



HAL
open science

Sparse Oracle Inequalities for Variable Selection via Regularized Quantization

Clément Levrard

► **To cite this version:**

Clément Levrard. Sparse Oracle Inequalities for Variable Selection via Regularized Quantization. 2015. hal-01005545v2

HAL Id: hal-01005545

<https://hal.science/hal-01005545v2>

Preprint submitted on 15 Apr 2015 (v2), last revised 6 Jul 2016 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sparse oracle inequalities for variable selection via regularized quantization

Clément Levrard

Abstract

We give oracle inequalities on procedures which combines quantization and variable selection via a weighted Lasso k -means type algorithm. The results are derived for a general family of weights, which can be tuned to size the influence of the variables in different ways. Moreover, these theoretical guarantees are proved to adapt the corresponding sparsity of the optimal codebooks, if appropriate. Even if there is no sparsity assumption on the optimal codebooks, our procedure is proved to be close to a sparse approximation of the optimal codebooks, as has been done for the Generalized Linear Models in regression. If the optimal codebooks have a sparse support, we also show that this support can be asymptotically recovered, giving an asymptotic upper bound on the probability of misclassification. These results are illustrated with Gaussian mixture models in arbitrary dimension with sparsity assumptions on the means, which are standard distributions in model-based clustering.

1 Introduction

Let P be a distribution over \mathbb{R}^d . Quantization is the problem of replacing P with a finite set of points, without losing too much information. To be more precise, if k denotes an integer, a k points quantizer Q is defined as a map from \mathbb{R}^d into a finite subset of \mathbb{R}^d with cardinality k . In other words, a k -quantizer divide \mathbb{R}^d into k groups, and assigns each group a representative, providing both a compression and a classification scheme for the distribution P .

The quantization theory was originally developed as a way to answer signal compression issues in the late 40's (see, e.g., [GG91]). However, unsupervised classification is also in the scope of its application. Isolating meaningful groups from a cloud of data is a topic of interest in many fields, from social science to biology.

Assume that P has a finite second moment, and let Q be a k points quantizer. The performance of Q in representing P is measured by the distortion

$$R(Q) = \int \|x - Q(x)\|^2 dP,$$

where $\int f$ means integration of f with respect to P . It is worth pointing out that many other distortion functions can be defined, using $\|x - Q(x)\|^r$ or more

general distance functions (see, e.g., [Fis10] or [GL00]). However, the choice of the Euclidean squared norm is convenient, since it allows to fully take advantage of the Euclidean structure of \mathbb{R}^d , as described in [Lev15]. Moreover, from a practical point of view, the k -means algorithm (see [Llo82]) is designed to minimize this squared-norm distortion and can be easily implemented.

Since the distortion is based on the Euclidean distance between a point and its image, it is well known that only nearest-neighbor quantizers are to be considered (see, e.g., [GL00] or [Pol82]). These quantizers are quantizers of the type $x \mapsto \operatorname{argmin}_{j=1,\dots,k} \|x - c_j\|$, where the c_i 's are elements of \mathbb{R}^d and are called code points. A vector of code points (c_1, \dots, c_k) is called a codebook, so that the distortion takes the form

$$R(\mathbf{c}) = P \min_{j=1,\dots,k} \|x - c_j\|^2.$$

It has been proved in [Pol81] that, whenever $P\|x\|^2 < \infty$, there exists optimal codebooks, denoted by \mathbf{c}^* .

Let X_1, \dots, X_n denote an independent and identically distributed sample drawn from P , and denote by P_n the associated empirical distribution, namely, for every measurable subset A , $P_n(A) = 1/n |\{i | X_i \in A\}|$. The aim is to design a codebook from this n -sample, whose distortion is as close as possible to the optimum $R(\mathbf{c}^*)$. The k -means algorithm provides the empirical codebook $\hat{\mathbf{c}}_n$, defined by

$$\hat{\mathbf{c}}_n = \operatorname{argmin} \frac{1}{n} \sum_{i=1}^n \min_{j=1,\dots,k} \|X_i - c_j\|^2 = \operatorname{argmin} P_n \min_{j=1,\dots,k} \|x - c_j\|^2.$$

Unfortunately, if $P^{(p)} \neq 0$, where $P^{(p)}$ denotes the marginal distribution of P on the p -th coordinate, then $\hat{\mathbf{c}}_n^{(p)} = (\hat{c}_1^{(p)}, \dots, \hat{c}_k^{(p)})$ may not be zero, even if the p -th coordinate has no influence on the classification provided by the k -means. For instance, if $\mathbf{c}^{*,(p)} = 0$, and $P^{(p)}$ has a density, then $\hat{\mathbf{c}}_n^{(p)} \neq 0$ almost surely. This suggests that the k -means algorithm does not provide sparse codebooks, even in the case where some variables plays no role in the classification, which can be detrimental to the computational tractability and to the interpretation of the corresponding clustering scheme in high-dimensional settings.

Consequently, when d is large, a variable selection procedure is usually performed preliminary to the k -means algorithm. The variable selection can be achieved using penalized BCCS strategies, as exposed in [CWLX14] or [WT10]. Though these procedures offer good performance in classifying the training sample X_1, \dots, X_n , under the assumption that the marginal distributions $P^{(p)}$ are independent, no theoretical result on the prediction performance has been given. An other way to perform variable selection can be to select coordinates whose empirical variances are larger than a determined ratio of the global variance, following the idea of [SB08]. This algorithm has shown good results on practical examples, such as curve clustering (see, e.g., [ABCP13]). However, there is no theoretical result on the prediction performance of the selected coordinates.

Algorithms combining variable selection through PCA and clustering via k -means, like RKM (Reduced k -means, introduced in [DSC94]) and FKM (Factorial k -means, introduced in [VK01]), are also very popular in practice. Some results on the performance in classifying the sample X_1, \dots, X_n have been derived in [TCKV10] under strong conditions on P . In addition, some asymptotic prediction results on these procedures have been established in [Ter14] and [Ter15], showing that both the resulting codebook and its distortion converge almost surely to respectively a minimizer of the distortion constrained on a lower-dimensional subspace of \mathbb{R}^d and the distortion of the latter, following the approach of [Pol81]. However, these methods could be insuitable for interpreting which variables are relevant for the clustering. In addition, no bounds on the excess distortion are available to our knowledge, and the choice of the dimension of the reduction space remains a hard issue, tackled in our procedure by a L_1 -type penalization.

In fact, excess risk bounds for procedures combining dimensionality reduction and clustering are mostly to be found in the model-based clustering literature (see, e.g., [MM13] for a L_0 -type penalization method, and [Mey13] for a L_1 -type penalization method). This approach, consisting in modeling P via a Gaussian mixture with sparse means through density estimation via constrained Maximum Likelihood Estimators, is clearly connected to ours. In fact, most of the derivation for the oracle inequalities stated in this paper use the same tools, drawn from empirical process theory. Nevertheless, no results on the convergence of the estimated means (i.e., model consistency) have been derived in this framework, and this model-based approach theoretically fails when P is not continuous, unlike k -means one (see, e.g., [Lev15]).

This paper exposes a theoretical study of a weighted Lasso type procedure adapted to k -means, as suggested in [SWF12]. Results are given for a general family of weights, encompassing the weights proposed in [SWF12] as well as those proposed in [vdG08] in a Generalized Linear Models for regression setting. To be more precise, we provide non-asymptotic excess distortion bounds along with model consistency results, under weaker conditions than ones required in [SWF12] (for instance, the coordinates are not assumed to be independent), and adapting the sparsity of the optimal codebooks. Interestingly, the excess distortion bounds are valid in the case where it may exist several optimal codebooks, contrary to results in [SWF12] and [vdG08]. These results are illustrated with Gaussian mixture distributions, often encountered in model-based clustering literature, showing at the same time that optimal codebooks can be proved to be unique for this type of distributions, under some conditions on the variances of the components of the mixture.

The paper is organized as follows. Some notation are introduced in Section 2, along with the Lasso k -means procedure and the different assumptions. The consistency and prediction results are gathered in Section 3, and the proof of these results are exposed in Section 4. At last, the proofs of some auxiliary results are given in the Appendix section.

2 Notation

Let x be in \mathbb{R}^d , then the p -th coordinate of x will be denoted by $x^{(p)}$. Throughout this paper, it is assumed that, for every $p = 1, \dots, d$, there exists a sequence M_p , such that $|x^{(p)}| \leq M_p$ P -almost surely. In other words P is assumed to have bounded marginal distributions $P^{(p)}$. To shorten notation, the Euclidean coordinate-wise product $\prod_{p=1}^d [-M_p, M_p]$ will be denoted by C . To frame quantization as a contrast minimization issue, let us introduce the following contrast function

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

where $\mathbf{c} = (c_1, \dots, c_k)$ denotes a codebook, that is a kd -dimensional vector. The risk $R(\mathbf{c})$ then takes the form $R(\mathbf{c}) = R(Q) = P\gamma(\mathbf{c}, \cdot)$, where we recall that Pf denotes the integration of the function f with respect to P . Similarly, the empirical risk $\hat{R}_n(\mathbf{c})$ can be defined as $\hat{R}_n(\mathbf{c}) = P_n\gamma(\mathbf{c}, \cdot)$, where P_n is the empirical distribution associated with X_1, \dots, X_n , in other words $P_n(A) = 1/n |\{i | X_i \in A\}|$, for every measurable subset $A \subset \mathbb{R}^d$. The usual k -means codebook $\hat{\mathbf{c}}_n$ is then defined as a minimizer of $\hat{R}_n(\mathbf{c})$.

It is worth pointing out that, since the support of P is bounded, then there exist such minimizers $\hat{\mathbf{c}}_n$ and \mathbf{c}^* (see, e.g., Corollary 3.1 in [Fis10]). In the sequel, the set of minimizers of the risk $R(\cdot)$ will be denoted by \mathcal{M} . Then, for any codebook \mathbf{c} , the loss $\ell(\mathbf{c}, \mathbf{c}^*)$ may be defined as the excess distortion, namely $\ell(\mathbf{c}, \mathbf{c}^*) = R(\mathbf{c}) - R(\mathbf{c}^*)$, for \mathbf{c}^* in \mathcal{M} .

