$, and $\sigma_- \geq c_-\sigma$. If $\tilde{B}/\sigma$ additionally satisfies $\frac{\tilde{B}^2}{\sigma^2} \geq 32 \log\left(\frac{2^{d+9} S_{d-1} k^2 c^{d-1}}{(1-\eta)\theta_{min} c_-^d}\right)$, choosing $\tau' = \frac{1}{8}$ in (6) leads to, for $t \leq \frac{\tilde{B}}{8}$,

$$\begin{aligned} p(t) &\leq t \frac{2k^2 M^{d-1} S_{d-1}}{(2\pi)^{\frac{d}{2}}(1-\eta)c_-^d \sigma^d} e^{-\frac{\tilde{B}^2}{32\sigma^2}} \\ &\leq t \frac{\theta_{min} M^{d-1}}{2^{d+8} c^{d-1} \sigma^d (2\pi)^{\frac{d}{2}}} \\ &\leq t \frac{\tilde{B}\theta_{min}}{(2\pi)^{\frac{d}{2}} 2^{d+8} M^2} \leq \frac{B p_{min}}{128 M^2}. \end{aligned}$$

Hence P satisfies a margin condition with radius $\tilde{B}/8$. Note that according to Proposition 5, no local minimizer of the distortion may be found in $\mathcal{B}(\mathbf{c}^*, r)$, for $\mathbf{c}^* \in \mathcal{M}$ and $r = \frac{Br_0}{4\sqrt{2M}}$. Note that $r \geq \frac{\tilde{B}^2}{64\sqrt{2}c\sigma}$ and $\|\mathbf{c}^* - \mathbf{m}\| \leq c_1\sigma\sqrt{kd}$. Thus, if $\frac{\sigma^2}{\tilde{B}^2} \leq \frac{1}{128\sqrt{2}c_1 c\sqrt{kd}}$, $\mathbf{c}^*$ is unique (up to relabeling). □

6.2. Proof of Proposition 4

Proof. Let $0 \leq t < \frac{1}{2}$, $\mathbf{c}^* \in \mathcal{M}$, and for short denote by r_{ij}, V_i, p_i the quantities $\|c_i^* - c_j^*\|$, $V_i(\mathbf{c}^*)$ and $p_i(\mathbf{c}^*)$. Also denote by u_{ij} the unit vector $\frac{c_j^* - c_i^*}{r_{ij}}$, $c_i^t = c + 2t(c_j^* - c_i^*)$, and by $H_{ij}^t = \left\{x | \quad \|x - c_i^t\| \leq \|x - c_j^*\|\right\}$. We design the quantizer Q_i^t as follows: for every $\ell \neq i, j$, $Q^t(V_\ell) = c_\ell^*$, $Q^t((V_i \cup V_j) \cap H_{ij}^t) = c_i^t$, and $Q^t((V_i \cup V_j) \cap (H_{ij}^t)^c) = c_j^*$. Then we may write

$$0 \leq R_{dist}(Q_i^t) - R_{dist}(\mathbf{c}^*) = 4p_i r_{ij}^2 t^2 + P\left((\|x - c_i^t\|^2 - \|x - c_j^*\|^2) 1_{V_j \cap H_{ij}^t}(x)\right). \tag{7}$$

On the other hand, straightforward calculation show that $V_j \cap H_{ij}^t = \left\{ x | \quad 0 \leq \left\langle x - \frac{c_i^* + c_j^*}{2}, u_{ij} \right\rangle \leq tr_{ij} \right\}$. Besides, for any $x \in V_j \cap H_{ij}^t$, denoting by s the quantity $\left\langle x - \frac{c_i^* + c_j^*}{2}, u_{ij} \right\rangle$, we have

$$\begin{aligned} \|x - c_i^t\|^2 - \|x - c_j^*\|^2 &= 2 \left\langle (1-2t)(c_j^* - c_i^*), x - \frac{c_i^* + c_j^*}{2} - t(c_j^* - c_i^*) \right\rangle \\ &= 2 \left[r_{ij} s(1-2t) - t(1-2t) r_{ij}^2 \right] \\ &= 2(1-2t)(s - tr_{ij}). \end{aligned}$$

Thus (7) may be written as

$$(1-2t) \int_0^{tr_{ij}} (tr_{ij} - s) dp_{ij}(s) \leq 2 p_i r_{ij} t^2.$$

Integrating by parts leads to $\int_0^{tr_{ij}} (tr_{ij} - s) dp_{ij}(s) = \int_0^{tr_{ij}} p_{ij}(u) du$. Thus

$$\int_0^{tr_{ij}} p_{ij}(\mathbf{c}^*, s) ds \leq 2t^2 r_{ij}(\mathbf{c}^*) \frac{p_i(\mathbf{c}^*)}{1-2t}.$$

The other inequalities follows from the same calculation, with the quantizer moving c_i^* to $c_i^* - 2t(c_j^* - c_i^*)$, and the quantizer moving c_i^* and c_j^* to $c_i^* + t(c_j^* - c_i^*)$ and $c_j^* + t(c_j^* - c_i)^*$, leaving the other cells V_ℓ unchanged. □

6.3. *Proof of Corollary 3*

Proof. According to [21, Lemma 4.4], if $R_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c}^*) \leq \delta$, then $\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\| \leq r$, with $r = \frac{Br_0}{4\sqrt{2}M}$. We may decompose the classification error as follows.

