Submitted to Bernoulli

Sparse oracle inequalities for variable selection via regularized quantization

CLÉMENT LEVRARD

Université Paris Diderot, 8 place Aurélie Nemours, 75013 Paris
E-mail: levrard@math.univ-paris-diderot.fr

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, suggesting that these procedures might be of particular interest in high dimensional settings. 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, providing an asymptotic consistency rate. 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.

Keywords: k-means, variable selection, sparsity, Lasso, oracle inequalities, clustering, high dimension.

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 loosing 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., [10]). 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

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

where Pf 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., [9] or [11]). 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 [15]. Moreover, from a practical point of view, the k-means algorithm (see [16]) 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., [11] or [21]). 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 [20] 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 [7] or [31]. Though these procedures offer good performance in classifying the 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 [23]. This algorithm has shown good results on practical examples, such as curve clustering (see, e.g., [1]). 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 [8]) and FKM (Factorial k-means, introduced in [30]), are also very popular in practice. Some results on the performance in classifying the

sample $X_1, \ldots, X_n$ have been derived in [27] under strong conditions on P. In addition, some asymptotic prediction results on these procedures have been established in [25] and [26], 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 [20]. However, these methods could be unsuitable 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., [18] for a L_0-type penalization method, and [19] 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., [15]).

This paper exposes a theoretical study of a weighted Lasso type procedure adapted to k-means, as suggested in [24]. Results are given for a general family of weights, encompassing the weights proposed in [24] as well as those proposed in [28] 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 [24] (for instance, the coordinates are not assumed to be independant), and adapting the sparsity of the optimal codebooks. From these non-asymptotic bounds, some asymptotic rates of convergence are derived when both the dimension and the sample size are large, showing that these Lasso type procedures may be suitable for high dimensional quantization. Interestingly, the excess distortion bounds are valid in the case where it may exist several optimal codebooks, contrary to results in [24] and [28]. 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 $p = 1, \ldots, d$, 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

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

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 $R(\mathbf{c}) = R(Q) = P\gamma(\mathbf{c}, .)$, 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 $\hat{R}_n(\mathbf{c}) = P_n\gamma(\mathbf{c}, .)$, 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 [9]). In the sequel, the set of minimizers of the risk $R(.)$ 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 $\mathcal{M}$.

From now on we assume that $k \geq 2$. 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

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

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

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

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

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

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

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

where 1_A denotes the indicator function associated with A. In the case where $P(W_i(\mathbf{c})) \neq 0$, for every $i = 1, \ldots, k$, 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 [12] that, for every Voronoi partition $W(\mathbf{c}^*)$ associated with an optimal codebook $\mathbf{c}^*$, and every $i = 1, \ldots, k$, $P(W_i(\mathbf{c}^*)) \neq 0$. Consequently, any optimal codebook satisfies the so-called centroid condition (see, e.g., Section 6.2 of [10]), 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 $\mathcal{M}$, $i \neq j$,

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

A proof of this statement can be found in Proposition 1 of [12]. According to [15], 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 [15], 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.

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

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

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

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 [24] 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., [2]), according to different criterions. Note that this kind of penalties is well-adapted to centered distributions. In practice, centering the data provides codebooks of the form $(\hat{c}_{n,\lambda,1} + \bar{X}, \dots, \hat{c}_{n,\lambda,k} + \bar{X})$ for the non centered distribution, where $\bar{X}$ denotes the empirical mean and $\hat{\mathbf{c}}_{n,\lambda}$ is hopefully sparse. From a theoretical point of view, deriving how close the codebook $\hat{\mathbf{c}}_{n,\lambda}$ computed on the centered data is to a codebook $\mathbf{c}^* - m$ would require a bound on $\|\bar{X} - m\|$, where m is the mean of P. In our framework, such a bound is typically of order $\sqrt{d/n}$ (see, e.g., Figure 1), hence might be unsuited for high dimensional settings. However, in some particular cases (for instance when the mean is sparse), other estimators of the means that are adapted to the high dimensional framework could be combined with our procedure.

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:

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

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

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

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 $S(\mathbf{c}) = \{j \in \{1,\dots,d\} \,|\mathbf{c}^{(j)} \neq 0\}$. 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 [13]. 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 [15].

As exposed in [15], Assumption 2 may be thought of as a margin condition for squared distance based quantization. Some examples of distributions satisfying (4) are given in [15], 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 [24] seems stronger than the condition required in Assumption 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 $\mathbf{c}^*(\mathbf{c}) \in \operatorname{argmin}_{\mathbf{c}^*} \|\mathbf{c} - \mathbf{c}^*\|$.

