Variable selection for k means quantization

Clément Levrard

Submitted to the Annals of Statistics

VARIABLE SELECTION FOR K-MEANS QUANTIZATION

By Clément Levrard

Université Paris Sud, UPMC and INRIA

Recent results in quantization theory provide theoretical bounds on the distortion of squared-norm based quantizers (see, e.g., [3] or [10]). These bounds are valid whenever the source distribution has a bounded support, regardless of the dimension of the underlying Hilbertian space.

However, it remains of interest to select relevant variable for quantization. This task is usually performed using coordinate energy-ratio thresholding (see, e.g., [1] or [17]), or maximizing a constrained empirical Between Cluster Sum of Squares criterion (see, e.g., [4] or [22]). This paper offers a Lasso type procedure to select the relevant variables for k-means clustering, as exposed in [18]. Moreover, some non-asymptotic convergence results on the distortion are derived for this procedure, along with consistency results toward sparse codebooks.

1. Introduction. Let P be a distribution over $\mathbb{R}^d$. Quantization is the issue 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.

The quantization theory was originally developed as a way to answer signal compression issues in the late 40's (see, e.g., [6]). 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., [5] or [7]). However, the choice of the Euclidean squared norm is convenient, since it allows to fully take

Keywords and phrases: k-means, variable selection, Lasso, margin condition

advantage of the Euclidean structure of $\mathbb{R}^d$, as described in [10]. Moreover, from a practical point of view, the k-means algorithm (see [11]) 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., [7] or [16]). These quantizers are quantizers of the type $x \mapsto \arg\min_{j=1,\dots,k} \|x - c_j\|$, 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 [15] 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 $P_n(A) = 1/n\,|\{i|X_i \in A\}|$, for every measurable subset 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

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

It is worth pointing out that, if $P^{(p)} \neq 0$, where $P^{(p)}$ denotes the marginal distribution of P on the j-th coordinate, then $\hat{\mathbf{c}}_n^{(p)} = (\hat{c}_1^{(p)}, \dots, \hat{c}_k^{(p)}) \neq 0$. This shows that the k-means algorithm does not provide sparse solutions, even if $P^{(j)}$ is a noise distribution.

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 [4] or [22]. Though these procedures offer good performance in classifying the sample $X_1, \dots, X_n$, under the assumption that the marginal distributions $P^{(j)}$ 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 [17]. 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.

This paper exposes a theoretical study of a Lasso type procedure combined with the k-means procedure, as suggested in [18]. Some results on the prediction performance and on the consistency to a sparse codebook are derived for this procedure, in the spirit of [21]. Some sparsity results on the empirical codebook are also given. It is worth pointing out that these results are valid when P satisfies a margin condition, as defined in [10], extending the scope of the asymptotic results proposed in [18].

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, technical proofs are to be found in Section 5.

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 exist 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 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

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

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 $R(\mathbf{c}) = R(Q) = P\gamma(\mathbf{c}, .)$, 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 $A \subset \mathbb{R}^d$.

It is worth pointing out that, if $P\|x\|^2 < \infty$, then there exist such minimizers $\hat{\mathbf{c}}_n$ and $\mathbf{c}^*$ (see, e.g., Theorem 4.12 in [7]). Throughout this paper it is assumed that there exists a unique optimal quantizer $\mathbf{c}^*$, up to relabeling code points.

To size the influence of the different coordinates on the quantization error, the following coordinate-wise quantization error and variance are introduced. Let $S \subset \{1, \dots, d\}$ denote a subset of coordinates, and $P^{(S)}$ denote the

marginal distribution of P over the set $\mathbb{R}^{|S|}$. We may define

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

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

$$\begin{cases} \sigma_S^2 &= \sigma_{S_1}^2 + \sigma_{S_2}^2, \\ \hat{\sigma}_S^2 &= \hat{\sigma}_{S_1}^2 + \hat{\sigma}_{S_2}^2, \\ R_S^* &\geq R_{S_1}^* + R_{S_2}^*, \\ \hat{R}_S^* &\geq \hat{R}_{S_1}^* + \hat{R}_{S_2}^*. \end{cases} \tag{1}$$

These elementary properties will be of importance when choosing which coordinate to select.

The following technical inequality is needed, in order to connect the loss $\ell(\mathbf{c}, \mathbf{c}^*)$ to the distance between codebooks.

DEFINITION 2.1. *Assume that there exists a unique optimal quantizer* $\mathbf{c}^*$. *Then* P *satisfies a margin condition if there exists* $\kappa_0 > 0$ *such that*

$$\forall \mathbf{c} \in C^k \quad \ell(\mathbf{c}, \mathbf{c}^*) \geq \kappa_0 \|\mathbf{c} - \mathbf{c}^*\|^2. \tag{2}$$

As exposed in [10], Definition 2.1 may be thought of as a margin condition in the framework of squared distance based quantization. Some examples of distributions satisfying (2) are given in [10]. Roughly, if P is well concentrated around k poles, then (2) will hold. It is also worth mentioning that the condition required in [18] is much stronger than the condition required in Definition 2.1, since it requires P 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. As exposed in [9], the condition mentioned above implies Definition 2.1.

The Lasso k-means procedure, introduced in [18], is defined as follows.

$$\hat{\mathbf{c}}_{n,\lambda} \in \arg\min_{\mathbf{c}\in C^k} P_n\gamma(\mathbf{c},.) + \lambda I(\mathbf{c}), \tag{3}$$

where $I(\mathbf{c})$ denotes a possibly weighted penalty function of the codebook $\mathbf{c}$. This paper provides results for two types of penalties $I(\mathbf{c})$: a Lasso type penalty where the weights are chosen to be 1, and a Weighted Lasso type penalty with adaptive weights.

Lasso type penalty

In this case the penalty function is chosen by

$$I(\mathbf{c}) = I_L(\mathbf{c}) = \sum_{p=1}^{d} \sqrt{c_1^{(p)2} + \ldots + c_k^{(p)2}}. \tag{4}$$

This L_1-type penalty is designed to drive the irrelevant (p)-th coordinates $c_1^{(p)}, \ldots, c_k^{(p)}$ together to zero, as exposed in [2]. The following proposition gives a theoretical guarantee on the coordinates which are not driven to zero.

PROPOSITION 2.1. *Let p be in $\{1, \ldots, d\}$. If*

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

then

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

Roughly, Proposition 2.1 ensures that the Lasso k-means procedure selects only variables whose empirical quantization error is small compared to its empirical variance. These variables may be interpreted as relevant variables for the empirical k-quantization error. However, when M_p is small, the choice of the penalty $I_L(\mathbf{c})$ will drive the (p)-th coordinates to 0, even if $P^{(p)}$ is supported on k points. This scaling issue can be addressed using a Weighted Lasso penalty, as done in [18].