Let c_1, \dots, c_k be a sequence of code points. A central role is played by the set of points which are closer to c_i than to any other c_j 's. To be more precise, the Voronoi cell, or quantization cell associated with c_i is the closed set defined by

$$V_i(\mathbf{c}) = \{x \in \mathbb{R}^d \mid \forall j \neq i \quad \|x - c_i\| \leq \|x - c_j\|\}.$$

It may be noted that $(V_1(\mathbf{c}), \dots, V_k(\mathbf{c}))$ does not form a partition of \mathbb{R}^d , since $V_i(\mathbf{c}) \cap V_j(\mathbf{c})$ may be non empty. To address this issue, the Voronoi partition associated with \mathbf{c} is defined as the sequence of subsets $W_i(\mathbf{c}) = V_i(\mathbf{c}) \setminus (\cup_{i>j} V_j(\mathbf{c}))$, for $i = 1, \dots, k$. It is immediate that the $W_i(\mathbf{c})$'s form a partition of \mathbb{R}^d , and that for every $i = 1, \dots, k$,

$$\bar{W}_i(\mathbf{c}) = V_i(\mathbf{c}),$$

where $\bar{W}_i(\mathbf{c})$ denotes the closure of the subset $W_i(\mathbf{c})$. The open Voronoi cell is defined the same way by

$$\overset{\circ}{V}_i(\mathbf{c}) = \{x \in \mathbb{R}^d \mid \forall j \neq i \quad \|x - c_i\| < \|x - c_j\|\},$$

and the following inclusion holds, for i in $\{1, \dots, k\}$,

$$\overset{\circ}{V}_i(\mathbf{c}) \subset W_i(\mathbf{c}) \subset V_i(\mathbf{c}).$$

The risk $R(\mathbf{c})$ then takes the form

$$R(\mathbf{c}) = \sum_{i=1}^k P(\|x - c_i\|^2 1_{W_i(\mathbf{c})}(x)),$$

where 1_A denotes the indicator function associated with A . In the case where $P(W_i(\mathbf{c})) \neq 0$, for every $i = 1, \dots, k$, it is clear that

$$P(\|x - c_i\|^2 1_{W_i(\mathbf{c})}(x)) \geq P(\|x - \eta_i\|^2 1_{W_i(\mathbf{c})}(x)),$$

with equality only if $c_i = \eta_i$, where η_i denotes the conditional expectation of P over the subset $W_i(\mathbf{c})$, that is

$$\eta_i = \frac{P(x 1_{W_i(\mathbf{c})}(x))}{P(W_i(\mathbf{c}))}.$$

Moreover, it is proved in Proposition 1 of [GLP07] that, for every Voronoi partition $W(\mathbf{c}^*)$ associated with an optimal codebook \mathbf{c}^* , and every $i = 1, \dots, k$, $P(W_i(\mathbf{c}^*)) \neq 0$. Consequently, any optimal codebook satisfies the so-called centroid condition (see, e.g., Section 6.2 of [GG91]), that is

$$\mathbf{c}_i^* = \frac{P(x 1_{W_i(\mathbf{c}^*)}(x))}{P(W_i(\mathbf{c}^*))}.$$

As a remark, the centroid condition ensures that $\mathcal{M} \subset C^k$, and, for every \mathbf{c}^* in \mathcal{M} , $i \neq j$,

$$\begin{aligned} P(V_i(\mathbf{c}^*) \cap V_j(\mathbf{c}^*)) &= P(\{x \in \mathbb{R}^d \mid \forall i' \quad \|x - c_i^*\| = \|x - c_j^*\| \leq \|x - c_{i'}^*\|\}) \\ &= 0. \end{aligned}$$

A proof of this statement can be found in Proposition 1 of [GLP07]. According to [Lev15], for every \mathbf{c}^* in \mathcal{M} , the following set is of special interest:

$$N_{\mathbf{c}^*} = \bigcup_{i \neq j} V_i(\mathbf{c}^*) \cap V_j(\mathbf{c}^*).$$

To be more precise, the key quantity is the margin function, which is defined as

$$p(t) = \sup_{\mathbf{c}^* \in \mathcal{M}} P(N_{\mathbf{c}^*}(t)),$$

where $N_{\mathbf{c}^*}(t)$ denotes the t -neighborhood of $N_{\mathbf{c}^*}$. As shown in [Lev15], bounds on this margin function (see Assumption 2 below) can provide interesting results on the convergence rate of the k -means codebook, along with basic properties of optimal codebooks.

In order to perform both variable selection and quantization, we introduce the Lasso k -means codebook $\hat{\mathbf{c}}_{n,\lambda}$ as follows.

$$\hat{\mathbf{c}}_{n,\lambda} \in \operatorname{argmin}_{\mathbf{c} \in C^k} P_n \gamma(\mathbf{c}, \cdot) + \lambda I_{\hat{w}}(\mathbf{c}), \quad (1)$$

where \hat{w} is a possibly random sequence of weights of size d , and $I_{\hat{w}}(\cdot)$ denotes the penalty function

$$I_{\hat{w}}(\mathbf{c}) = \sum_{p=1}^d \hat{w}_p \|\mathbf{c}^{(p)}\|. \quad (2)$$

Let us recall here that $\mathbf{c}^{(p)}$ denote the vector $(c_1^{(p)}, \dots, c_k^{(p)})$ made of the p -th coordinates of the different codepoints. The results exposed in the following section are illustrated with three sequences of weights, corresponding to different codebooks: the *plain Lasso* codebook, defined by the deterministic sequence $\hat{w}_p = 1$, the *normalized Lasso* codebook, defined by $\hat{w}_p = \hat{\sigma}_p$, and the *threshold Lasso* codebook, which is a slight modification of the original Lasso-type procedure mentioned in [SWF12] and is defined by $\hat{w}_p = 1/(\delta \vee \|\hat{\mathbf{c}}_n^{(p)}\|)$, where $\hat{\mathbf{c}}_n$ denotes the k -means codebook and δ a parameter to be tuned. It is likely that other families of weights may be of special interest, for instance combining normalization and threshold. Consequently the results are derived for an arbitrary family of weights satisfying some convergence conditions.

These L_1 -type penalties have been designed to drive the irrelevant (p)-th coordinates $c_1^{(p)}, \dots, c_k^{(p)}$ together to zero (see, e.g., [Bac08]), according to different criterions. To describe the influence of the different coordinates, the following notation are adopted. Let $S \subset \{1, \dots, d\}$ denote a subset of coordinates, then for any vector x in $(\mathbb{R}^d)^\ell$ and set $A \subset (\mathbb{R}^d)^\ell$, ℓ being a positive integer, x_S will denote the vector in $(\mathbb{R}^{|S|})^\ell$ corresponding to the coefficients of x on variables in S , and A_S will denote the set of such x_S , for x in A . Moreover, let P^S denote the marginal distribution of P over the set $\mathbb{R}^{|S|}$. We may then define the restricted distortions and variances as follows:

$$\begin{cases} \sigma_S^2 &= P^S \|x\|^2, \\ \hat{\sigma}_S^2 &= P_n^S \|x\|^2, \\ R_S^* &= \min_{\mathbf{c} \in C_S} P^S \gamma(\mathbf{c}, \cdot), \\ \hat{R}_S &= \min_{\mathbf{c} \in C_S} P_n^S \gamma(\mathbf{c}, \cdot), \end{cases}$$

where the vector x is element of $\mathbb{R}^{|S|}$. Elementary properties of the distortion show that, if $S = S_1 \cup S_2$, with empty intersection, then

$$\begin{cases} \sigma_S^2 &= \sigma_{S_1}^2 + \sigma_{S_2}^2, \\ \hat{\sigma}_S^2 &= \hat{\sigma}_{S_1}^2 + \hat{\sigma}_{S_2}^2, \\ R_S^* &\geq R_{S_1}^* + R_{S_2}^*, \\ \hat{R}_S &\geq \hat{R}_{S_1} + \hat{R}_{S_2}. \end{cases} \quad (3)$$

These elementary properties will be of importance when choosing which coordinate to select. A special attention will be paid to the subsets of variables formed by the support of codebooks. To be more precise, for every codebook \mathbf{c} in C^k , we define the support $S(\mathbf{c})$ of \mathbf{c} by $S(\mathbf{c}) = \{j \in \{1, \dots, d\} \mid \mathbf{c}^{(j)} \neq 0\}$. The following Proposition gives a first glance at which variables are in $S(\hat{\mathbf{c}}_{n,\lambda})$.

Proposition 2.1. *Let p be in $\{1, \dots, d\}$. If*

$$\sqrt{\hat{\sigma}_p^2 - \hat{R}_p} < \frac{\hat{w}_p \lambda}{2},$$

then

$$\hat{\mathbf{c}}_{n,\lambda}^{(p)} = (\hat{c}_{n,\lambda,1}^{(p)}, \dots, \hat{c}_{n,\lambda,k}^{(p)}) = (0, \dots, 0).$$

According to Proposition 2.1, the Lasso k -means procedures may be thought of as a multimodularity test on every coordinate, in the spirit of [JW14]. This result ensures that, if the distortion of the codebook $(0, \dots, 0)$ is close to the optimal empirical distortion, on the p -th coordinate, then the Lasso k -means will drive the p -th variable to 0. For the plain Lasso, the differences $\sqrt{\hat{\sigma}_p^2 - \hat{R}_p}$ are uniformly thresholded, whereas for the normalized Lasso, the threshold in λ is applied on the ratios $\hat{R}_p / \hat{\sigma}_p^2$. This point suggests that the normalized Lasso may succeed in recovering informative variables with small ranges.

We introduce now the assumptions which will be required to derive theoretical results on the performance of the Lasso codebooks. To deal with the case of possibly several optimal codebooks, we introduce the following structural assumption on P .

Assumption 1. *For every \mathbf{c}^* in \mathcal{M} and \mathbf{c} in C^k , if $S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)$, then $R(\mathbf{c}) > R(\mathbf{c}^*)$.*

Assumption 1 roughly requires that no optimal codebook has a support strictly contained in the support of another optimal codebook. This is obviously the case if P has a unique optimal codebook, up to relabeling.

Assumption 2 (Margin Condition). *There exists $r_0 > 0$ such that*

$$\forall t \leq r_0 \quad p(t) \leq c_0(P)t, \tag{4}$$

where $c_0(P)$ is a fixed constant, defined in [Lev15].

As exposed in [Lev15], Assumption 2 may be thought of as a margin condition for squared distance based quantization. Some examples of distributions satisfying (4) are given in [Lev15], including Gaussian mixtures under some conditions. Roughly, if P is well concentrated around k poles, then (4) will hold. It is also worth mentioning that the condition required in [SWF12] seems stronger than the condition required in Definition 2, since it requires P to have a unique optimal codebook, to be a mixture of components centered on the different optimal code points, and that the Hessian matrix of the risk function located at the optimal codebook is positive definite.

Moreover, Assumption 2 is a sufficient condition to ensure that some elementary properties that are often assumed are satisfied, as described in the following Proposition.

Proposition 2.2. *If P satisfies Assumption 2, then*

i) \mathcal{M} is finite,

ii) there exists $\kappa'_0 > 0$ such that, for every \mathbf{c} in C^k , $\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2 \leq \kappa'_0 \ell(\mathbf{c}, \mathbf{c}^*)$,

where $\mathbf{c}^*(\mathbf{c}) \in \arg\min_{\mathbf{c}^*} \|\mathbf{c} - \mathbf{c}^*\|$.

Moreover, if P satisfies Assumption 1, then there exists a constant κ''_0 such that, for every \mathbf{c}^* in \mathcal{M} and $S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)$, we have

$$\|\mathbf{c} - \mathbf{c}^*\|^2 \leq \kappa''_0 \ell(\mathbf{c}, \mathbf{c}^*).$$

The two first statements of Proposition 2.2 are to be found in Proposition 2.2 of [Lev15], the proof of the third statement is given in Section 4.2. Proposition 2.2 may be thought of as a generalization of the positive Hessian matrix condition of [Pol82] to the non-continuous case. It also allows to deal with the case where P has several optimal codebooks. In the following, we denote by κ_0 the quantity $\kappa'_0 \vee \kappa''_0$, whenever Assumption 2 and Assumption 1 are satisfied.

In addition to Assumption 2, we assume that the weights \hat{w}_p satisfy a uniform concentration inequality around some deterministic weights, as stated below.

Assumption 3 (Weights concentration). *There exists a sequence of deterministic weights $w_p > 0$, $p = 1, \dots, d$, and a constant $\kappa_1 < 1$ such that*

$$\mathbb{P} \left(\sup_{p=1, \dots, d} \left| \frac{\hat{w}_p}{w_p} - 1 \right| > \kappa_1 \right) := r_1(n) \xrightarrow{n \rightarrow \infty} 0. \quad (5)$$

Assumption 3 is obviously satisfied for the plain Lasso ($\hat{w}_p = 1$). The following proposition ensures that this statement remains true for the two other examples of weights. For any sequence w_p , we denote by $T(w)$ the quantity $\sup_{p=1, \dots, d} M_p/w_p$. With a slight abuse of notation, $T(\sigma)$ and $T(\delta)$ will refer to the sequences σ_p and $1/(\|\mathbf{c}^{*(p)}\| \vee \delta)$, where the latter is well defined when P has a unique optimal codebook.

