Robust Bregman Clustering

Claire Brécheteau, Aurélie Fischer, Clément Levrard

claire.brecheteau@ec-nantes.fr
aurelie.fischer@lpsm.paris
clement.levrard@lpsm.paris

May, 18, 2018

Abstract

Using a trimming approach, we investigate a k-means type method based on Bregman divergences for clustering data possibly corrupted with clutter noise. The main interest of Bregman divergences is that the standard Lloyd algorithm adapts to these distortion measures, and they are well-suited for clustering data sampled according to mixture models from exponential families. We prove that there exists an optimal codebook, and that an empirically optimal codebook converges a.s. to an optimal codebook in the distortion sense. Moreover, we obtain the sub-Gaussian rate of convergence for k-means $\frac{1}{\sqrt{n}}$ under mild tail assumptions. Also, we derive a Lloyd-type algorithm with a trimming parameter that can be selected from data according to some heuristic, and present some experimental results.

1 Introduction

Clustering is the problem of classifying data in groups of similar points, so that the groups are as homogeneous and at the same time as well separated as possible Duda et al. [2000]. There are no labels known in advance, so clustering is an unsupervised learning task. To perform clustering, one needs some distance-like function to serve as a proximity measure between points. A very famous clustering method is the standard k-means algorithm (see for instance Lloyd [1982]), based on the Euclidean distance.

Let $X_1, \ldots, X_n$ denote a sample of independent random observations with values in $\mathbb{R}^d$, with the same distribution as a generic random vector X with distribution P. For an integer $k \geq 1$, in order to group data items $X_1, \ldots, X_n$ in meaningful classes, the k-means procedure minimizes the so-called empirical distortion

$$R_n(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j \in [\![1,k]\!]} \|X_i - c_j\|^2,$$

over all possible cluster centers or codebooks $\mathbf{c} = (c_1, \ldots, c_k)$, with notation $[\![1, k]\!]$ for $\{1, 2, \ldots, k\}$. Here, P_n denotes the empirical measure associated to $X_1, \ldots, X_n$, defined by

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}},$$

for every Borel subset A of $\mathbb{R}^d$.

It has been shown by Banerjee et al. [2005b] that the standard k-means clustering algorithm (see for instance Lloyd [1982]) can be generalized to general Bregman divergences. Bregman divergences are a broad class of dissimilarity measures indexed by strictly convex functions. Introduced by Bregman [1967], they are useful in a wide range of areas, among which statistical learning and data mining (Banerjee et al. [2005b], Cesa-Bianchi and Lugosi [2006]), computational geometry Nielsen et al. [2007], natural sciences, speech

processing and information theory Gray et al. [1980]. Squared Euclidean, Mahalanobis, Kullback-Leibler and L^2 distances are particular cases of Bregman divergences.

A Bregman divergence is not necessary a true metric, since it may be asymmetric or fail to satisfy the triangle inequality. However, Bregman divergences fulfill an interesting projection property which generalizes the Hilbert projection on a closed convex set, as shown in Bregman [1967]. Moreover, Banerjee et al. [2005b] have established that there exists a relation between finite-dimensional Bregman divergences and exponential families. Although they are not true metrics, Bregman divergences satisfy some properties, such as non-negativity and separation, convexity in the first argument and linearity (see Banerjee et al. [2005b], Nielsen et al. [2007]). As Bregman divergences represent a natural tool to measure proximity between observations arising from various distributions, we use them here for clustering purpose.

The aim is to find a data-based codebook $\hat{\mathbf{c}}_n$ such that the clustering risk $R(\hat{\mathbf{c}}_n)$ gets close to the optimal risk $R_k^* := \inf_{\mathbf{c}} R(\mathbf{c})$ as the size of the data set grows.

The convergence properties of empirical distortion minimizers are now quite-well understood when the source distribution P is assumed to have a finite support Linder [2002], Fischer [2010], even in infinite-dimensional cases Biau et al. [2008], Levrard [2015]. In real data sets, the source signal is often corrupted by noise, violating in most cases the bounded support assumption. In practice, data are usually pre-processed via an outlier-removal step that requires an important quantity of expertise. From a theoretical point of view, this corruption issue might be tackled winsorizing or trimming classical estimators, or by introducing some new and robust estimators that can adapt heavy-tailed cases. Such estimators can be based on PAC-Bayesian or Median of Means techniques Catoni and Giulini [2018], Brownlees et al. [2015], Lecué and Lerasle [2017] for instance. In a nutshell, these estimator succeed in achieving subGaussian deviation bounds under mild tail condition such as bounded variances and expectations.