$$R_{classif}\left(\mathcal{C}(\mathbf{c}), \mathcal{C}(\mathbf{c}^*(\mathbf{c}))\right) = P\left(\bigcup_{j \neq i} V_j(\mathbf{c}^*) \cap V_i(\mathbf{c}) \right).$$

According to [21, Lemma 4.2],

$$\bigcup_{j \neq i} V_j(\mathbf{c}^*) \cap V_i(\mathbf{c}) \subset \mathcal{B}\left(N(\mathbf{c}^*(\mathbf{c})), \frac{4\sqrt{2}M}{B} \|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\| \right).$$

Thus, since P satisfies a margin condition with radius r_0,

$$\begin{aligned} R_{classif}\left(\mathcal{C}(\mathbf{c}), \mathcal{C}(\mathbf{c}^*(\mathbf{c}))\right) &\leq \frac{4\sqrt{2} p_{min}}{128M} \|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\| \\ &\leq \frac{\sqrt{p_{min}}}{16M} \sqrt{R_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c}^*)}, \end{aligned}$$

according to Proposition 5. □

6.4. Proof of Theorem 1

The proof of Theorem 1 relies on the techniques developed in the proof of [21, Theorem 3.1] and the following result from [7].

Theorem 4. *[7, Corollary 2.1] Assume that P is M-bounded. Then, for any $x > 0$, we have*

$$R_{dist}(\hat{\mathbf{c}}_n) - R_{dist}(\mathbf{c}^*) \leq \frac{12kM^2 + M^2\sqrt{2x}}{\sqrt{n}},$$

with probability larger than $1 - e^{-x}$.

We are now in position to prove Theorem 1.

Proof. Proof of Theorem 1 Assume that P satisfies a margin condition with radius r_0, and denote by $r = \frac{Br_0}{4\sqrt{2}M}$, $\delta = \frac{p_{min}}{2}r^2 \wedge \varepsilon$, where ε denotes the separation factor in Definition 1. For short denote, for any codebook $\mathbf{c} \in (\mathbb{R})^k$, by $\ell(\mathbf{c}, \mathbf{c}^*) = R_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c}^*)$. According to [21, Lemma 4.4], if $\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\| \geq r$, then $\ell(\mathbf{c}, \mathbf{c}^*) \geq \frac{p_{min}}{2}r^2 \wedge \varepsilon$. Hence, if $\ell(\mathbf{c}, \mathbf{c}^*) < \delta$, $\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\| < r$.

Using Theorem 4, we may write

$$\mathbb{P}\left(\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) > \delta\right) \leq e^{-\frac{n}{32M^4}\left((\delta - \frac{12kM^2}{\sqrt{n}})\right)^2}. \tag{8}$$

Now, for any $x > 0$ and constant C we have

$$\mathbb{P}\left[\left(\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) > C\frac{2}{p_{min}}\frac{\left(k + \log\left(\left|\bar{\mathcal{M}}\right|\right)\right)M^2}{n} + \frac{288M^2}{p_{min}n}x + \frac{64M^2}{n}x\right) \cap \left(\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) < \delta\right)\right]$$
$$\leq \mathbb{P}\left[\left(\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) > C\frac{2}{p_{min}}\frac{\left(k + \log\left(\left|\bar{\mathcal{M}}\right|\right)\right)M^2}{n} + \frac{288M^2}{p_{min}n}x + \frac{64M^2}{n}x\right) \cap \left(\hat{\mathbf{c}}_n \in \mathcal{B}(\mathcal{M}, r)\right)\right].$$

Proceeding as in the proof of [21, Theorem 3.1] entails, for every $x > 0$,

$$\mathbb{P}\left[\left(\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) > C\frac{2}{p_{min}}\frac{\left(k + \log\left(\left|\bar{\mathcal{M}}\right|\right)\right)M^2}{n} + \frac{288M^2}{p_{min}n}x + \frac{64M^2}{n}x\right) \cap \left(\hat{\mathbf{c}}_n \in \mathcal{B}(\mathcal{M}, r)\right)\right] \leq e^{-x}, \tag{9}$$

for some constant $C > 0$. Note that (8) and (9) are enough to give a deviation bound in probability. For the bound in expectation, set $\beta = \frac{2C\left(k + \log\left(\left|\bar{\mathcal{M}}\right|\right)\right)M^2}{np_{min}}$.

On one hand, Theorem 4 and (8) yield that

$$\begin{aligned}
\mathbb{E}(\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*)1_{\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*)>\delta}) &\leq \int_\delta^\infty \mathbb{P}(\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*)>u)du \\
&\leq \left[\frac{12kM^2}{\sqrt{n}} - \delta + \int_{\frac{12kM^2}{\sqrt{n}}}^\infty \mathbb{P}(\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*)>u)du\right] 1_{\delta<\frac{12kM^2}{\sqrt{n}}} \\
&\quad + \left[\int_0^\infty \mathbb{P}(\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*) - \frac{12kM^2}{\sqrt{n}} \geq (\delta - \frac{12kM^2}{\sqrt{n}}) + u)du\right] 1_{\delta\geq\frac{12kM^2}{\sqrt{*n}}} \\
&\leq \left[\frac{20kM^2}{\sqrt{n}} - \delta\right] 1_{\delta<\frac{12kM^2}{\sqrt{n}}} + \left[\int_0^\infty e^{-\frac{n}{32M^4}\left((\delta-\frac{12kM^2}{\sqrt{n}})+u\right)^2} du\right] 1_{\delta\geq\frac{12kM^2}{\sqrt{*n}}} \\
&\leq \left[\frac{20kM^2}{\sqrt{n}} - \delta\right] 1_{\delta<\frac{12kM^2}{\sqrt{n}}} + \left[e^{-\frac{n}{32M^4}\left((\delta-\frac{12kM^2}{\sqrt{n}})\right)^2}\frac{16M^2}{\sqrt{n}}\right] 1_{\delta\geq\frac{12kM^2}{\sqrt{n}}},
\end{aligned}$$