Weighted Lasso type penalty

The original procedure of Lasso k-means exposed in [18] is indeed a Weighted Lasso type procedure. However, different weights are proposed here. For these weights theoretical guarantees are provided on the convergence of the Lasso k-means estimator to a sparse codebook. The proposed penalty function is the following

$$\hat{I}_{WL}(\mathbf{c}) = \sum_{p=1}^{d} \hat{\sigma}_p \sqrt{c_1^{(p)2} + \ldots + c_k^{(p)2}},$$

where the empirical coordinate-wise variances are defined above. The following proposition gives a necessary condition for the p-th coordinate not to be driven to 0.

PROPOSITION 2.2. *Let p be in $\{1, \dots, d\}$. If*

$$\sqrt{1 - \frac{\hat{R}_p^*}{\hat{\sigma}_p^2}} < \frac{\lambda}{2},$$

then

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

The scaling issue mentioned above turns out to be addressed, since only the ratios between empirical variances and empirical k-quantization error are to be considered to determinate relevant variables. As in the Lasso penalty case, coordinates with large ratios between empirical k-quantization error over empirical variance will be driven to zero.

It is worth mentioning that in these two cases non-zero coordinates are only empirically characterized. The following section provides convergence results to sparse codebooks, along with prediction results.

3. Results.

3.1. *Lasso k-means distortion and consistency.* Throughout this subsection the penalty function $I(\mathbf{c})$ is chosen as $I_L(\mathbf{c})$. It is well known that Lasso type procedures may be thought of as model selection procedures over L_1 balls (see, e.g., [13]). This leads to the following result.

THEOREM 3.1. *Let M_∞ denote $\max_{p=1,\dots,d} M_p$. Choose*

$$\lambda \geq \frac{6kM_\infty\sqrt{2\log(d)}}{\sqrt{n}}\left(1 + \frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right),$$

for some $x > 0$. Then, for every $\varepsilon > 0$, with probability larger than $1 - \left(\frac{\sqrt{k}dM_\infty}{\varepsilon} + 1\right)e^{-x}$, we have

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \inf_{r>0} \inf_{I_L(\mathbf{c}) \leq r} \left(\ell(\mathbf{c}, \mathbf{c}^*) + \lambda(2r + 3)\varepsilon\right).$$

For any codebook $\mathbf{c}$, let $\|\mathbf{c}\|_0$ be defined as $|\{p|\mathbf{c}^{(p)} \neq (0, \dots, 0)\}|$. Furthermore, assume that P satisfies (2). Then the best sparse approximation of $\mathbf{c}^*$ at order λ is defined by

$$\mathbf{c}_\lambda^* \in \arg\min_{\mathbf{c} \in C^k} 3R(\mathbf{c}) + \frac{8\lambda^2}{\kappa_0}\|\mathbf{c}\|_0,$$

where κ_0 denotes the constant in (2). As in the empirical case of Proposition 2.1, the non-zero coordinates of $\mathbf{c}_\lambda^*$ may be characterized in the following way.

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

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

then

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

The proof of Proposition 3.1 is given in Section 4. Equipped with this proposition, we are now in position to state convergence results.

THEOREM 3.2. *Denote by $M_\infty = \max_{p=1,\dots,d}$. There exists a constant c_L such that, if*

$$\lambda \geq c_L M_\infty \frac{\sqrt{k\log(kd)}}{\sqrt{n}} \left(1 + \frac{\sqrt{\log(d\sqrt{n}) + x}}{\sqrt{k\log(kd)}}\right),$$

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

$$\lambda I(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}_\lambda^*) \leq \left(3R(\mathbf{c}_\lambda^*) + \frac{8\lambda^2}{3\kappa_0}\|\mathbf{c}_\lambda^*\|_0\right) \vee \lambda^2. \tag{5}$$

Moreover, on the same event, the following prediction result holds

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \frac{4\lambda^2\|\mathbf{c}^*\|_0}{\kappa_0} \vee (2\lambda^2). \tag{6}$$

Theorem 3.2 can be considered as an application of Theorem 2.1 in [21] to the framework of vector quantization. It also may be noticed that the constant c_L depends on the constant in Sudakov's minoration (see, e.g., Proposition 3.15 in [12]), hence no explicit calculation of c_L is given. The consistency result shows that, provided that λ is chosen large enough, $\hat{\mathbf{c}}_{n,\lambda}$ converges toward the sparse approximation $\mathbf{c}_\lambda^*$ at a rate smaller than $d\lambda$. This $d\lambda$ rate corresponds to the case where $\mathbf{c}_\lambda^* = \mathbf{c}^*$, and is clearly suboptimal. Consequently much smaller rates are expected. The prediction result provides a distortion rate smaller than $d\lambda^2$. When d is large, this rate is of little interest. However, if a standard k means algorithm is performed on the set S of variable selected by the Lasso k-means procedure, in the spirit of [14], then hopefully a distortion rate of $k|S|M_\infty^2/n$ could be attained, compared to the best codebook based on this subset (see, e.g., Theorem 3.1 in [10])). As announced in Section 2, when $X^{(p)}$ has a small range, then the p-th coordinate will be driven to 0 by the Lasso k-means procedure, regardless of its separation capacity. To address this scaling issue, some results are given for a Weighted Lasso k-means procedure in the following subsection.

3.2. *Weighted Lasso k-means distortion and consistency.* In this section the penalty function is $\hat{I}_{WL}(\mathbf{c}) = \sum_{p=1}^{d} \hat{\sigma}_p \sqrt{c_1^{(p)2} + \ldots + c_k^{(p)2}}$. The fact that the weights $\hat{\sigma}_p$ depends on the sample will cause several theoretical troubles. To address this issue, this penalty function is connected to a deterministic penalty function, namely

$$I_{WL}(\mathbf{c}) = \sum_{p=1}^{d} \sigma_p \sqrt{c_1^{(p)2} + \ldots + c_k^{(p)2}}.$$

Denote by T the quantity $\max_{p=1,\ldots,d} \frac{M_p}{\sigma_p}$. The following proposition relates $\hat{I}_{WL}$ to I_{WL}.