In the clustering framework, up to our knowledge, the only theoretically grounded robust procedure is to be found in Cuesta-Albertos et al. [1997], where a trimmed k-means heuristic is introduced. See also Garcí a Escudero et al. [2008] for heteroscedastic trimmed clustering. In some sense, this paper extends this trimming approach to the general framework of clustering with Bregman divergence. Our precise contributions are listed below.

1.1 Contribution

- We introduce and provide a robust Bregman clustering algorithm, based on the minimization of a trimmed distortion criterion R_h.

- We give theoretical evidences that such a criterion is adapted to heavy-tailed cases, for instance proving that minimizers exist whenever P has a first-order moment.

- We investigate the finite-sample behavior of trimmed empirical distortion minimizers, and show that our procedure attains state-of-the-art guarantees in terms of convergence towards the optimal. Namely, we derive sub-Gaussian deviation bounds for the excess trimmed distortion under bounded variance condition.

We also give numerical illustrations of the aforementioned results.

2 Clustering with trimmed Bregman divergence

2.1 Bregman divergences and distortion

A Bregman divergence is defined as follows.

Definition 1. *Let ϕ be a strictly convex $\mathcal{C}_1$ real-valued function defined on a convex set $\Omega \subset \mathbb{R}^d$. The Bregman divergence d_ϕ is defined for all $x, y \in \Omega$ by*

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla_y \phi, x - y \rangle.$$

For every codebook $\mathbf{c} = (c_1, c_2, \ldots, c_k) \in \Omega^{(k)}$, we set:

$$d_\phi(x, \mathbf{c}) := \min_{i \in [\![1,k]\!]} d_\phi(x, c_i).$$

Observe that, since ϕ is strictly convex, for all $x, y \in \mathbb{R}^d$, $d_\phi(x,y)$ is non-negative and equal to zero if and only if $x = y$ (see [Rockafellar, 1970, Theorem 25.1]). Note that by taking $\phi : x \mapsto \|x\|^2$, one gets $d_\phi(x,y) = \|x - y\|^2$.

Let us present a few other examples.

Examples

1. Exponential loss: $\phi : x \mapsto e^x$, from $\mathbb{R}$ to $\mathbb{R}$, leads to $d_\phi(x,y) = e^x - e^y - (x-y)e^y$.

2. Logistic loss: $\phi : x \mapsto x \ln x + (1-x) \ln(1-x)$, from $[0,1]$ to $\mathbb{R}$, leads to $d_\phi(x,y) = x \ln \frac{x}{y} + (1-x) \ln \left(\frac{1-x}{1-y} \right)$.

3. Kullback-Leibler: $\phi : x \mapsto \sum_{\ell=1}^d x_\ell \ln x_\ell$, from the $(d-1)-$simplex to $\mathbb{R}$, leads to $d_\phi(x,y) = \sum_{\ell=1}^d x_\ell \ln \frac{x_\ell}{y_\ell}$.

Let P be a distribution on $\mathbb{R}^d$, and $\mathbf{c}$ a codebook. The clustering performance of $\mathbf{c}$ will be measured via its *distortion*, namely

$$R(\mathbf{c}) = P d_\phi(u, \mathbf{c}),$$

where, for any function f and measure Q, $Q f(u)$ means integration of f with respect to Q. Whenever only $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ is available, we denote by $R_n(\mathbf{c})$ the corresponding empirical distortion (associated with P_n). In the case where P is a mixture of distributions belonging to an exponential family, there exists a natural choice of Bregman divergence, as detailed in Section 5. Standard Bregman clustering intends to infer a minimizer of R via minimizing R_n, and works well in the bounded support case (Fischer [2010]).

2.2 Trimmed optimal codebooks

As for classical mean estimation, plain k-means is sensitive to outliers. An attempt to address this issue is proposed in Cuesta-Albertos et al. [1997], Gordaliza [1991]: for a trim level $h \in (0, 1]$, both a codebook and a subset of P-mass larger than h (trimming set) are pursued. This heuristic can be generalized to our framework as follows.