Proposition 2.3.

For $\hat{w}_p = \hat{\sigma}_p$, if $1 > \kappa_1 > \frac{T^2(\sigma)\sqrt{\log(d)}}{\sqrt{2n}}$, then Assumption 3 holds with $w_p = \sigma_p$ and $r_1(n) = e^{-\left(\frac{\sqrt{2n}\kappa_1}{T^2(\sigma)} - \sqrt{\log(d)}\right)^2}$.

For $\hat{w}_p = 1/(\|\hat{\mathbf{c}}_n^{(p)}\| \vee \delta)$, let M be defined as $M = \sqrt{M_1^2 + \dots + M_d^2}$. If $1 > \kappa_1 > C_0 \frac{M\sqrt{k}}{\sqrt{n\delta}}$, for a fixed constant C_0 , Assumption 2 is satisfied, and \mathbf{c}^* is unique (up to relabeling), then Assumption 3 holds with $w_p = 1/(\|\mathbf{c}^{*(p)}\| \vee \delta)$ and $r_1(n) = e^{-\left(\frac{\sqrt{n\delta}\kappa_1}{C_0 M} - \sqrt{k}\right)^2}$.

The proof of Proposition 2.3 follows from standard concentration inequalities, and can be found in the Section 5.1 of the Appendix. At first sight, the assumption that \mathbf{c}^* is unique seems quite restrictive. However, as exposed in Section 3.3, it can be shown that Gaussian mixtures satisfy this property, provided that the variances of the components are small enough. In fact, if P has

several optimal codebooks, there is no intuition about toward which one $\hat{\mathbf{c}}_n$ will converge, hence the difficulty of defining deterministic limit weights for \hat{w}_p .

At last, we define the following quantities λ_0 and λ_1 which will play the role of minimal values for the regularization parameter λ , as exposed in [vdG08].

$$\begin{cases} \lambda_0 &= 16\sqrt{2\pi}\sqrt{\frac{k\log(kd)}{n}}T(w), \\ \lambda_1(x) &= e\lambda_0\left(1 + \sqrt{\frac{u+x}{k\log(kd)}}\right), \end{cases} \quad (6)$$

where $x > 0$ and $u = \log\left(\frac{\|w\|_2^2\sqrt{n}}{\sqrt{\log(kd)}}\right)$. These two quantities come from empirical process theory, their roles are explained in Section 4. Roughly, λ_0 is the minimal value of the regularization parameter which ensures that the empirical risk is close to the true risk uniformly on C^k , and $\lambda_1(x)$ is the minimal value which ensures that the deviation between empirical and true risk may be compared to the norm I_w uniformly on C^k .

3 Results

3.1 Oracle inequalities for the Lasso k -means

Following the approach of [MM11], Lasso type procedures may be thought of as model selection procedures over L_1 balls. Theorem 3.1 below is the adaptation of this idea for the Lasso k -means procedures.

Theorem 3.1. *Suppose that Assumption 3 is satisfied, for some constant $\kappa_1 < 1$, and choose*

$$\lambda \geq \frac{\lambda_1(x)}{1 - \kappa_1},$$

for some $x > 0$, where λ_1 is defined in (6). Then, with probability larger than $1 - r_1(n) - e^{-x}$, we have

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \inf_{r>0} \inf_{I_w(\mathbf{c}) \leq r} (\ell(\mathbf{c}, \mathbf{c}^*) + (3 - \kappa_1)\lambda(r \vee \lambda_0)).$$

A direct implication of Theorem 3.1 is that $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq 4\lambda(I_w(\mathbf{c}^*) \vee \lambda_0)$. Hence, choosing $\lambda \sim \lambda_1(x)$ gives a convergence rate for $\ell(\mathbf{c}, \mathbf{c}^*)$ of $T(w)/\sqrt{n}$, up to a $\log(n)$ factor. If $T(w)$ is fixed, i.e. does not depend on n , this rate is roughly the same as the rate of convergence of the k -means codebook without margin assumption, as shown in [BDL08]. It is the case for the plain and normalized Lasso, but also for the threshold Lasso if δ is chosen as $n^{-\alpha}$, $\alpha > 0$ for instance. Indeed, for n large enough, $T(n^{-\alpha}) = \inf_{\{p|\mathbf{c}^{*(p)} \neq 0\}} M_p / \|\mathbf{c}^{*(p)}\|$ is a fixed quantity. Since $I_w(\mathbf{c}^*) = |S(\mathbf{c}^*)|$ in this case, we get $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \lesssim |S(\mathbf{c}^*)|\lambda \sim |S(\mathbf{c}^*)|/\sqrt{n}$, up to a $\log(n)$ factor, showing that the distortion of the threshold Lasso adapts the sparsity of \mathbf{c}^* when n is large enough.

However, if P satisfies Assumption 2 and Assumption 1 in addition, further results may be derived, following the approach of [vdG08]. To this aim, we

defined, for a fixed codebook \mathbf{c}^* and a weight family w , the set of w -sparse approximations of \mathbf{c}^* at order λ by

$$\mathcal{M}_\lambda(\mathbf{c}^*) = \left\{ \operatorname{argmin}_{S(\mathbf{c}) \subset S(\mathbf{c}^*)} 3R(\mathbf{c}) + 8\kappa_0\lambda^2 \|w_{S(\mathbf{c})}\|^2 \right\},$$

where $\kappa_0 = \kappa'_0 \vee \kappa''_0$, as defined below Proposition 2.2. Then, the closest w -sparse approximation of \mathbf{c}^* may be defined as $\mathbf{c}_\lambda^*(\mathbf{c}^*) \in \operatorname{argmin}_{\mathbf{c} \in \mathcal{M}_\lambda(\mathbf{c}^*)} \|\mathbf{c} - \mathbf{c}^*\|$. With a slight abuse of notation, the w -sparse approximation of a codebook \mathbf{c} is defined by $\mathbf{c}_\lambda^*(\mathbf{c}) = \mathbf{c}_\lambda^*(\mathbf{c}^*(\mathbf{c}))$. It is immediate that, for the plain Lasso, $\|w_{S(\mathbf{c})}\| = |S(\mathbf{c})| = \|\mathbf{c}\|_0$, whereas for the normalized Lasso, $\|w_{S(\mathbf{c})}\| = \sigma_{S(\mathbf{c})}$. For the threshold Lasso, $\|w_{S(\mathbf{c})}\|$ has the slightly more intricate expression

$$\|w_{S(\mathbf{c})}\|^2 = \frac{1}{\delta^2} \left| S(\mathbf{c}) \cap \{j \mid \|\mathbf{c}^{*,(j)}\| \leq \delta\} \right| + \sum_{S(\mathbf{c}) \cap \{j \mid \|\mathbf{c}^{*,(j)}\| > \delta\}} \frac{1}{\|\mathbf{c}^{*,(j)}\|^2}.$$

However, it is easy to see that $\|w_{S(\mathbf{c})}\| \leq |S(\mathbf{c})|/\delta$. If we assume that the optimal codebooks \mathbf{c}^* are sparse, some guarantees on the support of the $\mathbf{c}_\lambda^*(\mathbf{c}^*)$'s may be given. To be more precise, the following subset of variables is introduced

$$(S^+)^c = \bigcap_{\mathbf{c}^* \in \mathcal{M}} S(\mathbf{c}^*)^c.$$

S^+ may be thought of as the generalized support over optimal codebooks, extending the definition of these sets from the unique optimal codebook case. If P has a unique optimal codebook, it is immediate that $S^+ = S(\mathbf{c}^*)$. However, even if all the codebook are sparse, S^+ may not be sparse. For instance, if $d = k = 2$ and P is a pointwise distribution with support $(-1, -1), (-1, 1), (1, -1), (1, 1)$ equally weighted, then every optimal codebook has at least one zero coordinate, whereas $S^+ = \{1, 2\}$.

From its definition, it is straightforward that $S(\mathbf{c}_\lambda^*(\mathbf{c}^*)) \subset S^+$, for \mathbf{c}^* in \mathcal{M} . Nevertheless, in the case where $S^+ = \{1, \dots, d\}$, the $\mathbf{c}_\lambda^*(\mathbf{c}^*)$'s may still have zero coordinates. In fact, $\mathbf{c}_\lambda^*(\mathbf{c}^*)$ may be thought of as tradeoff between distortion and size of the support, the latter being measured by $\|w\|_S^2$. As in the empirical case of Proposition 2.1, this tradeoff property may be illustrated in the following way.

Proposition 3.1. *Let p be in $\{1, \dots, d\}$. If*

$$\sigma_p^2 - R_p^* < \frac{8\lambda^2\kappa_0 w_p^2}{3},$$

then, for every \mathbf{c}^ in \mathcal{M} ,*

$$\mathbf{c}_\lambda^{*,(p)}(\mathbf{c}^*) = (0, \dots, 0).$$

The proof of Proposition 3.1 is given in Section 4.4. Proposition 3.1, as well as Proposition 2.1, may be thought of as a comparison between the risk of

optimal codebooks and the risk of the codebook $\mathbf{0}$, on the p -th variable. It is worth noticing that, for the plain Lasso and threshold Lasso, the comparison is based on the difference $\sigma_p^2 - R_p^*$, whereas for the normalized Lasso only the ratio R_p^*/σ_p^2 is to be considered. Once more, this point suggests that the sparse w -approximation may not be sensitive to coordinate-wise dilations in this case. We are now in position to state sharper oracle results, on both the excess distortion and the convergence of the Lasso k -means codebooks.

Theorem 3.2. *Suppose that Assumption 1, Assumption 2 and Assumption 3 are satisfied. If*

$$\lambda \geq \frac{2\lambda_1(x)}{1 - \kappa_1},$$

where λ_1 is defined in (6), then, with probability larger than $1 - r_1(n) - e^{-x}$, we have

$$\begin{aligned} \ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + \lambda(1 - \kappa_1)I_w(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}_\lambda^*(\hat{\mathbf{c}}_{n,\lambda})) &\leq \left[3\ell(\mathbf{c}_\lambda^*, \mathbf{c}^*) \right. \\ &\left. + 8\kappa_0\lambda^2 \|w_{S(\mathbf{c}_\lambda^*(\hat{\mathbf{c}}_{n,\lambda}))}\|^2 \vee 3\lambda\lambda_0 \right]. \end{aligned} \quad (7)$$

A consequence of Theorem 3.2 is $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq 8\kappa_0\lambda^2 \|w_{S(\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}))}\|^2 \vee 3\lambda\lambda_0$, which provides an oracle inequality adapting the sparsity of the $\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda})$'s. For instance, considering the plain Lasso, provided that $\mathbf{c}^* \neq \mathbf{0}$ for some \mathbf{c}^* , (7) leads to $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq 8(\kappa_0 \vee 1)|S(\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}))|\lambda^2 \leq 8(\kappa_0 \vee 1)|S^+|\lambda^2$. However, Theorem 3.2 also deals with the case where the \mathbf{c}^* 's are not sparse, comparing the Lasso k -means codebook $\hat{\mathbf{c}}_{n,\lambda}$ to the closest sparse w -approximations, for which Proposition 3.1 yields a reduced support whenever λ is large enough.

Theorem 3.2 may be considered as an application of Theorem 2.1 in [vdG13] to the k -means case, with a slight improvement in the analysis of the complexity term (see Section 5.3 in the Appendix for more details). The numerical constants in Theorem 3.2 have been arbitrarily fixed for clarity sakeness, note however that a more general version of Theorem 3.2 can be derived the same way as Theorem 2.1 in [vdG13].