PROPOSITION 3.2. *Suppose that* $2n > T^2\sqrt{\log(d)}$. *Then, for every* $y < \log(d)\left(\frac{2n}{T^2\sqrt{\log(d)}} - 1\right)^2$, *we have, with probability larger than* $1 - e^{-y}$, *for all* $\mathbf{c}$ *in* C^k,

$$\sqrt{1-\alpha(y)} I_{WL}(\mathbf{c}) \leq \hat{I}_{WL}(\mathbf{c}) \leq \sqrt{1+\alpha(y)} I_{WL}(\mathbf{c}), \tag{7}$$

where $\alpha(y) = \frac{T^2\sqrt{\log(d)}}{\sqrt{2n}}\left(1 + \sqrt{\frac{y}{\log(d)}}\right)$.

The proof of Proposition 3.2 is given in Section 4. Proposition 3.2 ensures that, provided that enough sample points are at disposal to correctly estimates the coordinate-wise variances, the data-driven penalty function $\hat{I}_{WL}(\mathbf{c})$ should be close to the deterministic penalty function $I_{WL}(\mathbf{c})$. Equipped with this proposition, some results can be derived for the k-means procedure with penalty $I_{WL}(\mathbf{c})$ which can be related to results for the Weighted Lasso k-means procedure we propose. This is the idea motivating the following results.

THEOREM 3.3. *Let* T *denote* $\max_{p=1,\ldots,d} \frac{M_p}{\sigma_p}$. *Let* $x > 0$, *and suppose that* $2n > T^2\sqrt{\log(d)}$. *Choose*

$$y < \log(d)\left(\frac{2n}{T^2\sqrt{\log(d)}} - 1\right)^2.$$

Suppose that

$$\lambda \geq \frac{1}{\sqrt{1-\alpha(y)}} \frac{6kM_\infty\sqrt{2\log(d)}}{\sqrt{n}} \left(1 + \frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right),$$

where $\alpha(y)$ is defined in Proposition 3.2. Then, for every $\varepsilon > 0$, with probability larger than $1 - e^{-y} - \left(\frac{\sqrt{k}\sigma^2 T}{\varepsilon} + 1\right) e^{-x}$, we have

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \inf_{r>0} \inf_{I_{WL}(\mathbf{c}) \leq r} \left(\ell(\mathbf{c}, \mathbf{c}^*) + \lambda(2r+4)\varepsilon\right).$$

As for the Lasso k-means case, Proposition 3.3 proves that the Weighted Lasso k-means codebook performs well in distortion compared to optimal codebooks over L_1-balls. As in Proposition 3.1, it is worth mentioning that Proposition 3.3 is valid even when P does not satisfy (2). The proof of Proposition 3.3 is postponed to Section 4.

For any codebook $\mathbf{c}$, let $S(\mathbf{c})$ be define as the set of coordinates p such that $(c_1^{(p)}, \ldots, c_k^{(p)}) \neq (0, \ldots, 0)$. As done in the previous section, let $\mathbf{c}_\lambda^*$ be defined as the sparse approximation of $\mathbf{c}^*$ at order λ, by

$$\mathbf{c}_\lambda^* = \arg\min_{\mathbf{c} \in C^k} 3R(\mathbf{c}) + \frac{8(1+\alpha)\lambda^2 \sigma^2_{S(\mathbf{c})}}{\kappa_0},$$

where α is a parameter which will be chosen as $\alpha(y)$, for some $y > 0$. The non-zero coordinates of $\mathbf{c}_\lambda^*$ may be characterized in the following way.

PROPOSITION 3.3. *Let p be in $\{1, \ldots, d\}$. If*

$$1 - \frac{R_p^*}{\sigma_p^2} < \frac{8(1+\alpha)\lambda^2}{3\kappa_0},$$

then

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

It is worth mentioning that the threshols takes into account only ratios of the type k-quantization error over variances, avoiding scaling issues. Equipped with this sparse approximation of $\mathbf{c}^*$, we are now in position to state the consistency and prediction results for the Weighted Lasso k-means procedure.

THEOREM 3.4. *Suppose that $2n > T^2\sqrt{\log(d)}$. Choose*

$$y < \log(d) \left(\frac{2n}{T^2\sqrt{\log(d)}} - 1\right)^2,$$

and $x > 0$. There exists a constant c_{WL} (the same as c_L), such that, if

$$\lambda \geq \frac{1}{\sqrt{1-\alpha(y)}} c_{WL} \sqrt{\frac{k\log(kd)}{n}} \left(1 + \sqrt{\frac{\log(\sigma^2\sqrt{n}) + x}{k\log(kd)}}\right),$$

where $\alpha(y)$ is defined in Proposition 3.2, then, with probability larger than $1 - e^{-x} - e^{-y}$, we have

$$\sqrt{1-\alpha(y)}\lambda I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda) \leq \left[3\ell(\mathbf{c}^*_\lambda, \mathbf{c}^*) + \frac{8(1+\alpha(y))\lambda^2\sigma^2_{S(\mathbf{c}^*_\lambda)}}{\kappa_0}\right] \vee \left[(1-\alpha(y))\lambda^2\right]. \tag{8}$$

Furthermore, on the same event, the following prediction result holds

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \left[\frac{8(1+\alpha(y))\sigma^2_{S(\mathbf{c}^*)}\lambda^2}{\kappa_0}\right] \vee \left[\sqrt{1-\alpha(y)}\lambda^2\right]. \tag{9}$$

As for the Lasso k-means case, Theorem 3.4 ensures that $\hat{\mathbf{c}}_{n,\lambda}$ is close to its sparse approximation, in the sense of I_{WL}, with a rate possibly much smaller than $\lambda\sigma^2$. This rate correspond to the case where the sparse approximation of $\mathbf{c}^*$ is $\mathbf{c}^*$. This leads to expect much smaller rates for the deviation between $\hat{\mathbf{c}}_{n,\lambda}$ and $\mathbf{c}^*_\lambda$. However, the prediction result is much more interesting, since it guarantees a distortion rate of $\sigma^2\lambda^2$ for the Weighted Lasso k-means procedure. As mentioned below Theorem 3.2, it is likely that this distortion rate could be improved by performing a standard k-means procedure on the set S of selected variables, possibly leading to a distortion rate of $k\sigma^2_S T^2_S/n$ (see, e.g., Theorem 3.1 in [10]), compared to the optimal codebook with support S.

4. Proofs. In this section the results are derived for a general penalty function

$$I_w(\mathbf{c}) = \sum_{p=1}^{d} w_p \sqrt{c_1^{(p)2} + \ldots + c_k^{(p)2}},$$

for any positive sequence $(w_p)_{p=1,\ldots,d}$. In the Lasso case, $w_p = 1$, whereas in the Weighted Lasso case $w_p = \hat{\sigma}_p$.