where we used $\sqrt{\pi}\leq 2$ and $(a+b)^2\geq a^2+b^2$ whenever a,$b\geq 0$. On the other hand, (9) entails

$$\begin{aligned}
\mathbb{E}(\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*)1_{\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*)\leq\delta}) &\leq (\beta-\delta)1_{\delta<\beta} + \left[\beta + \int_\beta^\infty \mathbb{P}((\ell(\hat{\mathbf{c}}_n,\mathbf{c}^*)\geq u)\cap(\hat{\mathbf{c}}_n\in\mathcal{B}(\mathcal{M},r)))du\right] 1_{\delta\leq\beta} \\
&\leq \beta + \frac{252M^2}{np_{min}},
\end{aligned}$$

where we used $p_{min}\leq 1$. Collecting the pieces gives the result of Theorem 1. □

6.5. *Proof of Proposition 7*

Proof. Assume that $dim(\mathcal{H}) = d$, and let $z_1,\dots,z_k$ be in $\mathcal{B}(0,M-\Delta/8)$ such that $\|z_i - z_j\|\geq\Delta$, and $\Delta\leq 2M$. Then slightly anticipating we may choose

$$\Delta \leq \frac{3M}{4k^{1/d}}.$$

Let $\rho = \Delta/8$, and for $\sigma\in\{-1,1\}^k$ and $\delta\leq 1$ denote by P_σ the following distribution. For any $A\subset\mathcal{H}$, and $i=1,\dots,k$,

$$P_\sigma(A\cap\mathcal{B}(z_i,\rho)) = \frac{1}{2\rho k}\left[(1+\sigma_i\delta)\lambda_1(e_1^*(A-z_i)\cap[0,\rho] + (1-\sigma_i\delta)\lambda_1(e_1^*(A-z_i)\cap[-\rho,0]\right],$$

where e_1^* denotes the projection onto the first coordinate and λ_1 denote the 1-dimensional Lebesgue measure. Note that for every i, $P_\sigma(\mathcal{B}(z_i,\rho)) = 1/k$. We let $\mathbf{c}_\sigma$ denote the codebook whose codepoints are $c_{\sigma,i} = z_i + \sigma_i\delta/2$. For such distributions P_σ's, it is shown in Section 6.6 that

$$\begin{cases} p_{min} & = & \frac{1}{k}, \\ B & \geq & \frac{3\Delta}{4}, \\ r_0 & \geq & \frac{\Delta}{4}, \\ \varepsilon & \geq & \frac{\Delta^2}{96k}. \end{cases}$$

Half of the proof of Proposition 7 is based on the following Lemma. For simplicity, we write $R(\hat{\mathbf{c}}, P_\sigma)$ for the distortion of the codebook $\hat{\mathbf{c}}$ when the distribution is P_σ.

Lemma 3. *For every σ, σ' in $\{-1,+1\}^k$,*

$$R(\mathbf{c}_{\sigma'}, P_\sigma) - R(\mathbf{c}_\sigma, P_\sigma) = \frac{2\delta^2\rho^2}{k} H(\sigma,\sigma') = \frac{2}{k}\|\mathbf{c}_\sigma - \mathbf{c}'_\sigma\|^2,$$

where $H(\sigma,\sigma') = \sum_{i=1}^k |\sigma_i - \sigma'_i|/2$. Moreover, for every codebook $\hat{\mathbf{c}}$ there exist $\hat{\sigma}$ such that, for all σ,

$$R(\hat{\mathbf{c}}, P_\sigma) - R(\mathbf{c}_\sigma, P_\sigma) \geq \frac{1}{4k}\|\mathbf{c}_{\hat{\sigma}} - \mathbf{c}_\sigma\|^2.$$

Lemma 3, whose proof is to be found in Section 6.6, ensures that our distortion estimation problem boils down to a σ estimation problem. Namely, we may deduce that

$$\inf_{\hat{Q}} \sup_{\sigma} \mathbb{E}(R(\hat{Q}, P_\sigma) - R(\mathbf{c}_\sigma, P_\sigma)) \geq \frac{\delta^2\rho^2}{4k} \inf_{\hat{\sigma}} \sup_{\sigma} H(\hat{\sigma},\sigma)).$$

The last part of the proof derives from the following.

Lemma 4. *If $k \geq n$ and $\delta \leq 1/2\sqrt{k/n}$, then, for every σ and σ' such that $H(\sigma,\sigma') = 1$,*

$$h^2(P_\sigma^{\otimes n}, P_{\sigma'}^{\otimes n}) \leq 1/4,$$

where h^2 denotes the Hellinger distance.