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 [15], 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 [21] 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 exist deterministic weights $w_p > 0$, $p = 1, \ldots, d$, and a constant $0 \leq \kappa_1 < 1$ such that*

$$\mathbb{P}\left(\sup_{p=1,\ldots,d} \left| \frac{\hat{w}_p}{w_p} - 1 \right| > \kappa_1 \right) := r_1(n) \underset{n\to\infty}{\longrightarrow} 0. \tag{5}$$

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 $\sup_{p=1,\ldots,d} M_p/w_p$. 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 $M = \sqrt{M_1^2 + \ldots + M_d^2}$. 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 $r_1(n) = e^{-\left(\frac{n\delta^2\kappa_1^2}{C_0^2 M^2} - k\right)}$.

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 3.4, 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 [28].

$$\begin{cases} \lambda_0 &= 8\sqrt{2\pi}\sqrt{\frac{k\log(kd)}{n}}T(w), \\ \lambda_1(x) &= e\lambda_0\left(1+\sqrt{\frac{u+x}{k\log(kd)}}\right), \end{cases} \tag{6}$$

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

We recall here that $k\geq 2$. The case $k=1$ may be treated as a special case of the standard Lasso estimator for linear regression (see, e.g., Chapter 2 of [6]).

3.1. Sparsity adaptive slow rate of convergence for the distortion

Following the approach of [17], 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 $1-r_1(n)-e^{-x}$, for every $\mathbf{c}^*$ in $\mathcal{M}$, we have

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

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 $\ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*)$ of order $T(w)/\sqrt{n}$, 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 [3].

Besides, some asymptotic results for $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*)$ when both d and n are large may also be deduced from Theorem 3.1, as stated by the following corollary.

Corollary 3.1. *Let* $\mathbf{c}^*$ *be in* $\mathcal{M}$ *and denote by* d^* *the quantity* $|S(\mathbf{c}^*)|$. *Assume that* $\max_{p=1,\dots,d} M_p = O(1)$, $n^{-1}\log(d) \to 0$, *and* $n^{-1}\lambda^{-2}\log(d\sqrt{n}) \to 0$.

For $\hat{w}_p = 1$, $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) = O_P(\lambda d^*)$.

For $\hat{w}_p = \hat{\sigma}_p$, *if we further assume* $\max_{p=1,\dots,d} \sigma_p = O(1)$ *and* $1 = O(\min_{p=1,\dots,d} \sigma_p)$, *then* $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) = O_P(\lambda d^*)$.

This result may be compared for instance with Theorem 4.1 of [22], in the framework of high dimensional regression. In this case an asymptotic convergence rate of $d^*\lambda$ may be similarly derived under the same assumptions (up to a $\log(n)$ factor) that $\log(d)n^{-1} \to 0$ and $\lambda^{-2}n^{-1} \to 0$. This shows that the optimal distortion may be asymptotically attained for dimension d of order e^{n^κ}, with $\kappa < 1$, choosing λ of order $n^{\frac{\kappa'-1}{2}}$, with $\kappa < \kappa' < 1$.

Moreover, Corollary 3.1 can provide a convergence rate of order $O(d^*\log(d)n^{-1/2})$ for the excess distortion of these Lasso-type procedures, up to a $\log(n)$ factor, hence adapting the sparsity of the optimal codebooks. In comparison to the $O(dn^{-1/2})$ rate that can be derived for the excess distortion of the k-means codebook (see, e.g., [3]), this suggests that regularized k-means might outperform standard k-means whenever $d^* << d$ and d is large. Some numerical illustration of this point is given below.

Numerical illustration: We consider the Gaussian mixture distributions with 4 components, each of them having covariance matrix I_d (identity matrix on $\mathbb{R}^d$), and with the following means:

$$\mu_1 = (\overbrace{0.8,\dots,0.8}^{5}, \overbrace{-0.8,\dots,-0.8}^{5}, \overbrace{0,\dots,0}^{d-10}), \qquad \mu_3 = -\mu_1,$$
$$\mu_2 = (\overbrace{0.8,\dots,0.8}^{10}, \overbrace{0,\dots,0}^{d-10}), \qquad \mu_4 = -\mu_2.$$

The weights of the mixture are chosen as $(0.3, 0.2, 0.2, 0.3)$. For d growing from 10 to 500, we compute the plain Lasso k-means codebooks with regularization parameter $\lambda(d) = 1.5 \times \log(d)/\sqrt{n}$, in the cases $n = 50$ and $n = 200$. Note that, since Gaussian mixture distributions have not a bounded support, this example does not fall in the scope of Theorem 3.1. This issue might be bypassed considering truncated Gaussian mixture distributions, as exposed in Section 3.4.

Following the approach of *Algorithm 1* of [24], the codebooks are computed using a Lloyd's-type algorithm: for any initial codebook, we update the assignments of data points to the closest code point and then update the code points to minimize the penalized squared distances to the previously assigned data points, using the Karush-Kuhn-Tucker condition that is necessary and sufficient when assignments are fixed. This procedure is repeated until convergence. Since every iteration decreases the penalized empirical distortion, the outcome of such an algorithm is clearly a local minimum of the penalized empirical distortion. This suggests that an effective global minimization of the penalized empirical distortion could be achieved by the comparison of the outcomes of several