4.1. *Proof of Proposition 2.1 and Proposition 2.2.* Let $V_1, \ldots, V_k$ be a Voronoi partition associated with $\hat{\mathbf{c}}_{n,\lambda}$, and let $\hat{L}$ be the matrix of assignments, defined by

$$\hat{L}_{i,j} = 1_{X_i \in V_j}.$$

Suppose that $\hat{\mathbf{c}}^{(p)}_{n,\lambda} \neq 0$, where through this subsection $\hat{\mathbf{c}}^{(p)}_{n,\lambda}$ will denote the column vector $(\hat{c}^{(p)}_{n,\lambda,1}, \ldots, \hat{c}^{(p)}_{n,\lambda,k})^t$, and denote by $\hat{X}^{(p)}$ the column vector $(X_1^{(p)}, \ldots, X_k^{(p)})^t$. Then the Karush-Kuhn-Tucker condition, for the empirical

risk strategy penalized with $I_w(\hat{\mathbf{c}}_{n,\lambda})$, implies that (see, e.g., the proof of Theorem 2 in [18])

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

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

$$\left(\sum_{X_i\in V_1}(X_i^{(p)}-\hat{c}_{n,\lambda,1}^{(p)}),\ldots,\sum_{X_i\in V_k}(X_i^{(p)}-\hat{c}_{n,\lambda,k}^{(p)})\right),$$

it may be noted that

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

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

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

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

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

Since $\hat{R}_p(\hat{\mathbf{c}}_{n,\lambda})-\hat{R}_p(Q_1)\le\hat{\sigma}_p^2-\hat{R}_p$, (10) ensures that

$$\frac{\lambda w_p}{2}\le\sqrt{\hat{\sigma}_p^2-\hat{R}_p}.$$

Taking $w_p=1$ gives the result of Proposition 2.1 and $w_p=\hat{\sigma}_p$ gives the result of Proposition 2.2.

4.2. *Proof of Proposition 3.2.* Heffding's inequality ensures that, for every $p = 1, \dots, d$, $\frac{\hat{\sigma}_p^2}{\sigma_p^2} - 1$ is a subgaussian random variable with variance bounded by $\frac{T^4}{4n}$. For a comprehensive introduction to subgaussian random variables and its application to empirical processes theory, the interested reader is referred to [12]. Applying Theorem 3.12 in [12] and a bounded difference concentration inequality (see, e.g., Theorem 5.1 in [12]) yields, with probability larger than $1 - e^{-y}$,

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

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

4.3. *Proof of Proposition 3.1.* Let S be a subset of $\{1, \dots, d\}$, and let p be in S such that

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

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

$$\mathbf{c}_S^* = \arg \min_{S(\mathbf{c})=S} R(\mathbf{c}).$$

Then, according to (1), 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

$$3R(\mathbf{c}_{S\setminus\{p\}}^*) + \frac{8\lambda^2}{\kappa_0} \|\mathbf{c}_{S\setminus\{p\}}^*\|_0 < 3R(\mathbf{c}_S^*) + \frac{8\lambda^2}{\kappa_0} \|\mathbf{c}_S^*\|_0.$$

4.4. *Proof of Proposition 3.3.* Adopting the notation of the previous subsection, let p be in S such that $1 - \frac{R_p^*}{\sigma_p^2} < \frac{8(1+\alpha)\lambda^2}{\kappa_0}$. Then, it can be derived the same way as in the previous subsection that

$$R(\mathbf{c}_{S\setminus\{p\}}^*) - R(\mathbf{c}_S^*) \leq \sigma_p^2 - R_p^*.$$

This leads to

$$3R(\mathbf{c}_{S\setminus\{p\}}^*) + \frac{8\lambda^2(1+\alpha)}{\kappa_0} \sigma_{S\setminus\{p\}}^2 < 3R(\mathbf{c}_S^*) + \frac{8\lambda^2(1+\alpha)}{\kappa_0} \sigma_S^2.$$

4.5. *Proof of Theorem 3.1.* As in the proof of Proposition 2.1, throughout this subsection, the penalty function is chosen as $I_w(\mathbf{c})$, for a sequence of weights w. Let $T(w)$ denote the quantity

$$T(w) = \max_{p=1,\dots,d} \frac{M_p}{w_p}.$$

Then $T(w) = M_\infty$ in the Lasso case and $T(w) = T$ in the Weighted Lasso case. 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, $I_w(\mathbf{c}) \leq \bar{M}(w)$.

Let $\bar{\gamma}$ be defined as

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

for every $\mathbf{c}$ in C^k and x in $\mathbb{R}^d$. The following proposition, inspired from Theorem 2.1 in [3], offers an upper bound on the deviations between P_n and P on the set of possible $\bar{\gamma}$ constrained by $I_w(\mathbf{c})$.

PROPOSITION 4.1. *Suppose that w is deterministic. Let $x > 0$. Then, with probability larger than $1 - e^{-x}$, we have*

$$\sup_{I_w(\mathbf{c}) \leq r} (P - P_n)\bar{\gamma}(\mathbf{c}, .) \leq r\frac{6kT(w)\sqrt{2\log(d)}}{\sqrt{n}} \left(1 + \frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right).$$

It is worth mentioning that the requirements that w is deterministic prevents from directly choosing $w_p = \hat{\sigma}_p$. This issue will be addressed in the following subsection. Now choose $\lambda \geq \frac{6kT(w)\sqrt{2\log(d)}}{\sqrt{n}} \left(1 + \frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right)$, and let $\varepsilon > 0$. Define $K(\varepsilon) = \lceil \frac{\bar{M}(w)}{\varepsilon} \rceil$, that is the smallest integer larger than $\frac{\bar{M}(w)}{\varepsilon}$, and $\hat{m} = \lceil \frac{I_w(\hat{\mathbf{c}}_{n,\lambda})}{\varepsilon} \rceil$. Then, applying a union bound to Proposition 4.1, it follows that, with probability larger than $1 - K(\varepsilon)e^{-x}$, for all $m = 1, \dots, K(\varepsilon)$,

$$\sup_{I_w(\mathbf{c}) \leq m\varepsilon} (P - P_n)\bar{\gamma}(\mathbf{c}, .) \leq m\varepsilon\frac{6kT(w)\sqrt{2\log(d)}}{\sqrt{n}} \left(1 + \frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right).$$

On this event, we have

$$\begin{aligned} P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda}, .) + \lambda I_w(\hat{\mathbf{c}}_{n,\lambda}) &\leq \inf_{r>0} \inf_{I_w(\mathbf{c}) \leq r} (P_n\bar{\gamma}(\mathbf{c}, .) + \lambda r) \\ &\leq \inf_{m=1,\dots,K(\varepsilon)} \inf_{I_w(\mathbf{c}) \leq m\varepsilon} (P_n\bar{\gamma}(\mathbf{c}, .) + \lambda m\varepsilon). \end{aligned}$$