At last, it is worth pointing out that $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq 8\kappa_0\lambda^2 \|w_{S(\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}))}\|^2 \vee 3\lambda\lambda_0$ provides a convergence rate in $1/n$, up to a $\log(n)$ factor, when $\lambda \sim \lambda_1$ and w does not depend on n . Interestingly, this rate is the convergence rate of the k -means codebook $\hat{\mathbf{c}}_n$, when P satisfies a margin condition, as described in [Lev15]. For the plain Lasso, if $S(\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda})) = \{1, \dots, d\}$, this inequality may even be written as $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \kappa_0 \frac{kM^2}{n} \log(n) \log(kd)$, where M^2 is the radius of the smallest Euclidean ball containing C . Since $\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq C_0\kappa_0 \frac{kM^2}{n}$, according to Theorem 3.1 in [Lev15], the distortions of the Lasso k -means and the k -means codebooks may be compared when $n \gg \log(kd)$.

Note that for the threshold Lasso, the choice $\delta \sim n^{-\alpha}$, $0 \leq \alpha < 1/2$ still leads to convergence rates in $|S(\mathbf{c}^*)|\log(n)/n$, when \mathbf{c}^* is unique. This result may be then compared with the asymptotic convergence rate of $O(n\lambda^2)$ given in Theorem 1 of [SWF12], which holds with stronger requirements on P and when

$n\lambda$ tends to infinity. In this particular case of a unique optimal codebook, further consistency results may be derived, as described in the following subsection.

3.2 Consistency of the threshold Lasso k -means

Throughout this subsection we assume that there exists a unique optimal codebook \mathbf{c}^* , up to relabeling. Let j be in $S(\mathbf{c}^*)^c$. Then Theorem 2 in [SWF12] established that $\hat{\mathbf{c}}_{n,\lambda}^{(j)} \rightarrow 0$ in probability under strong assumptions on P . To be more precise, it is assumed in [SWF12] that $P|_{V_j^*} = c_j^* + \varepsilon_j$, where $P|_{V_j^*}$ denotes the conditional law of P on the optimal Voronoi cell centered at the j -th optimal code point c_j^* , ε_j has independent coordinates, and the ε_j 's are independent. Theorem 3.3 below gives a generalization of this result, along with a convergence rate for $\mathbb{P}(\hat{\mathbf{c}}_{n,\lambda}^{(j)} \neq 0)$.

Theorem 3.3. *Suppose that Assumption 2 and Assumption 3 are satisfied. For $\hat{w}_p = 1/(\delta \vee \|\hat{\mathbf{c}}_n^{(p)}\|)$, with $\delta = n^{-\alpha}$, $0 < \alpha < 1/2$, choose the regularization parameter as $\lambda \sim n^{-\alpha}$. Then, for every j in $S(\mathbf{c}^*)^c$, we have*

$$\mathbb{P}\left(\hat{\mathbf{c}}_{n,\lambda}^{(j)} \neq 0\right) \underset{n \rightarrow \infty}{=} O\left(e^{-n^{1-2\alpha}}\right). \quad (8)$$

Moreover, for every j in $S(\mathbf{c}^*)$, we have

$$\mathbb{P}\left(\hat{\mathbf{c}}_{n,\lambda}^{(j)} = 0\right) \underset{n \rightarrow \infty}{=} O\left(e^{-n^{1-2\alpha}}\right). \quad (9)$$

The choice $\lambda \sim n^{-\alpha}$ has been made to optimize the convergence rate of the misclassification probabilities. However, this choice leads to suboptimal convergence rate for $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*)$ in Theorem 3.2, which is then of order $n^{-2\alpha}$. In fact, we only need $\lambda \geq \lambda_1(y_n)$, for $\log(n) = o(y_n)$, and λ tending to 0 to ensure that this model consistency result holds. Thus, the choice $\lambda \gtrsim \lambda_1(\log(n)^{(1+\varepsilon)})$, for a positive ε , provides both model consistency and almost optimal convergence of $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*)$.

To apply Theorem 3.3, checking that \mathbf{c}^* is unique remains a hard issue when $d > 1$. The following section gives an example where this problem can be tackled using straightforward consequences of Assumption 2.

3.3 Quasi-Gaussian mixture example

The aim of this section is to illustrate the consistency results exposed in Theorem 3.3 with a typical clustering example, namely Gaussian mixture distributions. In general, a Gaussian mixture distribution \tilde{P} may be defined by its density

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

where \tilde{k} denotes the number of components of the mixture, and the θ_i 's denote the weights of the mixture, which satisfy $\sum_{i=1}^{\tilde{k}} \theta_i = 1$. Moreover, the m_i 's denote

the means of the mixture, so that $m_i \in \mathbb{R}^d$, and the Σ_i 's are the $d \times d$ variance matrices of the components. In this case, we define the active set \tilde{S} as $S(\mathbf{m})$, where \mathbf{m} denotes the codebook with code points the m_i 's. For convenience it is assumed that $\tilde{S} = \{1, \dots, d'\}$, with $d' < d$. We restrict ourselves to the case where the number of components \tilde{k} is known, and match the size k of the codebooks.

Since the support of a Gaussian random variable is not bounded, we define the ‘‘quasi-Gaussian’’ mixture model as follows, truncating each Gaussian component. Let the density f of the distribution P be defined by

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

where N_i denotes a normalization constant for each Gaussian variable. To ensure that this model is close to the Gaussian mixture model, M has to be chosen large enough, say $M \geq 2 \sup_{j=1, \dots, k} \|m_j\|$ for instance. Let σ^2 and σ_-^2 denote the largest and smallest eigenvalues of the Σ_i 's. Then the following proposition states that, provided that σ is small enough, a model consistency result can be derived for the threshold Lasso.

Proposition 3.2. *Assume that $\sigma_- \geq c^- \sigma$, for some constant c^- . Then there exists a constant σ^+ such that, if $\sigma \leq \sigma^+$, then Assumption 2 holds, and \mathbf{c}^* is unique.*

Moreover, if $(\Sigma_i)_{pq} = 0$, for every (i, p, k) in $\{1, \dots, k\} \times \{1, \dots, d'\} \times \{d'+1, \dots, d\}$, then, for every $j \geq d'+1$, the threshold Lasso with exponent α satisfies

$$\mathbb{P} \left(\hat{\mathbf{c}}_{n, \lambda}^{(j)} \neq 0 \right) \underset{n \rightarrow \infty}{=} O \left(e^{-n^{1-2\alpha}} \right).$$

The assumption on the covariance matrices requires that the variable in \tilde{S} are independent from the variables in \tilde{S}^c . As shown in Section 4.7, this strong requirement ensures in fact that the support of the optimal codebook is contained in \tilde{S} . Proposition 3.2 then directly follows from Theorem 3.3.

The first part of Proposition 3.2 extends the results of Proposition 3.2 in [Lev15] to arbitrary dimension d . It also enhances the results of this Proposition, showing that there exists a unique optimal codebook instead of finitely many ones.

It is worth noting that Proposition 3.2 is valid when $k = \tilde{k}$. When k differs from \tilde{k} , Assumption 2 may not be satisfied. For instance, if P is rotationally symmetric and $k, d \geq 2$, then \mathcal{M} cannot be finite, which contradicts Proposition 2.2.

4 Proofs

4.1 Proof of Proposition 2.1

Let W_1, \dots, W_k be the Voronoi partition associated with $\hat{\mathbf{c}}_{n,\lambda}$, and let $L(\hat{\mathbf{c}}_{n,\lambda})$ be the matrix of assignments, defined by

$$L(\hat{\mathbf{c}}_{n,\lambda})_{i,j} = 1_{X_i \in W_j}.$$

Suppose that $\hat{\mathbf{c}}_{n,\lambda}^{(p)} \neq 0$, where $\hat{\mathbf{c}}_{n,\lambda}^{(p)}$ denotes the vector $(\hat{c}_{n,\lambda,1}^{(p)}, \dots, \hat{c}_{n,\lambda,k}^{(p)})^t$, and denote by $X^{(p)}$ the vector $(X_1^{(p)}, \dots, X_k^{(p)})^t$. Then the Karush-Kuhn-Tucker condition, for this penalized empirical risk minimization strategy, ensures that (see, e.g., the proof of Theorem 2 in [SWF12])

$$\frac{-2}{\sqrt{n}} L(\hat{\mathbf{c}}_{n,\lambda})^t \left(X^{(p)} - L(\hat{\mathbf{c}}_{n,\lambda}) \hat{\mathbf{c}}_{n,\lambda}^{(p)} \right) + \sqrt{n} \lambda \frac{\hat{w}_p \hat{\mathbf{c}}_{n,\lambda}^{(p)}}{\|\hat{\mathbf{c}}_{n,\lambda}^{(p)}\|} = 0. \quad (10)$$

Since $L(\hat{\mathbf{c}}_{n,\lambda})^t \left(X^{(p)} - L(\hat{\mathbf{c}}_{n,\lambda}) \hat{\mathbf{c}}_{n,\lambda}^{(p)} \right)$ is the following vector of size k

$$\left(\sum_{X_i \in W_1} (X_i^{(p)} - \hat{c}_{n,\lambda,1}^{(p)}), \dots, \sum_{X_i \in W_k} (X_i^{(p)} - \hat{c}_{n,\lambda,k}^{(p)}) \right),$$

it may be noted that

$$\left\| L(\hat{\mathbf{c}}_{n,\lambda})^t \left(X^{(p)} - L(\hat{\mathbf{c}}_{n,\lambda}) \hat{\mathbf{c}}_{n,\lambda}^{(p)} \right) \right\|^2 = \sum_{j=1}^k n_j^2 (\bar{c}_j^{(p)} - \hat{c}_{n,\lambda,j}^{(p)})^2,$$

where n_j denotes the number of sample vectors X_i 's in W_j , and \bar{c}_j denotes the empirical mean of the sample over the set W_j , that is $\bar{c}_j = \frac{1}{n_j} \sum_{X_i \in W_j} X_i$. Denote by \hat{p}_j the empirical weight of W_j , i.e. $\hat{p}_j = n_j/n$, then

$$\frac{1}{n^2} \left\| L(\hat{\mathbf{c}}_{n,\lambda})^t \left(X^{(p)} - L(\hat{\mathbf{c}}_{n,\lambda}) \hat{\mathbf{c}}_{n,\lambda}^{(p)} \right) \right\|^2 \leq \sum_{j=1}^k \hat{p}_j (\bar{c}_j^{(p)} - \hat{c}_{n,\lambda,j}^{(p)})^2,$$

where $\hat{p}_j \leq 1$ has been used. Let Q_1 be the quantizer which maps W_j to \bar{c}_j , then it is easy to see that

$$\sum_{j=1}^k \hat{p}_j (\bar{c}_j^{(p)} - \hat{c}_{n,\lambda,j}^{(p)})^2 = \hat{R}_p(\hat{\mathbf{c}}_{n,\lambda}) - \hat{R}_p(Q_1),$$

where we recall that $\hat{R}_p(Q) = P_n^{(p)} \|x - Q(x)\|^2$, for any quantizer Q .