Thus, if we choose $\delta = \frac{\sqrt{k}}{2\sqrt{n}}$, $\Delta =$ and $\rho = \frac{\Delta}{8}$, a direct application of [35, Theorem 2.12] yields

$$\inf_{\hat{Q}} \sup_{\sigma} \mathbb{E}(R(\hat{Q}, P_\sigma) - R(\mathbf{c}_\sigma, P_\sigma)) \geq \frac{9}{2^{16}} M^2 \frac{k^{1-\frac{2}{d}}}{n}.$$

□

6.6. Intermediate results for Section 6.5

First we prove Lemma 3.

Proof. Proof of Lemma 3 We let I_i denote the 1 dimensional interval $[z_i - \rho e_1, z_i + \rho e_1]$, and V_i the Voronoi cell associated with z_i. At last, for a quantizer Q we denote by $R_i(Q, P_\sigma)$ the contribution of I_i to the distortion, namely $R_i(Q, P_\sigma) = P_\sigma\|x - Q(x)\|^2 1_{V_i}(x) = P_\sigma\|x - Q(x)\|^2 1_{I_i}(x)$. Since $\Delta/2 - 3\rho > 0$, $I_i \subset V_i(\mathbf{c}_\sigma)$, for every i and σ. According to the centroid condition (Proposition 2), if $|Q(I_i)| = 1$, that is only one codepoint is associated with I_i, then

$$R(Q, P_\sigma) = R(\mathbf{c}_\sigma, P_\sigma) + \sum_{i=1}^k P_\sigma(I_i)\|Q(I_i) - c_{\sigma,i}\|^2, \tag{10}$$

hence the first part of Lemma 3, with Q associated to $\mathbf{c}_{\sigma'}$.

Now let $\mathbf{c}$ be a codebook, and denote by Q the associated quantizer. Denote by $n_i = |Q(I_i)|$, $n_i^{in} = |Q(I_i) \cap V_i|$ and $n_i^{out} = |Q(I_i) \cap V_i^c$. If $n_i^{out} \geq 1$, then there exists $x_0 \in I_i$ such that $\|Q(x_0) - x_0\| \geq \Delta/2 - \rho$. Then, for any $x \in I_i$ it holds $\|Q(x) - x\| \geq \|Q(x) - x_0\| - 2\rho \geq \Delta/2 - 3\rho$. We deduce that for such an i, and every σ,

$$R_i(Q, \sigma) \geq \frac{1}{k} \left\| \frac{\Delta}{2} - 3\rho^2 \right\| = \frac{\rho^2}{k}.$$

The second base inequality is that, for every Q such that $Q(I_i) = z_i$, and every σ,

$$R_i(Q, \sigma) = \frac{\rho^2}{3k}.$$

We are now in position to build a new quantizer $\tilde{Q}$ that outperforms Q.

- If $n_i^{in} = 1$ and $n_i^{out} = 0$, then $\tilde{Q}(I_i) = \pi_{I_i}(Q(I_i))$, where π_{I_i} denote the projection onto I_i.
- If $n_i^{out} \geq 1$, then $\tilde{Q}(I_i) = z_i$.
- If $n_i^{in} \geq 2$ and $n_i^{out} = 0$, then $\tilde{Q}(I_i) = z_i$.

Such a procedure defines a k-point quantizer $\tilde{Q}$ that sends every I_i onto I_i. Moreover, we may write, for every σ

$$\begin{aligned} R(Q, P_\sigma) &= \sum_{n_i^{in}=1, n_i^{out}=0} R_i(Q, P_\sigma) + \sum_{n_i^{out} \geq 1} R_i(Q, P_\sigma) + \sum_{n_i^{out}=0, n_i^{in} \geq 2} R_i(Q, P_\sigma) \\ &\geq \sum_i R_i(\tilde{Q}, P_\sigma) + \left|\{i | n_i^{out} \geq 1\}\right| \frac{2\rho^2}{3k} - \left|\{i | n_i^{out} = 0, n_i^{in} \geq 2\}\right| \frac{\rho^2}{3k}. \end{aligned}$$

Since $|\{i|n_i^{out} \geq 1\}| \geq \left|\{i|n_i^{out} = 0, n_i^{in} \geq 2\}\right|$, we have $R(Q, P_\sigma) \geq R(\tilde{Q}, P_\sigma)$, for every σ. Note that such a quantizer $\tilde{Q}$ is indeed a nearest-neighbor quantizer, with images $\tilde{c}_i \in I_i$. For such a quantizer $\tilde{\mathbf{c}}$, (10) yields, for every σ,

$$R(\tilde{\mathbf{c}}, P_\sigma) - R(\mathbf{c}_\sigma, P_\sigma) = \frac{\|\tilde{\mathbf{c}} - \mathbf{c}_\sigma\|^2}{k}.$$

Now, if $\mathbf{c}_{\hat{\sigma}}$ denotes $\arg\min_{\mathbf{c}_\sigma} \|\mathbf{c}_\sigma - \tilde{\mathbf{c}}\|$, then, for every σ we have

$$\|\tilde{\mathbf{c}} - \mathbf{c}_\sigma\| \geq \frac{\|\mathbf{c}_{\hat{\sigma}} - \mathbf{c}_\sigma\|}{2}.$$