It follows that

$$\begin{aligned} P\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) &\leq P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) + \lambda\hat{m}\varepsilon \\ &\leq P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) + \lambda I_w(\hat{\mathbf{c}}_{n,\lambda}) + \lambda\varepsilon \\ &\leq \inf_{m=1,\dots,K(\varepsilon)} \inf_{I_w(\mathbf{c})\leq m\varepsilon} \left(P\bar{\gamma}(\mathbf{c},.) + \lambda(2m+1)\varepsilon\right). \end{aligned}$$

Adding $-P\bar{\gamma}(\mathbf{c}^*,.)$ on both sides leads to

$$\begin{aligned} \ell(\hat{\mathbf{c}}_{n,\lambda},\mathbf{c}^*) &\leq \inf_{m=1,\dots,K(\varepsilon)} \inf_{I_w(\mathbf{c})\leq m\varepsilon} \left(\ell(\mathbf{c},\mathbf{c}^*) + \lambda(2m+1)\varepsilon\right) \\ &\leq \inf_{r>0} \inf_{I_w(\mathbf{c})\leq r} \left(\ell(\mathbf{c},\mathbf{c}^*) + \lambda(2m+3)\varepsilon\right). \end{aligned}$$

Choosing $w_p = 1$ concludes the proof for the Lasso k-means procedure.

4.6. *Proof of Theorem 3.3.* The proof of Theorem 3.3 is almost the same as the proof of Theorem 3.1, with weights $w_p = \sigma_p$, leading to $T(w) = T$. To avoid confusion, $I_{WL}(\mathbf{c})$ will denote $I_w(\mathbf{c})$ with weights $w_p = \sigma_p$, and $\hat{I}_{WL}(\mathbf{c})$ will denote $I_w(\mathbf{c})$ with weights $w_p = \hat{\sigma}_p$. Let λ be larger than $\frac{1}{\sqrt{1-\alpha(y)}} \frac{6kM_\infty\sqrt{2\log(d)}}{\sqrt{n}} \left(1 + \frac{1}{2k}\sqrt{\frac{x}{\log(d)}}\right)$, then, with probability larger than $1 - e^{-y} - \left(\frac{\sqrt{k}dT}{\varepsilon} + 1\right) e^{-x}$ we have, for every $\mathbf{c}$ in C^k,

$$P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) + \lambda\hat{I}_{WL} \leq P_n\bar{\gamma}(\mathbf{c},.) + \sqrt{1+\alpha(y)}\lambda I_{WL}(\mathbf{c}).$$

It follows that

$$\begin{aligned} &P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) + \lambda\hat{I}_{WL}(\hat{\mathbf{c}}_{n,\lambda}) \\ &\leq \inf_{m=1,\dots,K(\varepsilon)} \inf_{I_{WL}(\mathbf{c})\leq m\varepsilon} P_n\bar{\gamma}(\mathbf{c},.) + \sqrt{1+\alpha(y)}m\varepsilon \\ &\leq \inf_{m=1,\dots,K(\varepsilon)} \inf_{I_{WL}(\mathbf{c})\leq m\varepsilon} P\bar{\gamma}(\mathbf{c},.) + (\sqrt{1+\alpha(y)}m\varepsilon + \sqrt{1-\alpha(y)}m\varepsilon \\ &\leq \inf_{r>0} \inf_{I_{WL}(\mathbf{c})\leq r} P\bar{\gamma}(\mathbf{c},.) + \sqrt{1+\alpha(y)}(r+\varepsilon) + \sqrt{1-\alpha(y)}(r+\varepsilon), \end{aligned}$$

where the middle inequality follows from Proposition 4.1. On the other, it may be written that

$$\begin{aligned} P\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) &\leq P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) + \sqrt{1-\alpha(y)}\lambda\hat{m}\varepsilon \\ &\leq P_n\bar{\gamma}(\hat{\mathbf{c}}_{n,\lambda},.) + \sqrt{1-\alpha(y)}\lambda\varepsilon + \lambda\hat{I}_{WL}(\hat{\mathbf{c}}_{n,\lambda}), \end{aligned}$$

according to Proposition 3.2. Combining these two inequalities and taking into account that $\sqrt{1-\alpha(y)} + \sqrt{1+\alpha(y)} \leq 2$ leads to the result.

4.7. *Proof of Theorem 3.2.* As done in the previous subsection, the results are derived for a generic penalty function

$$I_w(\mathbf{c}) = \sum_{p=1}^{d} w_p \sqrt{c_1^{(p)} 2 + \ldots + c_k^{(p)2}}.$$

The main argument of this proof relies on a comparison between $(P - P_n)(\bar{\gamma}(\mathbf{c}, .) - \bar{\gamma}(\mathbf{c}', .))$ and $I_w(\mathbf{c} - \mathbf{c}')$, stated in the following proposition.

PROPOSITION 4.2. *Suppose that w is deterministic. Denote by u the quantity $\log\left(\frac{\|w\|^2\sqrt{n}}{\sqrt{\log(kd)}}\right)$. There exists a constant $L > 1$ such that, if we denote by*

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

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 $1 - e^{-x}$,

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

where we recall that $\bar{M}(w) = \sqrt{k}\|w\|^2 T(w)$.

The proof of Proposition 4.2 relies on Section 3.4 in [21], and is postponed to the next section. The consistency result also relies on the following Lemma, which connects the L_1 penalty to the size of the support. For any subset $S \subset \{1, \ldots, d\}$ and vector x in $\mathbb{R}^d$, the truncated vector x_S is defined by

$$x_S^{(p)} = x^{(p)} 1_{p \in S}.$$

Moreover, let $S(\mathbf{c})$ denote the support of $\mathbf{c}$, that is the set of coordinates such that $(c_1^{(p)}, \ldots, c_k^{(p)}) \neq (0, \ldots, 0)$. At last, for a fixed $\mathbf{c}'$ in C^k, following the notation of [21], 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 [20].