For a measure Q on $\mathbb{R}^d$, we write $Q \ll P$ (i.e., Q is a sub-measure of P) if $Q(A) \leq P(A)$ for every Borel set A, let $\mathcal{P}_h$ denote the set $\mathcal{P}_h = \{\frac{1}{h} Q \mid Q \ll P, Q(\mathbb{R}^d) = h\}$, and $\mathcal{P}_{+h} = \cup_{s \geq h} \mathcal{P}_s$. In analogy with Cuesta-Albertos et al. [1997] optimal trimming sets and codebooks are designed to achieve the optimal h-trimmed k-variation,

$$V_{k,h} = \inf_{\tilde{P} \in \mathcal{P}_{+h}} \inf_{\mathbf{c} \in \Omega^k} \tilde{P} d_\phi(u, \mathbf{c}) := \inf_{\tilde{P} \in \mathcal{P}_{+h}} \inf_{\mathbf{c} \in \Omega^k} R(\tilde{P}, \mathbf{c}), \tag{1}$$

that is the best possible k-point distortion based on a normalized sub-measure of P.

If $\mathbf{c}$ is a fixed codebook, we denote by $B_\phi(\mathbf{c}, r)$ (resp $\bar{B}_\phi(\mathbf{c}, r)$) the open (resp. closed) Bregman ball with radius r, $\left\{ x \mid \sqrt{d_\phi(x, \mathbf{c})} < r \right\}$ (resp. $\leq$), and by $r_h(\mathbf{c})$ the smallest radius $r \geq 0$ such that

$$P(B_\phi(\mathbf{c}, r)) \leq h \leq P(\bar{B}_\phi(\mathbf{c}, r)). \tag{2}$$

We also denote by $r_{n,h}(\mathbf{c})$ the radius when the distribution is P_n. Note that $r_{n,h}(\mathbf{c})^2$ is the Bregman divergence to the $\lceil nh \rceil$ nearest-neighbor of $\mathbf{c}$ in $\mathbb{X}_n$. Now, if $\mathcal{P}_h(\mathbf{c})$ is defined as the set of measures $\tilde{P}$ in $\mathcal{P}_h$ that coincides with P on $B_\phi(\mathbf{c}, r_h(\mathbf{c}))$, with support included in $\bar{B}_\phi(\mathbf{c}, r_h(\mathbf{c}))$. An easy result is

3

Lemma 2. *For all* $\mathbf{c} \in \Omega^{(k)}$, $h \in (0,1]$, $\tilde{P} \in \mathcal{P}_h$ *and* $\tilde{P}_{\mathbf{c}} \in \mathcal{P}_h(\mathbf{c})$,

$$R(\tilde{P}_{\mathbf{c}}, \mathbf{c}) \leq R(\tilde{P}, \mathbf{c}).$$

Equality holds if and only if $\tilde{P} \in \mathcal{P}_h(\mathbf{c})$.

This lemma is a straightforward generalisation of results in [Cuesta-Albertos et al., 1997, Lemma 2.1], Gordaliza [1991] or Chazal et al. [2011]. As a consequence, for any codebook $\mathbf{c} \in \Omega^{(k)}$ we may restrict our attention to sub-measures in $\mathcal{P}_h(\mathbf{c})$.

Definition 3. *For* $\mathbf{c} \in \Omega^{(k)}$, *the* h-*trimmed distortion of* $\mathbf{c}$ *is defined by*

$$R_h(\mathbf{c}) = hR(\tilde{P}_{\mathbf{c}}, \mathbf{c}),$$

where $\tilde{P}_{\mathbf{c}} \in \mathcal{P}_h(\mathbf{c})$.

Note that since $R(\tilde{P}_{\mathbf{c}}, \mathbf{c})$ does not depend on the $\tilde{P}_{\mathbf{c}}$ whenever $\tilde{P}_{\mathbf{c}} \in \mathcal{P}_h(\mathbf{c})$, $R_h(\mathbf{c})$ is well-defined. As well, $R_{n,h}(\mathbf{c})$ will denote the trimmed distortion corresponding to the distribution P_n. Another simple property of sub-measures can be translated in terms of trimmed distortion.

Lemma 4. *Let* $0 < h < h' < 1$ *and* $\mathbf{c} \in \Omega^{(k)}$. *Then*

$$\frac{R_h(\mathbf{c})}{h} \leq \frac{R_{h'}(\mathbf{c})}{h'}.$$

Moreover, equality holds if and only if $P(B_\phi(\mathbf{c}, r_{h'}(\mathbf{c}))) = 0$.

As well, this lemma generalize previous results in Cuesta-Albertos et al. [1997], Gordaliza [1991]. Combining Lemma 2 and Lemma 4 ensure that the h-trimmed k-variation may be achieved via minimizing a h-trimmed distortion.