Thus, recalling our initial codebook $\mathbf{c}$, for every σ, $R(\mathbf{c}, P_\sigma) - R(\mathbf{c}_\sigma, P_\sigma) \geq \frac{1}{4k}\|\mathbf{c}_{\hat{\sigma}} - \mathbf{c}_\sigma\|^2$. □

6.7. *Proof of Proposition 8*

Let $\mathbf{c}^* \in \mathcal{M}$ and $i \neq j$ such that $\|c_i^* - c_j^*\| = B$. We denote by $Q_{i,j}$ the $(k-1)$-points quantizer that maps $V_\ell(\mathbf{c}^*)$ onto c_ℓ^*, for $\ell \neq i, j$, and $V_i(\mathbf{c}^*) \cup V_j(\mathbf{c}^*)$ onto

$\frac{c_i^*+c_j^*}{2}$. Then $R_{dist}(Q_{i,j}) - R_{dist}(\mathbf{c}^*) = (p_i(\mathbf{c}^*) + p_j(\mathbf{c}^*))\frac{B^2}{4} \leq \frac{B^2}{4}$. Thus, denoting by $\mathbf{c}^{*,(k-1)}$ an optimal $(k-1)$-points quantizer, $R_{dist}(\mathbf{c}^{*,(k-1)}) - R_{dist}(\mathbf{c}^*) \leq \frac{B^2}{4}$. Since $\varepsilon \leq R_{dist}(\mathbf{c}^{*,(k-1)}) - R_{dist}(\mathbf{c}^*)$, the first part of Proposition 8 follows.

For the same optimal codebook $\mathbf{c}^*$, we denote for short by $p(t)$ the quantity

$$p(t) = P\left(\left\{x| \quad 0 \leq \left\langle x - \frac{c_i + c_j}{2}, \frac{c_i - c_j}{r_{i,j}(\mathbf{c}^*)} \right\rangle \leq t\right\} \cap V_i(\mathbf{c})\right),$$

and by $p_i = p_i(\mathbf{c}^*)$. According to Proposition 2, we have

$$\begin{aligned} p_i \frac{B}{2} &= P\left(\left\langle x - \frac{c_i^* + c_j^*}{2}, \frac{c_i^* - c_j^*}{r_{i,j}} \right\rangle 1_{V_i(\mathbf{c}^*)}(x)\right) \\ &= \int_0^{2M} t dp(t). \end{aligned} \tag{11}$$

Assume that $r_0 > 2B$. According to (11), we may write, integrating by parts,

$$\begin{aligned} p_i \frac{B}{2} &\geq \int_0^{2B} t dp(t) \\ &\geq 2Bp(2B) - \int_0^{2B} p(t)dt \\ &\geq 2Bp(2B) - \frac{B^3 p_{min}}{64M^2}. \end{aligned}$$

Hence $p(2B) \leq \frac{p_i}{3}$. On the other hand, (11) also yields that

$$\begin{aligned} p_i \frac{B}{2} &\geq \int_{2B}^{2M} t dp(t) \\ &\geq 2B\left(p_i - p(2B)\right) \\ &\geq p_i \frac{4B}{3}, \end{aligned}$$

hence the contradiction.

6.8. *Proof of Proposition 9*

The proof of Proposition 9 is based on the following Lemma, that connects the clusterability assumption introduced in Definition 3 to another clusterability definition introduced in [18].

Lemma 5. *[32, Lemma 10] Assume that there exist d_{rs}'s, $r \neq s$, such that, for any $r \neq s$ and $x \in V_s(\hat{\mathbf{c}}_n)$,*

$$P_n\left(\{x| \quad \|x_{rs} - \hat{c}_r\| \leq \|x_{rs} - \hat{c}_s\| + d_{rs}\}\right) < \hat{p}_{min},$$

where x_{rs} denotes the projection of x onto the line joining $\hat{c}_r$ and $\hat{c}_s$. Then, for all $r \neq s$,

$$\|\hat{c}_r - \hat{c}_s\| \geq d_{rs}.$$

Proof. Proof of Proposition 9 Now let $x \in \{x| \quad \|x_{rs} - \hat{c}_r\| \leq \|x_{rs} - \hat{c}_s\| + d_{rs}\} \cap V_s(\hat{\mathbf{c}}_n)$, for $d_{rs} \leq 2M$. Then

$$\|x - \hat{c}_r\| \leq \|x - \hat{c}_s\| + d_{rs}$$
$$\|x - \hat{c}_s\| \leq \|x - \hat{c}_r\|.$$

Taking squares of both inequalities leads to

$$\left\langle \hat{c}_s - \hat{c}_s, x - \frac{\hat{c}_r + \hat{c}_s}{2} \right\rangle \geq 0$$
$$2\left\langle \hat{c}_s - \hat{c}_s, x - \frac{\hat{c}_r + \hat{c}_s}{2} \right\rangle \leq d_{rs}^2 + 2d_{rs}\|x - \hat{c}_r\| \leq 8Md_{rs}.$$