LEMMA 4.1. *Let $\mathbf{c}'$ be a fixed codebook. Then, for every $\mathbf{c}$ in C^k and $\delta > 0$,*

$$2\lambda I_{w,1}(\mathbf{c}-\mathbf{c}') \leq \frac{1}{\delta}\ell(\mathbf{c},\mathbf{c}^*) + \frac{1}{\delta}\ell(\mathbf{c}',\mathbf{c}^*) + \frac{2\delta\lambda^2}{\kappa_0}\|w_{S(\mathbf{c}')}\|^2. \tag{12}$$

The proof of Lemma 4.1 can be found in [20]. 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 $2ab \leq \frac{\kappa_0}{\delta}a^2 + \frac{\delta}{\kappa_0}b^2$, and applying (2) leads to

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

□

Now turn to the case where $w = 1$, so that $\|w_{S(\mathbf{c}')}\|^2 = \|\mathbf{c}'\|_0$, and choose $\lambda \geq 2\lambda_1$. Let $\mathbf{c}'$ be a fixed codebook, to be chosen later. The fundamental Lasso inequality yields

$$P_n\gamma(\hat{\mathbf{c}}_{n,\lambda},.) + \lambda I_L(\hat{\mathbf{c}}_{n,\lambda}) \leq P_n\gamma(\mathbf{c}',.) + \lambda I_L(\mathbf{c}',.),$$

so that

$$P\gamma(\hat{\mathbf{c}}_{n,\lambda},.) + \lambda I_L(\hat{\mathbf{c}}_{n,\lambda}) \leq P\gamma(\mathbf{c}',.) + \lambda I_L(\mathbf{c}',.) + (P-P_n)(\gamma(\hat{\mathbf{c}}_{n,\lambda},.) - \gamma(\mathbf{c}',.)).$$

Splitting $I_L(\hat{\mathbf{c}}_{n,\lambda})$ in $I_{L,1}(\hat{\mathbf{c}}_{n,\lambda}) + I_{L,2}(\hat{\mathbf{c}}_{n,\lambda})$, it may be easily derived that $I_L(\mathbf{c}') - I_{L,1}(\hat{\mathbf{c}}_{n,\lambda}) \leq I_{L,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}')$ and $I_{L,2}(\hat{\mathbf{c}}_{n,\lambda}) = I_{L,2}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}')$. It follows that

$$\begin{aligned} P\gamma(\hat{\mathbf{c}}_{n,\lambda},.) + \lambda I_{L,2}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \leq P\gamma(\mathbf{c}',.) &+ \lambda I_{L,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \\ &+ (P-P_n)(\gamma(\hat{\mathbf{c}}_{n,\lambda},.) - \gamma(\mathbf{c}',.)). \end{aligned}$$

Consequently, Proposition 4.2 yields, with probability larger than $1 - e^{-x}$,

$$\begin{aligned}\ell(\mathbf{c}, \mathbf{c}^*) &+ \lambda I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \\ &\le \ell(\mathbf{c}', \mathbf{c}^*) + 2\lambda I_{L,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') + (P - P_n)(\gamma(\hat{\mathbf{c}}_{n,\lambda}, .) - \gamma(\mathbf{c}', .)) \\ &\le \ell(\mathbf{c}', \mathbf{c}^*) + \lambda_1(I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \vee \lambda_0) + 2\lambda I_{L,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}'). \qquad (13)\end{aligned}$$

Hence, applying Lemma 4.1 with $\delta = 2$ leads to

(14)
$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + 2\lambda I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \le 3\ell(\mathbf{c}', \mathbf{c}^*)\frac{8\lambda^2}{\kappa_0}\|\mathbf{c}'\|_0 + 2\lambda_1(I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \vee \lambda_0).$$

If $I(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \le \lambda_0$, then it is clear that $\lambda I(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \le \lambda^2$. Otherwise, we have

$$2\lambda I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \le 3\ell(\mathbf{c}', \mathbf{c}^*)\frac{8\lambda^2}{\kappa_0}\|\mathbf{c}'\|_0 + 2\lambda_1 I(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}').$$

Since $\lambda \ge 2\lambda_1$, the consistency result easily follows, taking $\mathbf{c}' = \mathbf{c}^*_\lambda$. Let us turn to the prediction result.

If $I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \le \lambda_0$, then $I_{L,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \le \lambda_0$. Consequently, taking $\mathbf{c}' = \mathbf{c}^*$ in (13) yields

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \le \lambda_0\lambda_1 + 2\lambda\lambda_0.$$

Since $\lambda_0 \le \lambda_1 \le \lambda/2$, it may be easily derived that $\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \le 2\lambda^2$. If $I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) > \lambda_0$, then taking $\mathbf{c}' = \mathbf{c}^*$ in (14) ensures that

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + 2(\lambda - \lambda_1)I_L(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \le \frac{8\lambda^2}{\kappa_0}\|\mathbf{c}^*\|_0.$$

4.8. *Proof of Theorem 3.4.* Throughout this subsection, the sequence w will be chosen as $w_p = \sigma_p$, so that $T(w) = T$ and $\bar{M}(w) = \sqrt{k}\sigma^2 T$. Choose $\lambda \ge \frac{2}{\sqrt{1-\alpha(y)}}\lambda_1$, where λ_1 is defined in Proposition 4.2. By definition of the Weighted Lasso k-means procedure, we have

$$P_n\gamma(\hat{\mathbf{c}}_{n,\lambda}, .) + \lambda\hat{I}_{WL}(\hat{\mathbf{c}}_{n,\lambda}) \le P_n\gamma(\mathbf{c}', .) + \lambda\hat{I}_{WL}(\mathbf{c}').$$

As in the previous subsection, this leads to

$$\begin{aligned}\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + \lambda\hat{I}_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') &\le \ell(\mathbf{c}', \mathbf{c}^*) + 2\lambda\hat{I}_{WL,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \\ &\quad + (P - P_n)(\gamma(\hat{\mathbf{c}}_{n,\lambda}, .) - \gamma(\mathbf{c}', .)).\end{aligned}$$

Using Proposition 3.2 and Proposition 4.2, it easily follows that, with probability larger than $1 - e^{-x} - e^{-y}$,

(15)
$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + \lambda\sqrt{1-\alpha(y)} I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \leq \ell(\mathbf{c}', \mathbf{c}^*) + \lambda_1 (I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}') \vee \lambda_0) + 2\sqrt{1+\alpha(y)} I_{WL,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}').$$

Now, applying Lemma 4.1 with $\delta = \frac{1}{2\sqrt{1+\alpha(y)}}$ and choosing $\mathbf{c}' = \mathbf{c}^*_\lambda$ leads to

$$\frac{1}{2\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*)} + \lambda\sqrt{1-\alpha(y)} I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda) \leq \frac{3}{2}\ell(\mathbf{c}^*_\lambda, \mathbf{c}^*) + \frac{4(1+\alpha(y))\lambda^2 \sigma^2_{S(\mathbf{c}^*_\lambda)}}{\kappa_0} + \lambda_1 (I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda) \vee \lambda_0).$$