Proposition 5.

$$V_{k,h} = \frac{1}{h} \inf_{\mathbf{c} \in \Omega^{(k)}} R_h(\mathbf{c}).$$

This proposition is an extension of [Cuesta-Albertos et al., 1997, Proposition 2.3]. Therefore, a good trimmed codebook in terms of k-variation should minimize R_h.

Definition 6. *A* h-*trimmed* k-*optimal codebook is any element* $\mathbf{c}^*$ *in* $\arg\min_{\mathbf{c} \in \Omega^{(k)}} R_h(\mathbf{c})$.

As exposed below, under mild assumptions on P and ϕ, such a trimmed k-optimal codebook exists.

Theorem 7. *Assume that* $P\|u\| < +\infty$, ϕ *is* $\mathcal{C}^2$ *and strictly convex and* $F_0 = \overline{conv(supp(P))} \subset \overset{o}{\Omega}$, *then the set* $\arg\min_{\mathbf{c} \in \Omega^{(k)}} R_h(\mathbf{c})$ *is not empty.*

Note that Theorem 7 only requires $P\|u\| < +\infty$. This can be compared with the standard squared Euclidean distance case, where $P\|u\|^2 < +\infty$ is required for R to have minimizers. From now on we denote by $\mathbf{c}_h^*$ a minimizer of R_h, and by $\hat{\mathbf{c}}_n$ a minimizer of the empirical trimmed distortion $R_{n,h}$.

2.3   Bregman-Voronoi cells and centroid condition

Similarly to the Euclidean case, the clustering associated with a codebook $\mathbf{c}$ will be given by a tesselation of the ambient space. To be more precise, for $\mathbf{c} \in \Omega^{(k)}$ and $i \in [\![1, k]\!]$, the Bregman-Voronoi cell associated with c_i is $V_i(\mathbf{c}) = \{x \mid \forall j \neq i \quad d_\phi(x, c_i) \leq d_\phi(x, c_j)\}$. Some further results on the geometry of Bregman Voronoi cells might be found in Nielsen et al. [2007]. Since the $V_i(\mathbf{c})$'s do not form a partition, $W_i(\mathbf{c})$ will denote a subset of $V_i(\mathbf{c})$ so that $(W_1(\mathbf{c}), \ldots, W_k(\mathbf{c}))$ is a partition of $\mathbb{R}^d$ (for instance break the ties of the V_i's with respect to the lexicographic rule).

4

Proposition 8. *Let* $\mathbf{c} \in \Omega^{(k)}$ *and* $P_{\mathbf{c}} \in \mathcal{P}_h(\mathbf{c})$. *Assume that for all* $i \in [\![1,k]\!]$, $P_{\mathbf{c}}(W_i(\mathbf{c})) > 0$, *and denote by* $\mathbf{m}$ *the codebook that consists in local means of* $P_{\mathbf{c}}$, *i.e.* $m_i = P_{\mathbf{c}}(u\mathbb{1}_{W_i(\mathbf{c})}(u))/P_{\mathbf{c}}(W_i(\mathbf{c}))$. *Then*

$$R_h(\mathbf{c}) \geq R_h(\mathbf{m}),$$

with equality if and only if for all i *in* $[\![1,k]\!]$, $c_i = m_i$.

Proposition 8 emphasizes the key property that Bregman divergences are minimized by expectations (this is not the case for L_1 distance for instance), and are the only loss functions satisfying that property; see Banerjee et al. [2005a]. Thus, a straightforward approach to minimize R_h can be based on an iterative scheme, as depicted below.

3   Description of the algorithm

The following algorithm is inspired by the trimmed version of the Lloyd's algorithm; see Cuesta-Albertos et al. [1997], but is also a generalization of the Bregman clustering algorithm for uniform finite-supported measures; see [Banerjee et al., 2005b, Algorithm 1]. We assume that we observe $\{X_1, \ldots, X_n\} = \mathbb{X}_n$, and that the mass parameter h equals $\frac{q}{n}$ for some positive integer q. We also let C_j denote the subset of $[\![1,n]\!]$ corresponding to the j-th cluster.

Algorithm 1. *Bregman trimmed k-means*

- ***Input*** $\{X_1, \ldots, X_n\} = \mathbb{X}_n$, q, k

- ***Initialization*** *Sample* $c_1, c_2, \ldots c_k$ *from* $\mathbb{X}_n$ *without replacement,* $\mathbf{c}^{(0)} \leftarrow (c_1, \ldots, c_k)$.

- ***Iterations*** *Repeat until stabilization of* $\mathbf{c}^{(t)}$.