We deduce from above that $d(x, \partial V_s(\hat{\mathbf{c}}_n)) \leq \frac{8M}{B} d_{rs}$, hence $x \in \mathcal{B}(N(\hat{\mathbf{c}}_n), \frac{8M}{B} d_{rs})$. Set

$$d_{rs} = \frac{2Mf}{\sqrt{n_{min}}} \geq f\sqrt{\hat{R}_{dist}(\hat{\mathbf{c}}_n)}\left(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}}\right),$$

and assume that $\hat{p}\left(\frac{16M^2 f}{\sqrt{n\hat{p}_{min}}B}\right) \leq \hat{p}_{min}$. Then Lemma 5 entails that for all $\hat{\mathbf{c}}_n$ minimizing $\hat{R}_{dist}$ and $r \neq s$, $\|\hat{c}_r - \hat{c}_s\| \geq d_{rs}$. Hence $X_1, \ldots, X_n$ is f-clusterable. □

6.9. Proof of Theorem 3

Proof. Assume that P satisfies a margin condition with radius r_0. For short we denote $R_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c}^*)$ by $\ell(\mathbf{c}, \mathbf{c}^*)$. As in the proof of Theorem 1, according to (8), (9), choosing $x = p\log(n)$, for n large enough, it holds, for every minimizer $\hat{\mathbf{c}}_n$ of $\hat{R}_{dist}$,

$$\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq C\frac{M^2 p\log(n)}{np_{min}}$$
$$\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \geq \frac{p_{min}}{2}\|\hat{\mathbf{c}}_n - \mathbf{c}^*(\hat{\mathbf{c}}_n)\|^2,$$

with probability larger than $1 - n^{-p} - e^{-\frac{n}{32M^4}\left((\delta - \frac{12kM^2}{\sqrt{n}})\right)^2}$. On this probability event we may thus write

$$\|\hat{\mathbf{c}}_n - \mathbf{c}^*(\hat{\mathbf{c}}_n)\| \leq C\frac{\sqrt{p\log(n)}}{p_{min}\sqrt{n}}. \tag{12}$$

Since $N(\hat{\mathbf{c}}_n) \subset \mathcal{B}(N(\mathbf{c}^*(\hat{\mathbf{c}}_n, \sqrt{2}\|\hat{\mathbf{c}}_n - \mathbf{c}^*(\hat{\mathbf{c}}_n)\|)$, we get

$$\hat{p}(t) \leq p\left(t + \sqrt{2}C\frac{\sqrt{p\log(n)}}{p_{min}\sqrt{n}}\right) \tag{13}$$
$$\leq \frac{Bp_{min}t}{128M^2} + C\frac{B\sqrt{p\log(n)}}{M\sqrt{n}},$$

when n is large enough so that $r_n < r_0$ and for $t \leq r_0 - r_n$, with $r_n = C\frac{\sqrt{p\log(n)}}{p_{min}\sqrt{n}}$. It remains to connect $\hat{p}_{min}$ and $\hat{B}$ with their deterministic counterparts. First, it is straightforward that

$$\hat{B} \geq B - \sqrt{2}r_n \geq \frac{B}{2}, \tag{14}$$

for n large enough. The bound for $\hat{p}_{min}$ is slightly more involved. Let i and $\hat{\mathbf{c}}_n$ such that $\hat{p}_{min} = P_n\left(V_i(\hat{\mathbf{c}}_n)\right)$. Then we may write

$$\begin{aligned}\hat{p}_{min} &= P_n\left(V_i(\hat{\mathbf{c}}_n)\right) \\ &= P_n\left(V_i(\mathbf{c}^*(\hat{\mathbf{c}}_n))\right) - P_n\left(V_i(\mathbf{c}^*(\hat{\mathbf{c}}_n)) \cap V_i(\hat{\mathbf{c}}_n)^c\right) + P_n\left(V_i(\mathbf{c}^*(\hat{\mathbf{c}}_n))^c \cap V_i(\hat{\mathbf{c}}_n)\right).\end{aligned}$$

According to [21, Lemma 4.2], $V_i(\mathbf{c}^*(\hat{\mathbf{c}}_n))\Delta V_i(\hat{\mathbf{c}}_n) \subset \mathcal{B}\left(N(\mathbf{c}^*(\hat{\mathbf{c}}_n), \frac{4\sqrt{2}M}{B}r_n\right)$, where Δ denotes the symmetric difference. Hoeffding's inequality gives

$$\left|(P_n - P)\bigcup_{\mathbf{c}^*\in\mathcal{M}} N(\mathbf{c}^*, \frac{4\sqrt{2}M}{B}r_n)\right| \leq \sqrt{\frac{2p\log(n)}{n}},$$

with probability larger than $1 - n^{-p}$. Hence

$$\begin{aligned}P_n\left(V_i(\mathbf{c}^*(\hat{\mathbf{c}}_n))\Delta V_i(\hat{\mathbf{c}}_n)\right) &\leq p\left(\frac{4\sqrt{2}M}{B}r_n\right) + \sqrt{\frac{2p\log(n)}{n}} \\ &\leq C\frac{\sqrt{p\log(n)}}{p_{min}\sqrt{n}},\end{aligned}$$

for n large enough so that $\frac{4\sqrt{2}M}{B}r_n \leq r_0$. Concerning $P_n\left(V_i(\mathbf{c}^*(\hat{\mathbf{c}}_n))\right)$, using Hoeffding's inequality again we may write

$$\begin{aligned}P_n\left(V_i(\mathbf{c}^*(\hat{\mathbf{c}}_n))\right) &\geq p_{min} - \sup_{\mathbf{c}^*\in\bar{\mathcal{M}}, i=1,\dots,k} |(P_n - P)V_i(\mathbf{c}^*)| \\ &\geq p_{min} - \sqrt{\frac{2(p\log(n) + \log(k|\bar{\mathcal{M}}|)}{n}},\end{aligned}$$

with probability larger than $1 - n^{-p}$. We deduce that for n large enough,

$$\hat{p}_{min} \geq p_{min} - C\frac{\sqrt{p\log(n)}}{p_{min}\sqrt{n}} \geq \frac{p_{min}}{2},$$

for n large enough. Thus, (13) gives

$$\hat{p}(t) \leq \frac{\hat{B}\hat{p}_{min}}{512M^2}t + C\frac{\hat{B}\sqrt{p\log(n)}}{M\sqrt{n}}.$$