Recalling that $\lambda \geq \frac{2}{\sqrt{1-\alpha(y)}}\lambda_1$ and $\lambda_1 \geq \lambda_0$, if $I(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda) \leq \lambda_0$, then

$$\lambda I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda) \leq \sqrt{1-\alpha(y)}\lambda^2.$$

Otherwise, we have

$$\lambda I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*_\lambda) \leq \frac{1}{\sqrt{1-\alpha(y)}} \left[3\ell(\mathbf{c}^*_\lambda, \mathbf{c}^*) + \frac{8(1+\alpha(y))\lambda^2 \sigma^2_{S(\mathbf{c}^*_\lambda)}}{\kappa_0} \right].$$

Let us turn now to the prediction result. Suppose that $I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \leq \lambda_0$. Then, if $\mathbf{c}' = \mathbf{c}^*$, the Lasso inequality combined with Proposition 4.2 ensures that

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + \lambda \hat{I}_{WL,2}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \leq \lambda_1\lambda_0 + \lambda \hat{I}_{WL,1}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*),$$

which leads to, applying Proposition 3.2,

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \lambda_1\lambda_0 + \sqrt{1+\alpha(y)}\lambda\lambda_0 \leq \frac{\lambda^2}{2}\left(\sqrt{(1-\alpha(y))(1+\alpha(y))} + \frac{1}{2}(1-\alpha(y)) \right).$$

Since $\sqrt{1-\alpha(y)} + \sqrt{1+\alpha(y)} \leq 2$, it is easy to see that

$$\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) \leq \sqrt{1-\alpha(y)}\lambda^2.$$

Now if $I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) > \lambda_0$, choosing $\mathbf{c}' = \mathbf{c}^*$ in (15) and applying Lemma 4.1, with $\delta = \frac{1}{2\sqrt{1+\alpha(y)}}$, leads to

$$\frac{1}{2}\ell(\hat{\mathbf{c}}_{n,\lambda}, \mathbf{c}^*) + \frac{\sqrt{1-\alpha(y)}}{2}\lambda I_{WL}(\hat{\mathbf{c}}_{n,\lambda} - \mathbf{c}^*) \leq \frac{4(1+\alpha(y))\sigma^2_{S(\mathbf{c}^*)}\lambda^2}{\kappa_0}.$$

5. Technical proofs.

5.1. *Proof of Proposition 4.1.* This proof is a slight modification of a result in [3], namely Lemma 4.3. Introducing some independent Rademacher variables ε_i, $i = 1, \dots, n$, such that $\varepsilon_i = \pm 1$ with probability $1/2$, and applying the symmetrization principle (see, e.g., Section 2.2 in [8]) leads to

$$\mathbb{E} \sup_{I_w(\mathbf{c}) \le r} (P - P_n)\bar{\gamma}(\mathbf{c}, .) \le 2\mathbb{E}_X \mathbb{E}_\varepsilon \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \min_{j=1,\dots,k} -2 \langle X_i, c_j \rangle + \|c_j\|^2,$$

where $\mathbb{E}_Z$ means expectation with respect to the law of Z, for some random variable Z. Let us denote by $I_w(c)$ the norm $I_w(c) = \sum_{p=1}^{d} w_p |c_p|$, for a code point c in C. Proceeding by induction on k as done in Lemma 4.3 *ii)* in [3], we may write

$$\mathbb{E}_\varepsilon \sup_{I_w(c) \le r} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \min_{j=1,\dots,k} -2 \langle X_i, c_j \rangle + \|c_j\|^2 \le 2k \left[\mathbb{E}_\varepsilon \sup_{I_w(c) \le r} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle X_i, c \rangle + \frac{rT(w)}{2\sqrt{n}} \right].$$

At last, it is immediate that

$$\mathbb{E}_{X,\varepsilon} \sup_{I(c) \le r} \left\langle c, \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\rangle \le r \mathbb{E}_{X,\varepsilon} \sup_{p=1,\dots,d} \left| \sum_{i=1}^{n} \varepsilon_i \frac{X_i^{(p)}}{w_p} \right|.$$

When $X_1, \dots, X_n$ is fixed, Hoeffding's inequality ensures that, for every $p = 1, \dots, d$, $\sum_{i=1}^{n} \varepsilon_i \frac{X_i^{(p)}}{w_p}$ is subgaussian with variance $\frac{T(w)^2}{n}$. For a comprehensive introduction to subgaussian variables and its application to empirical processes theory the interested reader is referred to [12]. Applying Theorem 3.12 of [12] ensures that

$$\mathbb{E}_\varepsilon \sup_{p=1,\dots,d} \left| \sum_{i=1}^{n} \varepsilon_i \frac{X_i^{(p)}}{w_p} \right| \le \sqrt{2 \log(d)} \frac{T(w)}{\sqrt{n}},$$

which leads to

$$\mathbb{E} \sup_{I_w(\mathbf{c}) \le r} (P - P_n)\bar{\gamma}(\mathbf{c}, .) \le \frac{2kT(w)}{\sqrt{n}} r + \frac{4k\sqrt{2 \log(d)} T(w)}{\sqrt{n}} r.$$

Applying a bounded difference concentration inequality such as Theorem 5.1 in [12] leads to the desired result.

5.2. *Proof of Proposition 4.2.* For a fixed $\mathbf{c}'$ in $\mathbf{c}^k$, denote by $Z_r(\mathbf{c}')$ the following random variable

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

The following proposition offers a 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 there exists a constant $L>1$ such that, with probability larger than $1-e^{-x}$,*

$$Z_r(\mathbf{c}') \leq 16L\sqrt{\frac{k\log(kd)}{n}}rT(w)\left(1+\frac{1}{\sqrt{2}L}\sqrt{\frac{x}{k\log(kd)}}\right).$$

The proof of Proposition 5.1 can be found in the next subsection. Proposition 4.2 derives from a peeling argument, as in Section 3.4 of [21], combined with Proposition 5.1. Let a be such that $e^{-(a-1)}2\bar{M}\leq\lambda_0$, and take $u_0=\log(a)$. Then it is easy to see that $u_0\leq u$, where u is defined in Proposition 4.2. We may write

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

where the last inequality follows from Proposition 5.1 and the fact that $L>1$. 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 [10], and mainly relies on Talagrand's generic chaining principle (see, e.g., [19]). 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 $\mathbb{R}^d$,

$$\left|-2\left\langle x,c_j\right\rangle+\|c_j\|^2+2\left\langle x,c_j'\right\rangle-\|c_j'\|^2\right|\leq 4rT,$$

which leads to

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

As a consequence, a bounded difference concentration inequality (see, e.g., Theorem 5.1 in [12]) yields, with probability larger than $1-e^{-x}$,