    - $NN_q^{(t)} \leftarrow$ *indices of the* q *smallest values of* $d_\phi(x, \mathbf{c}^{(t-1)})$, $x \in \mathbb{X}_n$.
    - *For* $j = 1, \ldots, k$, $C_j^{(t)} \leftarrow W_j(\mathbf{c}^{(t-1)})$.
    - *For* $j = 1, \ldots, k$, $c_j^{(t)} \leftarrow \dfrac{\sum_{X \in C_j^{(t)} \cap NN_q^{(t)}} X}{\left| C_j^{(t)} \cap NN_q^{(t)} \right|}$.

- ***Output*** $\mathbf{c}^{(t)}$, $C_1^{(t)}, \ldots, C_k^{(t)}$.

As for every EM-type algorithm, initialization is a crucial point that will not be investigated in this paper. Note however that Bregman adaptations of approximate minimization methods such as k-means ++ Arthur and Vassilvitskii [2007] could be an efficient way to address the initialization issue, at least in practice. An easy consequence of Proposition 8 for the empirical measure P_n associated with $\mathbb{X}_n$ is the following. For short we denote by $R_{n,h}$ the trimmed distortion associated with P_n.

Proposition 9. *Algorithm 1 converges to a local minimum of the function* $R_{n,h}$.

It is worth mentioning that in full generality the output of Algorithm 1 is not a global minimizer of $R_{n,h}$. However, it is likely that suitable clusterability assumptions as in Kumar and Kannan [2010], Tang and Monteleoni [2016], Levrard [2018] would lead to further guarantees on such an output.

4   Convergence of a trimmed empirical distortion minimizer

This section is devoted to investigate the convergence of a minimizer $\hat{\mathbf{c}}_n$ of the empirical trimmed distortion $R_{n,h}$. Throughout this section ϕ is assumed to be $\mathcal{C}^2$, and $F_0 = \overline{conv(supp(P))} \subset \mathring{\Omega}$. That is, the closure of the convex hull of the support of P is a subset of the interior of Ω.

We begin with a generalization of [Cuesta-Albertos et al., 1997, Theorem 3.4] whenever P has a unique optimal trimmed codebook $\mathbf{c}_h^*$.

Theorem 10. *Assume that there exists a unique minimizer $\mathbf{c}_h^*$ of R_h. If P is continuous and satisfies $P\|u\|^p < \infty$ for some $p > 2$, then:*

$$\lim_{n \to +\infty} D(\hat{\mathbf{c}}_n, \mathbf{c}_h^*) = 0 \ a.e.$$

where $D(\mathbf{c}, \mathbf{c}') = \min_{\sigma \in \Sigma_k} \max_{i \in [\![1,k]\!]} |c_i - c'_{\sigma(i)}|$ and Σ_k denotes the set of all permutations of $[\![1,k]\!]$. Moreover,

$$\lim_{n \to +\infty} R_{n,h}(\hat{\mathbf{c}}_n) = R_h(\mathbf{c}_h^*) \ a.e.$$

Observe that slightly milder conditions are required for the trimmed distortion of $\hat{\mathbf{c}}_n$ to converge towards the optimal at a parametric rate.

Theorem 11. *Assume that $P\|u\|^p < \infty$, where $p \geq 2$. Then, for n large enough, with probability larger than $1 - n^{-\frac{p}{2}} - 2e^{-x}$, we have*

$$R_h(\hat{\mathbf{c}}_n) - R_h(\mathbf{c}_h^*) \leq \frac{C_P}{\sqrt{n}}(1 + \sqrt{x}).$$

Note that Theorem 11 does not require a unique trimmed optimal codebook, and only requires an order 2 moment condition for $\hat{\mathbf{c}}_n$ to achieve a sub-Gaussian rate in terms of trimmed distortion. This condition is in line with the order 2 moment condition required in Brownlees et al. [2015] for a robustified estimator of $\mathbf{c}^*$ to achieve similar guarantees, as well as the finite-variance condition required in Catoni and Giulini [2018] in a mean estimation framework. To derive results in expectation, a technical additional condition is needed.

Corollary 12. *If in addition there exist $c_0 \in \mathring{\Omega}$ and ψ a convex function such that*

$$\sup_{c \in B_\phi(c_0,t) \cap F_0} \|\nabla_c \phi\| \leq \psi(t),$$

with $P\|u\|^2 \psi^2(\frac{k\|u\|}{h}) < +\infty$, $P\|u\|^2 < +\infty$ and $P\psi^2(\frac{k\|u\|}{h}) < +\infty$, then

$$\mathbb{E}\left(R_h(\hat{\mathbf{c}}_n) - R_h(\mathbf{c}_h^*)\right) \leq \frac{C_P}{\sqrt{n}}.$$

Note that such a function ψ exists in most of the classical cases. However the moment condition required by Corollary 12 might be quite stronger than the order 2 condition of Theorem 11, as illustrated below.