For n large enough so that $C\frac{\hat{B}\sqrt{p\log(n)}}{M\sqrt{n}} \leq \frac{\hat{p}_{min}}{2}$, Proposition 9 ensures that $X_1, \dots, X_n$ is $8\sqrt{p_{min}n}$-clusterable.

According to Theorem 2, on this probability event, at most $\frac{1}{2p_{min}}$ points are misclassified by $\hat{\mathbf{c}}_{KM,n}$ compared to $\hat{\mathbf{c}}_n$. Thus, denoting by $n_j = nP_nV_j(\hat{\mathbf{c}}_n)$ and $\hat{n}_j = n(P_nV_j(\hat{\mathbf{c}}_{KM,n}))\hat{c}_{KM,j}$, we may write

$$\sum_{j=1}^{k} n_j\|\hat{c}_j - \hat{c}_{KM,j}\| \leq \sum_{j=1}^{k} \|n_j\hat{c}_j - \hat{n}_j\hat{c}_{KM,j}\| + |n_j - \hat{n}_j|\,\|\hat{c}_{KM,j}\|$$
$$\leq \sum_{j=1}^{k} \left\| \sum_{i=1}^{n} X_i(1_{V_j(\hat{\mathbf{c}}_n)}(X_i) - 1_{V_j(\hat{\mathbf{c}}_{KM,n})}(X_i)) \right\| + \frac{M}{p_{min}},$$

since $\hat{\mathbf{c}}_{KM,n}$ and $\hat{\mathbf{c}}_n$ satisfy the centroid condition (Proposition 2). Thus,

$$\sum_{j=1}^{k} n_j\|\hat{c}_j - \hat{c}_{KM,j}\| \leq \frac{3M}{2p_{min}}.$$

At last, since for all $j = 1, \ldots, k$, $\frac{n_j}{n} \geq \hat{p}_{min} \geq \frac{p_{min}}{2}$, we deduce that

$$\|\hat{\mathbf{c}}_n - \hat{\mathbf{c}}_{KM,n}\| \leq \frac{3M}{np_{min}^2}.$$

□

6.10. *Proof of Corollary 4*

Proof. We recall that under the assumptions of Proposition 3, $\mathbf{c}^*$ is unique and P satisfies a margin condition with radius $\tilde{B}/8$. As in the proof of Theorem 3, we assume that

$$\|\hat{\mathbf{c}}_n - \mathbf{c}^*\| \leq C\frac{\sqrt{1\log(n)}}{p_{min}\sqrt{n}}.$$

This occurs with probability larger than $1 - n^{-1} - e^{-\frac{n}{32M^4}\left((\delta - \frac{12kM^2}{\sqrt{n}})\right)^2}$. It can be deduced from [7, Corollary 2.1] that, with probability larger than $1 - 2e^{-x}$,

$$\sup_{\mathbf{c}\in\mathcal{B}(0,M)^k} \left|\hat{R}_{dist}(\mathbf{c}) - R_{dist}(\mathbf{c})\right| \leq \frac{6kM^2 + 8M^2\sqrt{2x}}{\sqrt{n}}.$$

Therefore, for n large enough, it holds

$$\hat{R}_{dist}(\hat{\mathbf{c}}_n) \geq \frac{R_{dist}(\mathbf{c}^*)}{2},$$

with probability larger than $1 - 1/n$. On this probability event, a large enough n entails that every initialization of the Lloyd's algorithm is a good initialization. According to Theorem 2, we may write

$$\|\hat{\mathbf{c}}_{KM,n} - \mathbf{c}^*\| \leq C\frac{M\sqrt{\log(n)}}{\sqrt{n}}.$$

Since, according to [21, Lemma 4.2], $V_i(\mathbf{c}^*)\Delta V_i(\hat{\mathbf{c}}_{KM,n}) \subset \mathcal{B}\left(N(\mathbf{c}^*, \frac{4\sqrt{2}M}{B}\|\hat{\mathbf{c}}_{KM,n} - \mathbf{c}^*\|\right)$, the margin condition entails that

$$R_{classif}\left(\mathcal{C}(\hat{\mathbf{c}}_{KM,n}), \mathcal{C}(\mathbf{c}^*)\right) \leq CM\sqrt{\frac{\log(n)}{n}} \leq C\sigma\sqrt{\frac{\log(n)}{n}}.$$

Using Markov's inequality yields the same result in expectation. It remains to note that in the case $k = 2$, $\Sigma_i = \sigma^2 I_d$ and $p_1 = p_2 = \frac{1}{2}$, though $\mathbf{c}^*$ may differ from $\mathbf{m}$, we have $\mathcal{C}(\mathbf{c}^*) = \mathcal{C}(\mathbf{m})$. □

References

[1] Antoniadis, A., Brosat, X., Cugliari, J. and Poggi, J.-M. (2011). Clustering functional data using wavelets Research Report No. RR-7515.