Denote by $\hat{\mathbf{c}}_{n,\lambda}^{(-p)}$ the codebook that has p -th coordinate 0 and the same coordinates as $\hat{\mathbf{c}}_{n,\lambda}$ otherwise, and let Q_2 be the quantizer which maps W_j to $\hat{c}_{n,\lambda,j}^{(-p)}$. Then, by definition, $\hat{R}(\hat{\mathbf{c}}_{n,\lambda}) + I_{\hat{w}}(\hat{\mathbf{c}}_{n,\lambda}) \leq \hat{R}(\hat{\mathbf{c}}_{n,\lambda}^{(-p)}) + \lambda I_{\hat{w}}(\hat{\mathbf{c}}_{n,\lambda}^{(-p)})$, and

$\hat{R}(\hat{\mathbf{c}}_{n,\lambda}^{(-p)}) \leq \hat{R}(Q_2)$. Thus, direct calculation leads to $\hat{\sigma}_p^2 \geq \hat{R}_p(\hat{\mathbf{c}}_{n,\lambda})$, according to (3). Since $\hat{R}_p(\hat{\mathbf{c}}_{n,\lambda}) - \hat{R}_p(Q_1) \leq \hat{\sigma}_p^2 - \hat{R}_p$, (10) ensures that

$$\frac{\lambda \hat{w}_p}{2} \leq \sqrt{\hat{\sigma}_p^2 - \hat{R}_p}.$$

4.2 Proof of Proposition 2.2

As mentioned below Proposition 2.2, a proof of the two first statements can be found in the proof of Proposition 2.2 in [Lev15]. The last statement follows from the compactness of $\{\mathbf{c} \in C^k | S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)\}$ and the fact that \mathcal{M} is finite, knowing that $R(\cdot)$ is continuous (see, e.g., Lemma 4.1 in [Lev15]). This ensures that $\inf_{\mathbf{c}^* \in \mathcal{M}} \inf_{S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)} \ell(\mathbf{c}, \mathbf{c}^*) \geq c > 0$, for some constant c , whenever Assumption 1 is satisfied. Since C^k is bounded, $\sup_{\mathbf{c}^* \in \mathcal{M}, S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)} \|\mathbf{c} - \mathbf{c}^*\|^2 / \ell(\mathbf{c}, \mathbf{c}^*)$ is finite.

4.3 Proof of Theorem 3.1

We recall here that $T(w)$ denotes the quantity $T(w) = \max_{p=1,\dots,d} M_p/w_p$. Let also $\bar{M}(w)$ be defined as $\sqrt{k}\|w\|^2 T(w)$. It is immediate that, for every \mathbf{c} in C^k , $I_w(\mathbf{c}) \leq \bar{M}(w)$. Moreover, we define $\bar{\gamma}$ by

$$\bar{\gamma}(\mathbf{c}, x) = \min_{j=1,\dots,k} -2 \langle x, c_j \rangle + \|c_j\|^2,$$

for every \mathbf{c} in C^k and x in \mathbb{R}^d . The prediction results of this paper are based on the following concentration inequality, which connects the deviation of the empirical processes $(P - P_n)\bar{\gamma}(\mathbf{c}, \cdot)$ to the I_w norm.

Proposition 4.1. *Suppose that w is a deterministic sequence of weights. Denote by u the quantity $\log\left(\frac{\|w\|^2 \sqrt{n}}{\sqrt{\log(kd)}}\right)$. If we denote by*

$$\lambda_0 = 16\sqrt{2\pi} \sqrt{\frac{k \log(kd)}{n}} T(w),$$

then, for every $x > 0$, denoting by

$$\lambda_1 = e\lambda_0 \left(1 + \sqrt{\frac{u+x}{k \log kd}}\right),$$

we have, for any fixed \mathbf{c}' in C^k , with probability larger than $1 - e^{-x}$,

$$\sup_{I_w(\mathbf{c}-\mathbf{c}') \leq 2\bar{M}(w)} \frac{|(P - P_n)(\bar{\gamma}(\mathbf{c}, \cdot) - \bar{\gamma}(\mathbf{c}', \cdot))|}{I_w(\mathbf{c} - \mathbf{c}') \vee \lambda_0} \leq \lambda_1. \quad (11)$$

Proposition 4.1 may be thought of as a slight generalization of inequality (7) in [vdG13]. Its proof, given in Section 5.2 of the Appendix, differs from the

original one by the use of Gaussian complexities rather than Talagrand's generic chaining principle to derive a multivariate contraction principle. We are now in position to prove Theorem 3.1.

We recall here that $\lambda \geq \frac{\lambda_1}{1-\kappa_1}$. From Assumption 3, it easily follows that $(1-\kappa_1)I_w(\mathbf{c}) \leq I_{\hat{w}}(\mathbf{c}) \leq (1+\kappa_1)I_w(\mathbf{c})$, with probability larger than $1-r_1(n)$. On this event, we have, for every \mathbf{c} in C^k ,

$$P_n \bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda}, \cdot) + \lambda(1-\kappa_1)I_w(\hat{\mathbf{c}}_{n,\lambda}) \leq P_n \bar{\gamma}(\mathbf{c}, \cdot) + \lambda(1+\kappa_1)I_w(\mathbf{c}).$$

Using (11), with $\mathbf{c}' = \mathbf{0}$, it follows that

$$\begin{aligned} P \bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda}, \cdot) &\leq P \bar{\gamma}(\mathbf{c}, \cdot) + \lambda(1+\kappa_1)I_w(\mathbf{c}) + \lambda_1(I_w(\mathbf{c}) \vee \lambda_0) \\ &\quad + \lambda_1(I_w(\hat{\mathbf{c}}_{n,\lambda}) \vee \lambda_0) - \lambda(1-\kappa_1)I_w(\hat{\mathbf{c}}_{n,\lambda}). \end{aligned} \quad (12)$$

If $I_w(\hat{\mathbf{c}}_{n,\lambda}) > \lambda_0$, adding $-P \bar{\gamma}(\mathbf{c}^*, \cdot)$ on both sides leads to

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \ell(\mathbf{c}, \mathbf{c}^*) + \lambda(1+\kappa_1)I_w(\mathbf{c}) + \lambda_1(I_w(\mathbf{c}) \vee \lambda_0).$$

Hence

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \inf_{r>0} \inf_{I_w(\mathbf{c}) \leq r} \ell(\mathbf{c}, \mathbf{c}^*) + 2\lambda(r \vee \lambda_0).$$

If $I_w(\hat{\mathbf{c}}_{n,\lambda}) \leq \lambda_0$, (12) may be written

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \ell(\mathbf{c}, \mathbf{c}^*) + 2\lambda(I_w(\mathbf{c}) \vee \lambda_0) + \lambda(1-\kappa_1)\lambda_0,$$

hence

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \inf_{r>0} \inf_{I_w(\mathbf{c}) \leq r} \ell(\mathbf{c}, \mathbf{c}^*) + (3-\kappa_1)\lambda(r \vee \lambda_0).$$

4.4 Proof of Proposition 3.1

Let S be a subset of $\{1, \dots, d\}$, and let p be in S such that

$$\sigma_p^2 - R_p^* < \frac{8\kappa_0\lambda^2 w_p^2}{3}.$$

Denote by \mathbf{c}_S^* an optimal codebook with support S , that is

$$\mathbf{c}_S^* = \operatorname{argmin}_{S(\mathbf{c})=S} R(\mathbf{c}).$$

Then, according to (3), we may write

$$\begin{aligned} R(\mathbf{c}_{S \setminus \{p\}}^*) - R(\mathbf{c}_S^*) &\leq R_{S \setminus \{p\}}^* + \sigma_{(S \setminus \{p\})^c}^2 - (R_{S \setminus \{p\}}^* + R_p^*) - \sigma_{S^c}^2 \\ &\leq \sigma_p^2 - R_p^*. \end{aligned}$$

Therefore

$$3R(\mathbf{c}_{S \setminus \{p\}}^*) + 8\lambda^2\kappa_0\|w_{S \setminus \{p\}}\|^2 < 3R(\mathbf{c}_S^*) + 8\lambda^2\kappa_0\|w_S\|^2.$$

4.5 Proof of Theorem 3.2

Let \mathbf{c} be a fixed codebook in C^k , and \mathbf{c}' be another codebook in C^k . Following the notation of [vdG13], with a slight abuse of notation, we denote by $I_{w,1}(\mathbf{c}' - \mathbf{c})$ and $I_{w,2}(\mathbf{c}' - \mathbf{c})$ the quantities

$$\begin{cases} I_{w,1}(\mathbf{c}' - \mathbf{c}) &= I_w((\mathbf{c}' - \mathbf{c})_{S(\mathbf{c})}), \\ I_{w,2}(\mathbf{c}' - \mathbf{c}) &= I_w((\mathbf{c}' - \mathbf{c})_{S^c(\mathbf{c})}). \end{cases}$$

The following result is derived from Lemma A.4 in [vdG08].

Lemma 4.1. *Let \mathbf{c} and \mathbf{c}' be in C^k , and denote by $\mathbf{c}^* = \mathbf{c}^*(\mathbf{c}')$. If $S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)$ or $\mathbf{c}^*(\mathbf{c}) = \mathbf{c}^*$, for any $\delta > 0$, we have*

$$2\lambda I_{w,1}(\mathbf{c}' - \mathbf{c}) \leq \frac{1}{\delta} \ell(\mathbf{c}, \mathbf{c}^*) + \frac{1}{\delta} \ell(\mathbf{c}', \mathbf{c}^*) + 2\delta \kappa_0 \lambda^2 \|w_{S(\mathbf{c})}\|^2. \quad (13)$$

The proof of Lemma 4.1 can be found in [vdG08]. For the sake of completeness it is briefly recalled here.

Proof of Lemma 4.1. Using Cauchy-Schwarz inequality, it is easy to see that

$$\begin{aligned} 2\lambda I_{w,1}(\mathbf{c}' - \mathbf{c}) &\leq 2\lambda \sqrt{\sum_{p \in S(\mathbf{c})} w_p^2} \|\mathbf{c}' - \mathbf{c}\| \\ &\leq 2\lambda \sqrt{\sum_{p \in S(\mathbf{c})} w_p^2} (\|\mathbf{c}' - \mathbf{c}^*\| + \|\mathbf{c} - \mathbf{c}^*\|). \end{aligned}$$

Using the inequality $2ab \leq \kappa_0 \delta a^2 + \frac{1}{\delta \kappa_0} b^2$ and Proposition 2.2 leads to

$$2\lambda I_{w,1}(\mathbf{c}' - \mathbf{c}) \leq \frac{1}{\delta} (\ell(\mathbf{c}, \mathbf{c}^*) + \ell(\mathbf{c}', \mathbf{c}^*)) + 2\delta \kappa_0 \lambda^2 \|w_{S(\mathbf{c})}\|^2.$$

□

Equipped with this Lemma, we are in position to prove Theorem 3.2. By definition, $\hat{\mathbf{c}}_{n,\lambda}$ satisfies

$$P_n \bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda}, \cdot) + \lambda I_{\hat{w}}(\hat{\mathbf{c}}_{n,\lambda}) \leq P_n \bar{\gamma}(\mathbf{c}, \cdot) + \lambda I_{\hat{w}}(\mathbf{c}).$$

Using Proposition 4.1, we get

$$\begin{aligned} P \bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda}, \cdot) + \lambda I_{\hat{w},2}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) &\leq P \bar{\gamma}(\mathbf{c}, \cdot) \\ &\quad + \lambda I_{\hat{w},1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) + \lambda_1 (\lambda_0 \vee I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c})). \end{aligned} \quad (14)$$

If $I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) > \lambda_0$, then adding $-P \bar{\gamma}(\mathbf{c}^*, \cdot)$ on both sides and using Assumption 3 leads to

$$\begin{aligned} \ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + (1 - \kappa_1) \lambda I_{w,2}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) &\leq \ell(\mathbf{c}, \mathbf{c}^*) \\ &\quad + (1 + \kappa_1) \lambda I_{w,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) + \lambda_1 I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}). \end{aligned}$$

Hence, if $\mathbf{c}^*(\mathbf{c}) = \mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}) = \mathbf{c}^*$ or $S(\mathbf{c}) \subsetneq S(\mathbf{c}^*)$,

$$\begin{aligned} \ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + [(1 - \kappa_1)\lambda - \lambda_1] I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) \\ \leq \ell(\mathbf{c}, \mathbf{c}^*) + 2\lambda I_{w,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) \\ \leq \frac{3}{2}\ell(\mathbf{c}, \mathbf{c}^*) + \frac{1}{2}\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + 4\kappa_0\lambda^2 \|w_{S(\mathbf{c})}\|^2, \end{aligned}$$

according to Lemma 4.1, with $\delta = 2$. Noting that $\lambda_1 \leq (1 - \kappa_1)\lambda/2$ yields

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + (1 - \kappa_1)\lambda I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) \leq 3\ell(\mathbf{c}, \mathbf{c}^*) + 8\kappa_0\lambda^2 \|w_{S(\mathbf{c})}\|^2.$$

If $I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) \leq \lambda_0$, then combining Assumption 2.3 with (14) entails

$$\begin{aligned} \ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + (1 - \kappa_1)\lambda I_w(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) &\leq \ell(\mathbf{c}, \mathbf{c}^*) + 2\lambda I_{w,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}) + \lambda_1\lambda_0 \\ &\leq \ell(\mathbf{c}, \mathbf{c}^*) + 3\lambda\lambda_0. \end{aligned}$$

Since $\mathbf{c}_\lambda^*(\hat{\mathbf{c}}_{n,\lambda})$ satisfies $S(\mathbf{c}_\lambda^*(\hat{\mathbf{c}}_{n,\lambda})) \subsetneq S(\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}))$ or $\mathbf{c}^*(\hat{\mathbf{c}}_{n,\lambda}) = \mathbf{c}^*(\mathbf{c}_\lambda^*(\hat{\mathbf{c}}_{n,\lambda}))$, choosing $\mathbf{c} = \mathbf{c}_\lambda^*$ gives the result.