1. In the plain k-means case $\phi(x) = \|x\|^2$ and $\Omega = \mathbb{R}^d$, we can choose $c_0 = 0$ and $\psi(t) = 2t$. Then the condition of Corollary 12 boils down to $P\|u\|^4 < +\infty$.

2. In the case where $\phi(x) = \exp(x)$, $\Omega = \mathbb{R}$, we may also choose $c_0 = 0$, and $\psi(t) = \exp(t)$. The condition of Corollary 12 may be written as $Pu^2 < +\infty$, $Pu^2 \exp(\frac{2ku}{h}) < +\infty$, and $P\exp(\frac{2ku}{h}) < +\infty$.

5 Numerical experiments

In this part, we apply Algorithm 1 to mixtures of distributions belonging to some exponential family. As presented in Banerjee et al. [2005b], a distribution from an exponential family may be associated to a Bregman divergence, by Legendre duality of convex functions. For a particular distribution, the corresponding Bregman divergence is more adapted for the clustering than other divergences.

Recall that an exponential family associated to a proper closed convex function ψ defined on an open parameter space $\Theta \subset \mathbb{R}^d$ is a family of distributions $\mathcal{F}_\psi = \{P_{\psi,\theta} \mid \theta \in \Theta\}$, such that, for all $\theta \in \Theta$, $P_{\psi,\theta}$, defined on $\mathbb{R}^d$, is absolutely continuous with respect to some distribution P_0, with Radon-Nikodym density $p_{\psi,\theta}$ defined for all $x \in \Omega$ by:

$$p_{\psi,\theta}(x) = \exp(\langle x, \theta \rangle - \psi(\theta)).$$

For this model, the expectation of $P_{\psi,\theta}$ may be expressed as $\mu(\theta) = \nabla_\theta \psi$. We define

$$\phi(\mu) = \sup_{\theta \in \Theta} \left\{ \langle \mu, \theta \rangle - \psi(\theta) \right\}.$$

By Legendre duality, for all μ such that ϕ is defined, we get:

$$\phi(\mu) = \langle \theta(\mu), \mu \rangle - \psi(\theta(\mu)),$$

with $\theta(\mu) = \nabla_\mu \phi$. The density of $P_{\psi,\theta}$ with respect to P_0 can be rewritten using the Bregman divergence associated to ϕ as follows:

$$p_{\psi,\theta}(x) = \exp(-d_\phi(x,\mu) + \phi(x)).$$

In the next experiments, we use Gaussian, Poisson, Binomial and Gamma mixture distributions and the corresponding Bregman divergences. Figure 1 presents the 4 densities together with the functions ψ and ϕ, as well as the associated Bregman divergences d_ϕ.

Distribution	$p_{\psi,\theta}(x)$	θ	$\psi(\theta)$
Gaussian	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$	$\frac{a}{\sigma^2}$	$\frac{\sigma^2}{2}\theta^2$
Poisson	$\frac{\lambda^x \exp(-\lambda)}{x!}$	$\log(\lambda)$	$\exp(\theta)$
Binomial	$\frac{N!}{x!(N-x)!}q^x(1-q)^{N-x}$	$\log\left(\frac{q}{1-q}\right)$	$N\log\left(1+\exp(\theta)\right)$
Gamma	$\frac{x^{k-1}\exp(-\frac{x}{b})}{\Gamma(k)b^k}$	$-\frac{k}{\mu}$	$k\log\left(-\frac{1}{\theta}\right)$

Distribution	μ	$\phi(\mu)$	$d_\phi(x,\mu)$
Gaussian	a	$\frac{1}{2\sigma^2}\mu^2$	$\frac{1}{2\sigma^2}(x-\mu)^2$
Poisson	λ	$\mu\log(\mu)-\mu$	$x\log\left(\frac{x}{\mu}\right)-(x-\mu)$
Binomial	Nq	$\mu\log\left(\frac{\mu}{N}\right)+(N-\mu)\log\left(\frac{N-\mu}{N}\right)$	$x\log\left(\frac{x}{\mu}\right)+(N-x)\log\left(\frac{N-x}{N-\mu}\right)$
Gamma	kb	$-k+k\log\left(\frac{k}{\mu}\right)$	$\frac{k}{\mu}\left(\mu\log\left(\frac{\mu}{x}\right)+x-\mu\right)$

Figure 1: Exponential family distributions and associated Bregman divergences.