[2] Antos, A., Györfi, L. and György, A. (2005). Individual convergence rates in empirical vector quantizer design. *IEEE Trans. Inform. Theory* **51** 4013–4022. MR2239017 (2007a:94125)

[3] Arthur, D. and Vassilvitskii, S. (2007). `k-means++`: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035. ACM, New York. MR2485254

[4] Auder, B. and Fischer, A. (2012). Projection-based curve clustering. *Journal of Statistical Computation and Simulation* **82** 1145-1168.

[5] Awasthi, P. and Sheffet, O. (2012). Improved spectral-norm bounds for clustering. In *Approximation, randomization, and combinatorial optimization. Lecture Notes in Comput. Sci.* **7408** 37–49. Springer, Heidelberg. MR3003539

[6] Azizyan, M., Singh, A. and Wasserman, L. (2013). Minimax Theory for High-dimensional Gaussian Mixtures with Sparse Mean Separation. In *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 2139–2147. Curran Associates, Inc.

[7] Biau, G., Devroye, L. and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory* **54** 781–790. MR2444554

[8] Brezis, H. (2011). *Functional analysis, Sobolev spaces and partial differential equations. Universitext.* Springer, New York. MR2759829 (2012a:35002)

[9] Bunea, F., Giraud, C., Royer, M. and Verzelen, N. (2016). PECOK: a convex optimization approach to variable clustering. *ArXiv e-prints.*

[10] Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Comput. Statist. Data Anal.* **14** 315–332. MR1192205

[11] Chou, P. A. (1994). The distortion of vector quantizers trained on n vectors decreases to the optimum as $\mathcal{O}_p(1/n)$. In *Proc. IEEE Int. Symp. Inf. Theory* 457.

[12] Dasgupta, S. and Schulman, L. (2007). A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *J. Mach. Learn. Res.* **8** 203–226. MR2320668

[13] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B* **39** 1–38.

[14] Fischer, A. (2010). Quantization and clustering with Bregman divergences. *J. Multivariate Anal.* **101** 2207–2221. MR2671211

[15] Gersho, A. and Gray, R. M. (1991). *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, MA, USA.

[16] Graf, S. and Luschgy, H. (2000). *Foundations of quantization for probability distributions. Lecture Notes in Mathematics* **1730**. Springer-Verlag, Berlin. MR1764176

[17] Kim, K., Zhang, S., Jiang, K., Cai, L., Lee, I.-B., Feldman, L. J. and Huang, H. (2007). Measuring similarities between gene expression profiles through new data transformations. *BMC Bioinformatics* **8** 29.

[18] Kumar, A. and Kannan, R. (2010). Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science—FOCS 2010* 299–308. IEEE Computer Soc., Los Alamitos, CA. MR3025203

[19] Kumar, A., Sabharwal, Y. and Sen, S. (2005). Linear time algorithms for clustering problems in any dimensions. In *Automata, languages and programming. Lecture Notes in Comput. Sci.* **3580** 1374–1385. Springer, Berlin. MR2184726

[20] Levrard, C. (2013). Fast rates for empirical vector quantization. *Electron. J. Statist.* **7** 1716–1746.

[21] Levrard, C. (2015). Nonasymptotic bounds for vector quantization in Hilbert spaces. *Ann. Statist.* **43** 592–619.

[22] Levrard, C. (2018). Sparse oracle inequalities for variable selection via regularized quantization. *Bernoulli* **24** 271–296.

[23] Linder, T. (2002). *Learning-Theoretic Methods in Vector Quantization* In *Principles of Nonparametric Learning* 163–210. Springer Vienna, Vienna.

[24] Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–137. MR651807

[25] Lu, Y. and Zhou, H. H. (2016). Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants. *ArXiv e-prints*.

[26] Mahajan, M., Nimbhorkar, P. and Varadarajan, K. (2012). The planar k-means problem is NP-hard. *Theoret. Comput. Sci.* **442** 13–21. MR2927097

[27] Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618 (2001i:62074)

[28] Ng, A. Y., Jordan, M. I. and Weiss, Y. (2001). On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* 849–856. MIT Press.

[29] Orhan, U., Hekim, M. and Ozer, M. (2011). EEG signals classification using the K-means clustering and a multilayer perceptron neural network

model. *Expert Syst. Appl.* **38** 13475-13481.
[30] Ostrovsky, R., Rabani, Y., Schulman, L. J. and Swamy, C. (2012). The effectiveness of Lloyd-type methods for the k-means problem. *J. ACM* **59** Art. 28, 22. MR3008400
[31] Pollard, D. (1982). A central limit theorem for k-means clustering. *Ann. Probab.* **10** 919–926. MR672292 (84c:60047)
[32] Tang, C. and Monteleoni, C. On Lloyd's Algorithm: New Theoretical Insights for Clustering in Practice. Supplementary material.
[33] 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.
[34] Tavazoie, S., Hugues, J. D., Campbell, M. J., Cho, R. J. and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature genetics* **22** 281-5.
[35] Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*, 1st ed. Springer Publishing Company, Incorporated.