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

It remains to bound from above $\mathbb{E}Z_r$. According to the symmetrization principle (see, e.g., Section 2.2 of [8]), we may write

$$\mathbb{E}Z_r \le 2\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{I_w(\mathbf{c}-\mathbf{c}')\le r} \frac{1}{n}\sum_{i=1}^n \varepsilon_i(\gamma(\mathbf{c},X_i)-\gamma(\mathbf{c}',X_i)),$$

where the ε_i's are independent Rademacher variables. Let $X_1,\dots,X_n$ be fixed, and define, for $\mathbf{c}$ such that $I_w(\mathbf{c}-\mathbf{c}')\le r$ the random variable

$$Y_{\mathbf{c}} = \sum_{i=1}^n \varepsilon_i\gamma(\mathbf{c},X_i).$$

Define the pseudo-distance $d_0(\mathbf{c},\mathbf{c}')$ by

$$d_0^2(\mathbf{c},\mathbf{c}') = \sum_{i=1}^n\sum_{j=1}^k 8\left\langle c_j-c_j'\right\rangle^2 + 2n\sum_{j=1}^k(\|c_j\|^2-\|c_j'\|^2)^2.$$

Since

$$(\gamma(\mathbf{c},X_i)-\gamma(\mathbf{c}',X_i))^2 \le \max_{j=1,\dots,k} 8\left\langle c_j-c_j'\right\rangle^2 + 2(\|c_j\|^2-\|c_j'\|^2)^2,$$

it is easy to see that, when $X_1,\dots,X_n$ is fixed, $Y_{\mathbf{c}_1}-Y_{\mathbf{c}_2}$ is a subgaussian random variable with variance smaller than $d_0^2(\mathbf{c}_1,\mathbf{c}_2)$. The main argument of our proof is the following Theorem 2.1.5 of [19].

THEOREM 5.1. *Let Y_v, $v\in\mathcal{V}$ denote a centered stochastic process indexed by $\mathcal{V}$, and X_v denote a centered Gaussian process indexed by the same set $\mathcal{V}$. Let d be a pseudo-distance over $\mathcal{V}$ such that*

i) $\forall\, v,v'\in\mathcal{V}$ $Y_v-Y_{v'}$ *is subgaussian with variance* $d^2(v,v')$,
ii) $\forall\, v,v'\in\mathcal{V}$ $\mathrm{Var}(X_v-X_{v'}) = d^2(v,v')$.

Then there exists a universal constant $L>1$ such that

$$\mathbb{E}\sup_{v\in\mathcal{V}}(Y_v-Y_{v_0}) \le L\mathbb{E}\sup_{v\in\mathcal{V}}(X_v-X_{v_0}),$$

where v_0 is a fixed element of $\mathcal{V}$.

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

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

where the ξ's and ξ''s are independent standard Gaussian random variables. It is worth noticing that, for all $\mathbf{c}_1$ and $\mathbf{c}_2$ in $\mathcal{V}$, $\mathrm{Var}(N_{\mathbf{c}_1} - N_{\mathbf{c}_2}) = d_0^2(\mathbf{c}_1, \mathbf{c}_2)$. Consequently, applying Theorem 5.1 to the set $\mathcal{V}$, equipped with the pseudo-distance d_0, yields

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

It follows that

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

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

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

It is worth noticing 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 [12] 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

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

Combining these two bounds leads to

$$\mathbb{E}Z_r(\mathbf{c}') \le 16L\sqrt{\frac{k\log(kd)}{n}}rT(w).$$

REFERENCES

[1] Antoniadis, A., Brossat, X., Cugliari, J. and Poggi, J.-M. (2013). Clustering functional data using wavelets. *Int. J. Wavelets Multiresolut. Inf. Process.* **11** 1350003, 30. MR3038615

[2] Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225. MR2417268 (2010a:68132)

[3] Biau, G., Devroye, L. and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory* **54** 781–790. MR2444554 (2009m:68221)

[4] Chang, X., Wang, Y., Li, R. and Xu, Z. (2014). Sparse K-Means with ℓ_∞/ℓ_0 Penalty for High-Dimensional Data Clustering. *ArXiv e-prints*.

[5] Fischer, A. (2010). Quantization and clustering with Bregman divergences. *J. Multivariate Anal.* **101** 2207–2221. MR2671211 (2012c:62188)

[6] Gersho, A. and Gray, R. M. (1991). *Vector quantization and signal compression*. Kluwer Academic Publishers, Norwell, MA, USA.

[7] Graf, S. and Luschgy, H. (2000). *Foundations of quantization for probability distributions. Lecture Notes in Mathematics* **1730**. Springer-Verlag, Berlin. MR1764176 (2001m:60043)

[8] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442 (2009h:62060)

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

[10] Levrard, C. (2014). Non Asymptotic Bounds for Vector Quantization. *ArXiv e-prints*.

[11] Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–137. MR651807 (84a:94012)

[12] Massart, P. (2007). *Concentration inequalities and model selection. Lecture Notes in Mathematics* **1896**. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MR2319879 (2010a:62008)

[13] Massart, P. and Meynet, C. (2011). The Lasso as an ℓ_1-ball model selection procedure. *Electron. J. Stat.* **5** 669–687. MR2820635 (2012m:62210)

[14] Massart, P. and Meynet, C. (2012). Some rates of convergence for the selected Lasso estimator. In *Algorithmic learning theory. Lecture Notes in Comput. Sci.* **7568** 17–33. Springer, Heidelberg. MR3042877

[15] Pollard, D. (1981). Strong consistency of k-means clustering. *Ann. Statist.* **9** 135–140. MR600539 (83c:62098)

[16] Pollard, D. (1982). A central limit theorem for k-means clustering. *Ann. Probab.* **10** 919–926. MR672292 (84c:60047)

[17] Steinley, D. and Brusco, M. J. (2008). Selection of variables in cluster analysis: an empirical comparison of eight procedures. *Psychometrika* **73** 125–144. MR2395296

[18] Sun, W., Wang, J. and Fang, Y. (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electron. J. Stat.* **6** 148–167. MR2879675

[19] Talagrand, M. (2005). *The generic chaining. Springer Monographs in Mathematics*. Springer-Verlag, Berlin Upper and lower bounds of stochastic processes. MR2133757 (2006b:60006)

[20] van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809 (2009h:62048)

[21] van de Geer, S. A. (2013). Generic chaining and the ℓ_1-penalty. *J. Statist. Plann. Inference* **143** 1001–1012. MR3029225

[22] Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.* **105** 713–726. MR2724855 (2011m:62219)