5.1 Trimming parameter selection

To assess the good behavior of our procedure with respect to outliers, we propose to perform the same experiment as in Banerjee et al. [2005b], but with a level of noise. First, we propose in this section a heuristic to select the trimming parameter q, that is, the number of points in the sample which are assigned to a cluster and not considered as noise. Our strategy is as follows: we let q vary from 1 to the sample size n, plot the curve $q \mapsto cost[q]$ where $cost[q]$ denotes the optimal empirical distortion at trimming level q, and choose $q^\star$ by seeking for a cut-point on the curve. Indeed, when the parameter q gets large enough, it is likely that the procedure begins to assign outliers to clusters, which dramatically deprecates the empirical distortion.

We consider mixture models of Gaussian, Poisson, Binomial and Gamma distributions. Each of the mixtures consists in three components with equal probabilities $\frac{1}{3}$. The means are set to 10, 20 and 40. We set the standard deviation for the Gaussian densities to 5, the number of trials for the Binomial distribution to 100, and the shape parameter for the Gamma distribution to 40. Each time, 100 points are sampled from the mixture distribution and 20 outliers are added, uniformly sampled on $[0, 200]$.

We use Algorithm 1 for each of these noisy mixture distributions, using the corresponding divergence, and also make the same experiment with the Cauchy distribution, using the squared Euclidean norm, that is, the Bregman divergence associated with the Gaussian distribution.

We get $q = 110$ for the Gaussian mixture, $q = 106$ for the Poisson mixture and $q = 105$ for the Cauchy mixture.The curves for selecting q are depicted in Figure 2 for Gaussian, Poisson and Cauchy mixtures. In Figure 3, we have plotted the clustering results associated to the selected parameters for these distributions.

Note that the selection of the trimming parameter works well in the Cauchy example and the associated clustering seems also good. Thus, the method can adapt to mixture models that are not necessarily from distributions in exponential families.
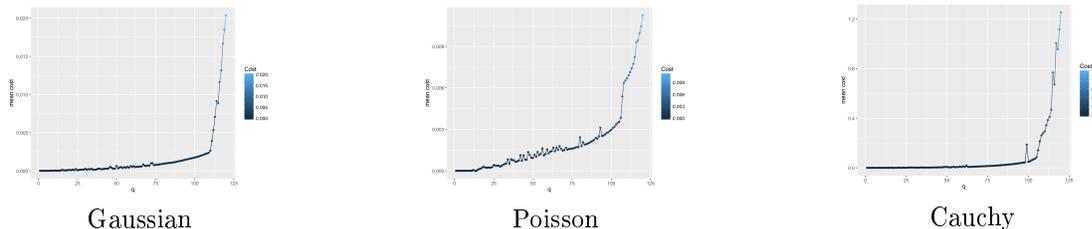

| Gaussian | Poisson | Cauchy |

Figure 2: Trimming parameter q selection.
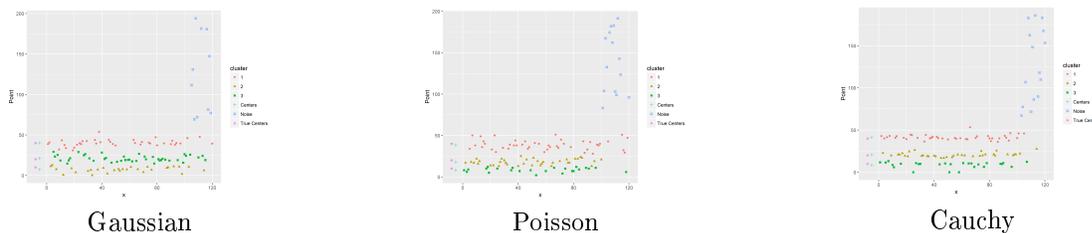

| Gaussian | Poisson | Cauchy |

Figure 3: Clustering associated to the selected parameter q.