4.6 Proof of Theorem 3.3

Let j be in $S(\mathbf{c}^*)^c$. Theorem 3.3 follows from the Karush-Kuhn-Tucker condition, as in [SWF12], combined with some standard deviation bounds, which are listed below. Throughout this derivation, K will denote a generic positive constant not depending on n .

Proposition 4.2. *For every x in \mathbb{R}^d , let $G^{(j)}(x, \mathbf{c}^*)$ denote the k dimensional vector*

$$\left(x^{(j)} 1_{W_1(\mathbf{c}^*)}(x), \dots, x^{(j)} 1_{W_k(\mathbf{c}^*)}(x) \right).$$

Then, we have

$$\begin{aligned} \mathbb{P} \left[\left| (P - P_n)G^{(j)}(\cdot, \mathbf{c}^*) \right| \geq Kn^{-\alpha} \right] &= O(e^{-n^{1-2\alpha}}), \\ \mathbb{P} \left[\sup_{\mathbf{c} \in C^k} \left| (P_n - P) \sum_{i \neq p} 1_{W_i(\mathbf{c}) \cap W_p(\mathbf{c}^*)} \right| \geq Kn^{-\alpha} \right] &= O(e^{-n^{1-2\alpha}}). \end{aligned}$$

Assume that $\hat{\mathbf{c}}_{n,\lambda}^{(j)} \neq 0$, then the K.K.T condition yields

$$\frac{2}{\sqrt{n}} \|\hat{L}\hat{X}^{(j)}\| = \left\| \sqrt{n}\lambda\hat{w}_j \frac{\hat{\mathbf{c}}_{n,\lambda}^{(j)}}{\|\hat{\mathbf{c}}_{n,\lambda}^{(j)}\|} + \frac{2}{\sqrt{n}} \hat{L}^t \hat{L} \hat{\mathbf{c}}_{n,\lambda}^{(j)} \right\| \geq \sqrt{n}\lambda\hat{w}_j,$$

since $\hat{L}^t \hat{L}$ is positive. According to Assumption 3, it follows that $\sqrt{n}\lambda\hat{w}_j \geq (1 - \kappa_1)n^{1/2}$, with probability larger than $1 - r_1(n)$, when n is large enough. On the other hand, we have

$$\begin{aligned} \|L(\hat{\mathbf{c}}_{n,\lambda})X^{(j)}\| &\leq \|L(\mathbf{c}^*)X^{(j)}\| + \|(L(\hat{\mathbf{c}}_{n,\lambda}) - L(\mathbf{c}^*))X^{(j)}\| \\ &\leq n\|P_n G^{(j)}(\cdot, \mathbf{c}^*)\| + M^{(j)}n P_n \sum_{i \neq p} 1_{W_i(\hat{\mathbf{c}}_{n,\lambda}) \cap W_p(\mathbf{c}^*)}. \end{aligned}$$

According to the centroid condition, $PG^{(j)}(\cdot, \mathbf{c}^*) = 0$. Thus, it follows from Proposition 4.2 that $\|P_n G^{(j)}(\cdot, \mathbf{c}^*)\| \leq Kn^{-\alpha}$ with probability $1 - O(e^{-n^{1-2\alpha}})$. Lemma 4.2 in [Lev15] ensures that

$$P \sum_{i \neq p} 1_{W_i(\hat{\mathbf{c}}_{n,\lambda}) \cap W_p(\mathbf{c}^*)} \leq p(K\|\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*\|) \leq K\|\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*\|,$$

according to Assumption 2. Moreover, since $\|w_{S(\mathbf{c}^*)}\|$ tends to $\sum_{i \in S(\mathbf{c}^*)} \frac{1}{\|\mathbf{c}^{*,(i)}\|}$, Theorem 3.2 yields $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq K\lambda^2$, with probability larger than $1 - r_1(n) - O(e^{-n^{1-2\alpha}})$, for n large enough. On the same event, Proposition 2.2 gives $\|\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*\| \leq K\lambda$, which leads to

$$P_n \sum_{i \neq p} 1_{W_i(\hat{\mathbf{c}}_{n,\lambda}) \cap W_p(\mathbf{c}^*)} \leq Kn^{-\alpha},$$

with probability larger than $1 - r_1(n) - O(e^{-n^{1-2\alpha}})$, according to Proposition 4.2. Then the K.K.T condition entails

$$(1 - \kappa_1)n^{1/2} \leq Kn^{1/2-\alpha},$$

with probability larger than $1 - r_1(n) - O(e^{-n^{1-2\alpha}})$, which is impossible when n is large enough. Noting that $r_1(n) = O(e^{-n^{1-2\alpha}})$, according to Proposition 3, gives the result.

Conversely, if j is in $S(\mathbf{c}^*)$ and $\hat{\mathbf{c}}_{n,\lambda}^{(j)} = 0$, then the K.K.T condition yields

$$\frac{2}{n} \|L(\hat{\mathbf{c}}_{n,\lambda}) \hat{X}^{(j)}\| \leq \lambda \hat{w}_j.$$

Since w_j tends to $1/\|\mathbf{c}^{*,(j)}\|$ and $\|PG^{(j)}(\cdot, \mathbf{c}^*)\| > 0$, according to the centroid condition, using the same concentration bounds as above leads to a contradiction.

4.7 Proof of Proposition 3.2

The first part of Proposition 3.2 is derived from the following Lemma, which extends Proposition 3.2 of [Lev15]. We denote by \tilde{B} the quantity $\inf_{i \neq j} \|m_i - m_j\|$.

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

$$R(\mathbf{m}) \leq \frac{\sigma^2 k \theta_{max} d}{(1 - \eta)}, \quad (15)$$

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

$$R(\mathbf{c}) > \frac{\tau^2 \tilde{B}^2 \theta_{min}}{4} \left(1 - \frac{2\sigma\sqrt{d}}{\sqrt{2\pi}\tau\tilde{B}} e^{-\frac{\tau^2 \tilde{B}^2}{4d\sigma^2}} \right)^d, \quad (16)$$

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

$$\forall t \leq \tau' \tilde{B} \quad p(t) \leq t \frac{2k^2 \theta_{\max} M^{d-1} S_{d-1}}{(2\pi)^{d/2} (1-\eta) c_-^d \sigma^d} e^{-\frac{[\frac{1}{2} - (2\tau + \tau')]^2 \tilde{B}^2}{2\sigma^2}}, \quad (17)$$

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

The proof of Lemma 4.2 follows from direct calculation, as in the proof of Proposition 3.2 of [Lev15]. For the sake of completeness it is given in Section 5.5 of the Appendix.

Let τ and τ' be positive quantities satisfying $2\tau + \tau' < 1/2$, and $\tau' > 8\sqrt{2}M\tau/(1-2\tau)\tilde{B}$. According to (15) and (16), if σ is small enough, then every optimal codebook \mathbf{c}^* satisfies $\sup_{j=1,\dots,k} \|m_j - c_j^*\| \leq \tau\tilde{B}$, up to relabeling code points.

On the other hand, (17) ensures that, for σ small enough, P satisfies Assumption 2 with radius $r_0 \geq \tau'\tilde{B}$. Let \mathbf{c}^* be an optimal codebook. According to *i*) of Proposition 2.2 in [Lev15], no other optimal codebook can be found in a ball of radius $(1-2\tau)\tilde{B}\tau'/4\sqrt{2}M$ centered at \mathbf{c}^* . Since $(1-2\tau)\tilde{B}\tau'/4\sqrt{2}M > 2\tau$, this proves that \mathbf{c}^* must be unique.

To apply Theorem 3.3, we need to show that $S(\mathbf{c}^*) \subset \tilde{S}$. Suppose that there exists $j \geq d'+1$ such that $\mathbf{c}^{*,(j)} \neq 0$. Let s denote the orthogonal transformation defined by $s(x_1, \dots, x_{d'}, x_{d'+1}, \dots, x_d) = (x_1, \dots, x_{d'}, -x_{d'+1}, \dots, -x_d)$. Since $(\Sigma_i)_{p,q} = 0$, for every (i, p, k) in $\{1, \dots, k\} \times \{1, \dots, d'\} \times \{d'+1, \dots, d\}$, P is invariant through composition by s . Hence $s(\mathbf{c}^*)$ is an optimal codebook, and $\mathbf{c}^* \neq s(\mathbf{c}^*)$, which contradicts the fact that \mathbf{c}^* is unique.

5 Appendix

5.1 Proof of Proposition 2.3

Hoeffding's inequality ensures that, for every $p = 1, \dots, d$, $\hat{\sigma}_p^2/\sigma_p^2 - 1$ is a subgaussian random variable with variance bounded by $T(\sigma)^4/4n$. For a comprehensive introduction to subgaussian random variables and its application to empirical process theory, the interested reader is referred to [Mas07]. Applying Theorem 3.12 in [Mas07] and a bounded difference concentration inequality (see, e.g., Theorem 5.1 in [Mas07]) yields, with probability larger than $1 - e^{-y}$,

$$\max_{p=1,\dots,d} \left| \frac{\hat{\sigma}_p^2}{\sigma_p^2} - 1 \right| \leq \frac{T(\sigma)^2 \sqrt{\log(d)}}{\sqrt{2n}} \left(1 + \sqrt{\frac{y}{\log(d)}} \right).$$

Noting that $2n > T(\sigma)^2 \sqrt{\log(d)}$ and $y < \log(d) \left(\frac{2n}{T(\sigma)^2 \sqrt{\log(d)}} - 1 \right)^2$ leads to the result.

For the threshold Lasso with unique optimal codebook, Theorem 3.1 in [Lev15], combined with Assumption 2, provides a constant C_0 such that

$$\|\hat{\mathbf{c}}_n - \mathbf{c}^*\| \leq C_0 M \sqrt{\frac{k}{n}} \left(1 + \frac{x}{k}\right),$$

with probability larger than $1 - e^{-x}$. Since, for every p in $\{1, \dots, d\}$,

$$\left| \frac{\|\mathbf{c}^{*,(p)}\| \vee \delta}{\|\hat{\mathbf{c}}_n^{(p)}\| \vee \delta} - 1 \right| \leq \frac{\left| \|\mathbf{c}^{*,(p)}\| \vee \delta - \|\hat{\mathbf{c}}_n^{(p)}\| \vee \delta \right|}{\delta} \leq \frac{\|\hat{\mathbf{c}}_n - \mathbf{c}^*\|}{\delta},$$

the results easily follows.

5.2 Proof of Proposition 4.1

For a fixed \mathbf{c} in \mathbf{c}^k , denote by $Z_r(\mathbf{c})$ the following random variable

$$Z_r(\mathbf{c}) = \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} |(P - P_n)(\bar{\gamma}(\mathbf{c}', \cdot) - \bar{\gamma}(\mathbf{c}, \cdot))|.$$

The following proposition gives a non-asymptotic bound on $Z_r(\mathbf{c})$.

Proposition 5.1. *Suppose that w is deterministic. Let $x > 0$, and \mathbf{c} be a fixed codebook. Then, with probability larger than $1 - e^{-x}$,*

$$Z_r(\mathbf{c}) \leq 16\sqrt{2\pi} \sqrt{\frac{k \log(kd)}{n}} r T(w) \left(1 + \frac{1}{4\sqrt{\pi}} \sqrt{\frac{x}{k \log(kd)}}\right).$$

The proof of Proposition 5.1 is postponed to the next subsection. Proposition 4.1 derives from a peeling argument, as in Section 3.4 of [vdG13], combined with Proposition 5.1. Let a be the smallest integer such that $e^{-(a-1)} 2\bar{M}(w) \leq \lambda_0$, and take $u_0 = \log(a)$ (we recall here that $\bar{M}(w) = \sqrt{k} \|w\|^2 T(w)$ is an upper bound on $I_w(\mathbf{c})$, for \mathbf{c} in \mathbf{C}^k). Then it is easy to see that $u_0 \leq u$, where u is defined in Proposition 4.1. We may write

$$\begin{aligned} & \mathbb{P} \left(\sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq 2\bar{M}(w)} \frac{|(P - P_n)(\bar{\gamma}(\mathbf{c}', \cdot) - \bar{\gamma}(\mathbf{c}, \cdot))|}{I_w(\mathbf{c}' - \mathbf{c}) \vee \lambda_0} \geq \lambda_1 \right) \\ & \leq \sum_{j=2}^a \mathbb{P} \left(\sup_{\substack{I_w(\mathbf{c}' - \mathbf{c}) \leq 2e^{-(j-1)} \bar{M}(w) \\ I_w(\mathbf{c}' - \mathbf{c}) \geq 2e^{-j} \bar{M}(w)}} \frac{|(P - P_n)(\bar{\gamma}(\mathbf{c}', \cdot) - \bar{\gamma}(\mathbf{c}, \cdot))|}{2e^{-j} \bar{M}(w)} \geq \lambda_1 \right) \\ & \quad + \mathbb{P} \left(\sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq \lambda_0} \frac{|(P - P_n)(\bar{\gamma}(\mathbf{c}', \cdot) - \bar{\gamma}(\mathbf{c}, \cdot))|}{2e^{-(a-1)} \bar{M}(w)} \geq \lambda_1 \right) \\ & \leq \sum_{j=1}^a \mathbb{P} \left(Z_{2e^{-(j-1)} \bar{M}(w)} \geq 2e^{-(j-1)} \bar{M}(w) \lambda_0 \left(1 + \sqrt{\frac{u+x}{k \log kd}}\right) \right) \\ & \leq ae^{-u} e^{-x}, \end{aligned}$$

where the last inequality follows from Proposition 5.1. Noticing that $ae^{-u} \leq 1$ proves the result.

5.3 Proof of Proposition 5.1

This proof is a slight modification of the proof of Theorem 3.1 in [Lev15], and mainly relies on the use of Gaussian complexities combined with Slepian's Lemma (see, e.g., Theorem 3.14 in [Mas07]). First, it may be easily noticed that, for every $j = 1, \dots, k$, if $I_w(\mathbf{c}' - \mathbf{c}) \leq r$, then, for all x in \mathbb{R}^d ,

$$|-2\langle x, c'_j \rangle + \|c'_j\|^2 + 2\langle x, c_j \rangle - \|c_j\|^2| \leq 4rT(w),$$

which leads to

$$\|\bar{\gamma}(\mathbf{c}', \cdot) - \bar{\gamma}(\mathbf{c}, \cdot)\|_\infty \leq 4rT(w).$$

As a consequence, a bounded difference concentration inequality (see, e.g., Theorem 5.1 in [Mas07]) yields, with probability larger than $1 - e^{-x}$,

$$Z_r(\mathbf{c}) \leq \mathbb{E}Z_r(\mathbf{c}) + 4rT(w)\sqrt{\frac{2x}{n}}.$$

It remains to bound from above $\mathbb{E}Z_r(\mathbf{c})$. According to the symmetrization principle (see, e.g., Section 2.2 of [Kol06]), combined with Lemma 4.5 of [LT91], we may write

$$\mathbb{E}Z_r(\mathbf{c}) \leq 2\sqrt{\frac{\pi}{2}}\mathbb{E}_X\mathbb{E}_g \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} \frac{1}{n} \sum_{i=1}^n g_i(\bar{\gamma}(\mathbf{c}', X_i) - \bar{\gamma}(\mathbf{c}, X_i)),$$

where the g_i 's are independent standard Gaussian variables. Let X_1, \dots, X_n and \mathbf{c} be fixed, and define, for \mathbf{c}' such that $I_w(\mathbf{c}' - \mathbf{c}) \leq r$ the Gaussian process

$$Y_{\mathbf{c}'} = \sum_{i=1}^n g_i(\bar{\gamma}(\mathbf{c}', X_i) - \bar{\gamma}(\mathbf{c}, X_i)).$$

Since, for every codebooks \mathbf{c}'_1 and \mathbf{c}'_2 ,

$$(\bar{\gamma}(\mathbf{c}'_1, X_i) - \bar{\gamma}(\mathbf{c}'_2, X_i))^2 \leq \max_{j=1, \dots, k} 8\langle c'_{1,j} - c'_{2,j}, X_i \rangle^2 + 2(\|c'_{1,j}\|^2 - \|c'_{2,j}\|^2)^2,$$

it is easy to see that

$$\text{Var}(Y_{\mathbf{c}'_1} - Y_{\mathbf{c}'_2}) \leq \sum_{i=1}^n \sum_{j=1}^k 8\langle c'_{1,j} - c'_{2,j}, X_i \rangle^2 + 2n \sum_{j=1}^k (\|c'_{1,j}\|^2 - \|c'_{2,j}\|^2)^2.$$

To derive bounds on the Gaussian complexity defined above, the following comparison result between Gaussian processes is needed.

Theorem 5.1 (Slepian's Lemma). *Let Y_t and N_t , t in \mathcal{V} , be some centered real Gaussian processes. Assume that*

$$\forall t_1, t_2 \in \mathcal{V} \quad \text{Var}(Y_{t_1} - Y_{t_2}) \leq \text{Var}(N_{t_1} - N_{t_2}),$$

then

$$\mathbb{E} \sup_{t \in \mathcal{V}} Y_t \leq 2\mathbb{E} \sup_{t \in \mathcal{V}} N_t.$$

A proof of Theorem 5.1 can be found in Theorem 3.14 of [Mas07]. Denote by \mathcal{V} the set of codebooks \mathbf{c}' in C^k such that $I_w(\mathbf{c}' - \mathbf{c}) \leq r$. Now introduce, for \mathbf{c}' such that $I_w(\mathbf{c}' - \mathbf{c}) \leq r$, the following Gaussian process

$$N_{\mathbf{c}'} = 2\sqrt{2} \sum_{i=1}^n \sum_{j=1}^k \langle c'_j - c_j, X_i \rangle \xi_{i,j} + \sqrt{2n} \sum_{j=1}^k (\|c'_j\|^2 - \|c_j\|^2) \xi'_j,$$

where the ξ 's and ξ' 's are independent standard Gaussian random variables. Note that, for all \mathbf{c}'_1 and \mathbf{c}'_2 in \mathcal{V} , $\text{Var}(Y_{\mathbf{c}'_1} - Y_{\mathbf{c}'_2}) \leq \text{Var}(N_{\mathbf{c}'_1} - N_{\mathbf{c}'_2})$. Consequently, applying Theorem 5.1 yields

$$\mathbb{E}_g \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} Y_{\mathbf{c}'} \leq 2\mathbb{E}_{\xi, \xi'} \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} N_{\mathbf{c}'}$$

It follows that

$$\begin{aligned} \mathbb{E}_{\xi, \xi'} \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} N_{\mathbf{c}'} &\leq \mathbb{E}_{\xi} \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} 2\sqrt{2} \sum_{i=1}^n \sum_{j=1}^k \langle c'_j - c_j, X_i \rangle \xi_{i,j} \\ &\quad + \mathbb{E}_{\xi'} \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} \sqrt{2n} \sum_{j=1}^k (\|c'_j\|^2 - \|c_j\|^2) \xi'_j. \end{aligned}$$

The first term of the right side can be bounded as follows.

$$\begin{aligned} &\mathbb{E}_{\xi} \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} 2\sqrt{2} \sum_{i=1}^n \sum_{j=1}^k \langle c'_j - c_j, X_i \rangle \xi_{i,j} \\ &\leq 2\sqrt{2} \mathbb{E}_{\xi} \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} \sum_{j=1}^k \left\langle c'_j - c_j, \sum_{i=1}^n \xi_{i,j} X_i \right\rangle \\ &\leq 2\sqrt{2} \mathbb{E}_{\xi} \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} \left(\sum_{j=1}^k \sum_{p=1}^d w_p |c'^{(p)}_j - c_j^{(p)}| \right) \max_{j,p} \left| \sum_{i=1}^n \frac{\xi_{i,j} X_i^{(p)}}{w_p} \right| \\ &\leq 2\sqrt{2kr} \mathbb{E}_{\xi} \max_{j=1, \dots, k, p=1, \dots, d} \left| \sum_{i=1}^n \frac{\xi_{i,j} X_i^{(p)}}{w_p} \right|. \end{aligned}$$

Note that, for every (j, p) , the random variable $\sum_{i=1}^n \frac{\xi_{i,j} X_i^{(p)}}{w_p}$ is Gaussian, with variance bounded by $nT^2(w)$. Consequently, applying Theorem 3.12 in [Mas07] gives

$$\mathbb{E}_{\xi} \max_{j=1, \dots, k, p=1, \dots, d} \left| \sum_{i=1}^n \frac{\xi_{i,j} X_i^{(p)}}{w_p} \right| \leq T(w) \sqrt{2n \log(kd)}.$$

In turn, the second term of the right side may be bounded by

$$\begin{aligned}
\mathbb{E}_{\xi'} \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} \sqrt{2n} \sum_{j=1}^k (\|c'_j\|^2 - \|c_j\|^2) \xi'_j \\
\leq \sqrt{2n} \mathbb{E}_{\xi'} \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} \sum_{j=1}^k \left(\sum_{p=1}^d w_p |c'_j{}^{(p)} - c_j{}^{(p)}| \frac{2M_p}{w_p} \right) |\xi'_j| \\
\leq 2\sqrt{2n} T(w) \mathbb{E}_{\xi'} \sup_{I_w(\mathbf{c}' - \mathbf{c}) \leq r} I(\mathbf{c}' - \mathbf{c}) \sqrt{\sum_{j=1}^k \xi_j'^2} \\
\leq 2T(w) r \sqrt{2nk}.
\end{aligned}$$

Combining these two bounds leads to

$$\mathbb{E} Z_r(\mathbf{c}) \leq 16\sqrt{2\pi} \sqrt{\frac{k \log(kd)}{n}} r T(w).$$

5.4 Proof of Proposition 4.2

For every u in \mathbb{R}^k , we denote by Y_u the random variable $\langle u, G^{(j)}(\cdot, \mathbf{c}^*) \rangle$, so that $\|(P - P_n)G^{(j)}(\cdot, \mathbf{c}^*)\| = \sup_{\|u\| \leq 1} (P - P_n)Y_u := Y$. Since, for every u such that $\|u\| \leq 1$, $|Y_u| \leq M_j$ a.s., a bounded difference inequality yields (see, e.g., Theorem 5.1 in [Mas07]), with probability larger than $1 - e^{-x}$,

$$Y \leq \mathbb{E}Y + M_p \sqrt{\frac{2x}{n}}.$$

An upper bound on $\mathbb{E}Y$ may be derived the same way as in the proof of Proposition 5.1: introducing some Rademacher random variables ε_i , $i = 1, \dots, n$, and using the symmetrization principle, we get

$$\begin{aligned}
\mathbb{E}Y &\leq 2\mathbb{E}_{X, \varepsilon} \sup_{\|u\| \leq 1} \left\langle u, \frac{1}{n} \sum_{i=1}^n \varepsilon_i G^{(j)}(X_i, \mathbf{c}^*) \right\rangle \\
&\leq \frac{2}{n} \sqrt{\mathbb{E}_{X, \varepsilon} \left\| \sum_{i=1}^n \varepsilon_i G^{(j)}(X_i) \right\|^2} \\
&\leq \frac{2M_p}{\sqrt{n}},
\end{aligned}$$

according to Cauchy-Schwarz and Jensen's inequalities. Choosing $x = n^{1-2\alpha}$ gives the first concentration inequality of Proposition 4.2.

Now consider the $\{0, 1\}$ -valued random variables $Y_{\mathbf{c}}$, indexed by C^k , defined by $Y_{\mathbf{c}} = 1_{\bigcup_{i \neq p} W_i(\mathbf{c}) \cap W_p(\mathbf{c}^*)}$. According to [Pol82], the sets $\{\bigcup_{i \neq p} W_i(\mathbf{c}) \cap W_p(\mathbf{c}^*) | \mathbf{c} \in C^k\}$ have finite VC-dimension, say D . Using the well-known Vapnik-Chervonenkis bound (see, e.g., [BBL05]), combined with a bounded difference

concentration inequality yields

$$P_n Y_{\mathbf{c}} \leq P Y_{\mathbf{c}} + K \sqrt{\frac{D}{n}} \left(1 + \sqrt{\frac{x}{D}} \right),$$

with probability larger than $1 - e^{-x}$, for every \mathbf{c} in C^k , and for some absolute constant K . Choosing $x = n^{1-2\alpha}$ provides the second inequality of Proposition 4.2.

5.5 Proof of Lemma 4.2

The proof of Proposition 4.2 follows from the Proof of Proposition 3.2 in [Lev15]. To give an upper bound on $R(\mathbf{m})$, we may write

$$\begin{aligned} & \sum_{i,j=1}^k \frac{\theta_i}{(2\pi)^{d/2} N_i \sqrt{|\Sigma_i|}} \int_{W_j(\mathbf{m})} \|x - m_j\|^2 e^{-\frac{1}{2}(x-m_i)^\top \Sigma_i^{-1}(x-m_i)} \mathbf{1}_{\mathcal{B}(0,M)}(x) dx \\ & \leq \sum_{i=1}^k \frac{\theta_i}{(1-\eta)(2\pi)^{d/2} \sqrt{|\Sigma_i|}} \int_{\mathbb{R}^d} \|x - m_i\|^2 e^{-\frac{1}{2}(x-m_i)^\top \Sigma_i^{-1}(x-m_i)} dx \\ & \leq \sum_{i=1}^k \frac{\theta_i}{(1-\eta)(2\pi)^{d/2}} \int_{\mathbb{R}^d} \|\sqrt{\Sigma_i} u\|^2 e^{-\frac{1}{2}\|u\|^2} du, \end{aligned}$$

where $\sqrt{\Sigma_i}$ denotes the square root of the matrix Σ_i . Since Σ_i has its largest eigenvalue bounded by σ^2 , it follows that $\|\sqrt{\Sigma_i} u\|^2 \leq \sigma^2 \|u\|^2$, for every u in \mathbb{R}^d . We deduce that

$$\begin{aligned} R(\mathbf{m}) & \leq \frac{\sigma^2 k \theta_{max}}{(2\pi)^{d/2} (1-\eta)} \int_{\mathbb{R}^d} \|u\|^2 e^{-\frac{1}{2}\|u\|^2} du \\ & \leq \frac{\sigma^2 k \theta_{max} d}{(1-\eta)}, \end{aligned}$$

which proves (15). Now let \mathbf{c} be a codebook, and let i be such that $\|c_i - m_j\| > \tau \tilde{B}$, for every j in $\{1, \dots, k\}$, with $\tau < 1/2$. Since $\mathcal{B}(m_i, \tau \tilde{B}/2) \subset \mathcal{B}(0, M)$, we may write

$$\begin{aligned} R(\mathbf{c}) & > \int_{\mathcal{B}(m_i, \tau \tilde{B}/2)} \left(\frac{\tau \tilde{B}}{2} \right)^2 \frac{\theta_i}{(2\pi)^{d/2} \sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(x-m_i)^\top \Sigma_i^{-1}(x-m_i)} dx \\ & > \frac{\theta_{min} \tau^2 \tilde{B}^2}{4(2\pi)^{d/2}} \int_{\sqrt{\Sigma_i}^{-1} \mathcal{B}(0, \tau \tilde{B}/2)} e^{-\frac{1}{2}\|u\|^2} du \\ & > \frac{\theta_{min} \tau^2 \tilde{B}^2}{4(2\pi)^{d/2}} \int_{\mathcal{B}(0, \frac{\tau \tilde{B}}{2\sigma})} e^{-\frac{1}{2}\|u\|^2} du. \end{aligned}$$

Since, for every positive r , $[-r/\sqrt{d}, r/\sqrt{d}]^d \subset \mathcal{B}(0, r)$, and $\int_r^\infty e^{-r^2/2} dr \leq e^{-r^2/2}/r$, it follows that

$$\begin{aligned} R(\mathbf{c}) &> \frac{\theta_{\min} \tau^2 \tilde{B}^2}{4(2\pi)^{d/2}} \left(2 \int_0^{\frac{\tau \tilde{B}}{2\sigma\sqrt{d}}} e^{-\frac{r^2}{2}} dr \right)^d \\ &> \frac{\tau^2 \tilde{B}^2 \theta_{\min}}{4} \left(1 - \frac{2\sigma\sqrt{d}}{\sqrt{2\pi}\tau\tilde{B}} e^{-\frac{\tau^2 \tilde{B}^2}{4d\sigma^2}} \right)^d, \end{aligned}$$

which proves (16). At last, let τ' be such that $2\tau + \tau' < \frac{1}{2}$, and let y be in $N_{\mathbf{c}^*}(\tau'\tilde{B})$. Then, for every i in $\{1, \dots, k\}$, we have $\|m_i - y\| \geq [\frac{1}{2} - (2\tau + \tau')]\tilde{B}$. Hence

$$f(y) \leq \frac{k\theta_{\max}}{(2\pi)^{d/2}(1-\eta)\sigma_-^d} e^{-\frac{[\frac{1}{2} - (2\tau + \tau')]^2 \tilde{B}^2}{2\sigma_-^2}}.$$

Since $\sigma_- \geq c_- \sigma$ and $\lambda(N_{\mathbf{c}^*}(t)) \leq 2tkM^{d-1}S_{d-1}$, straightforward calculation leads to (17).

References

- [ABCP13] Anestis Antoniadis, Xavier Brossat, Jairo Cugliari, and Jean-Michel Poggi. Clustering functional data using wavelets. *Int. J. Wavelets Multiresolut. Inf. Process.*, 11(1):1350003, 30, 2013.
- [Bac08] Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.
- [BBL05] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.
- [BDL08] Gérard Biau, Luc Devroye, and Gábor Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory*, 54(2):781–790, 2008.
- [CWLX14] Xiangyu Chang, Yu Wang, Rongjian Li, and Zongben Xu. Sparse K-Means with ℓ_∞/ℓ_0 Penalty for High-Dimensional Data Clustering. *ArXiv e-prints*, 2014.
- [DSC94] Geert De Soete and J. Douglas Carroll. K-means clustering in a low-dimensional euclidean space. In Edwin Diday, Yves Lechevalier, Martin Schader, Patrice Bertrand, and Bernard Burtschy, editors, *New Approaches in Classification and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 212–219. Springer Berlin Heidelberg, 1994.

- [Fis10] Aurélie Fischer. Quantization and clustering with Bregman divergences. *J. Multivariate Anal.*, 101(9):2207–2221, 2010.
- [GG91] Allen Gersho and Robert M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991.
- [GL00] Siegfried Graf and Harald Luschgy. *Foundations of quantization for probability distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.
- [GLP07] Siegfried Graf, Harald Luschgy, and Gilles Pagès. Optimal quantizers for Radon random vectors in a Banach space. *J. Approx. Theory*, 144(1):27–53, 2007.
- [JW14] Jiashun Jin and Wanjie Wang. Important Feature PCA for high dimensional clustering. *ArXiv e-prints*, July 2014.
- [Kol06] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.
- [Lev15] Clément Levrard. Nonasymptotic bounds for vector quantization in Hilbert spaces. *Ann. Statist.*, 43(2):592–619, 2015.
- [Llo82] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, 28(2):129–137, 1982.
- [LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.
- [Mas07] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
- [Mey13] Caroline Meynet. An ℓ_1 -oracle inequality for the Lasso in finite mixture Gaussian regression models. *ESAIM Probab. Stat.*, 17:650–671, 2013.
- [MM11] Pascal Massart and Caroline Meynet. The Lasso as an ℓ_1 -ball model selection procedure. *Electron. J. Stat.*, 5:669–687, 2011.
- [MM13] Cathy Maugis-Rabusseau and Bertrand Michel. Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM Probab. Stat.*, 17:698–724, 2013.
- [Pol81] David Pollard. Strong consistency of k -means clustering. *Ann. Statist.*, 9(1):135–140, 1981.

- [Pol82] David Pollard. A central limit theorem for k -means clustering. *Ann. Probab.*, 10(4):919–926, 1982.
- [SB08] Douglas Steinley and Michael J. Brusco. Selection of variables in cluster analysis: an empirical comparison of eight procedures. *Psychometrika*, 73(1):125–144, 2008.
- [SWF12] Wei Sun, Junhui Wang, and Yixin Fang. Regularized k -means clustering of high-dimensional data and its asymptotic consistency. *Electron. J. Stat.*, 6:148–167, 2012.
- [TCKV10] Marieke E. Timmerman, Eva Ceulemans, Henk A. L. Kiers, and Maurizio Vichi. Factorial and reduced K -means reconsidered. *Comput. Statist. Data Anal.*, 54(7):1858–1871, 2010.
- [Ter14] Yoshikazu Terada. Strong consistency of reduced K -means clustering. *Scand. J. Stat.*, 41(4):913–931, 2014.
- [Ter15] Yoshikazu Terada. Strong consistency of factorial K -means clustering. *Ann. Inst. Statist. Math.*, 67(2):335–357, 2015.
- [vdG08] Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.
- [vdG13] Sara A. van de Geer. Generic chaining and the ℓ_1 -penalty. *J. Statist. Plann. Inference*, 143(6):1001–1012, 2013.
- [VK01] Maurizio Vichi and Henk A. L. Kiers. Factorial k -means analysis for two-way data. *Comput. Statist. Data Anal.*, 37(1):49–64, 2001.
- [WT10] Daniela M. Witten and Robert Tibshirani. A framework for feature selection in clustering. *J. Amer. Statist. Assoc.*, 105(490):713–726, 2010.