5.2   Comparative performances of Bregman clustering for mixtures with noise

We build 100 120-samples with each time 100 points from the mixture distribution and 20 outliers uniformly sampled on $[0, 200]$, as in Section 5.1. Most of the time, with these amounts of points, our heuristic leads to a selection of q around 103. In the next experiments, we choose $q = 103$. To assess the performance of a partitioning, we compute the normalized mutual information NMI, as defined in Strehl and Ghosh [2002], between the true clusters and the clusters obtained with Algorithm 1. More specifically, the points corresponding to noise are assigned to one same cluster, and we compute the renormalized mutual information between the true partition, made of the three clusters arising from the mixture and the cluster of noisy points, and the clustering obtained with the trimming procedure with parameter $q = 103$. Confidence intervals for the normalized mutual information, with level around 5 percent, centered at the mean $mean(NMI)$, and with length $2 \times 1.96 \sqrt{\frac{var(NMI)}{100}}$, are derived in Figure 4. Note that, in general, the divergences associated to the distributions provide a better clustering, although it is less marked than for data without noise, as illustrated by the Binomial mixture case.

Model	$d_{Gaussian}$	$d_{Poisson}$	$d_{Binomial}$	d_{Gamma}
Gaussian	**0.6585 ± 0.0130**	0.6273 ± 0.0113	0.6370 ± 0.0108	0.5926 ± 0.0119
Poisson	0.6582 ± 0.0134	**0.6872 ± 0.0126**	0.6790 ± 0.0146	0.6500 ± 0.0120
Binomial	**0.7295 ± 0.0127**	0.7271 ± 0.01381	0.7290 ± 0.0125	0.6990 ± 0.0110
Gamma	0.7653 ± 0.0122	0.8122 ± 0.0110	0.7912 ± 0.0118	**0.8142 ± 0.0129**
Cauchy	**0.7074 ± 0.0149**	0.7204 ± 0.0128	0.6876 ± 0.01278	0.6729 ± 0.0116

Figure 4: Normalized mutual information for noisy data, adaptive choice of q.

We conduct the same experiment for mixtures in dimension 2, with the same parameters. More precisely, X_i is replaced by (X_i^1, X_i^2) where the X_i^j's are independent and from the same component of the original unidimensional mixture. We proceed as for dimension 1 with 100 points corresponding to signal and 20 additional points. The Bregman divergences used in the algorithm are the sum of the two coordinates of the corresponding 1-dimensional Bregman divergences. Using the same trimming parameter selection heuristic as above, we choose $q = 110$ for the Gaussian, Poisson and Gamma cases, and $q = 100$ for the Cauchy distribution and the Binomial case. The corresponding clustering results are depicted in Figure 5 for Gaussian, Poisson and Cauchy mixtures. In Figure 6, we computed the mutual information for the selected trimming parameter.

The heuristic for selecting the trimming set is not as effective as in dimension 1. It may come from the fact that the clutter noise is closer to the signal in dimension 2. The Bregman divergences for the Poisson, Binomial and Gamma distributions take into account the heteroscedasticity of the data, since the variance within a direction $(0, 1)$ or $(1, 0)$ depends on the mean within the direction. In the multidimensional Gaussian setting, Garcí a Escudero et al. [2008] proposed an algorithm, implemented in the R package tclust, that adapts to mixtures with different covariances matrices. It may be a future work to adapt our methods to non-Gaussian distributions with non-diagonal covariance matrices.
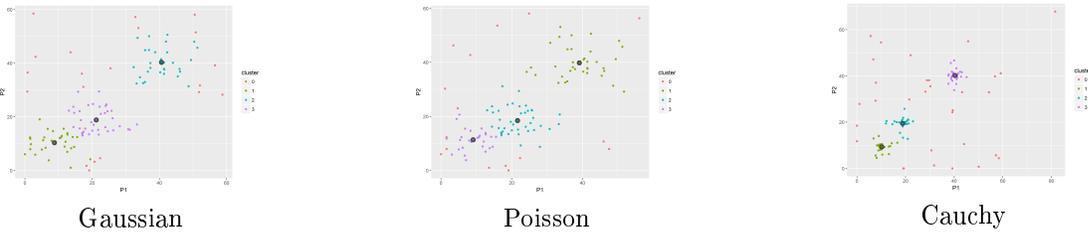

Gaussian   Poisson   Cauchy

Figure 5: Clustering associated to the selected parameter q (dimension 2).

Model	$d_{Gaussian}$	$d_{Poisson}$	$d_{Binomial}$	d_{Gamma}
Gaussian	$\mathbf{0.6543 \pm 0.0107}$	0.6356 ± 0.0094	0.6426 ± 0.0106	0.5732 ± 0.0017
Poisson	0.7237 ± 0.0095	$\mathbf{0.7450 \pm 0.0091}$	0.7432 ± 0.0085	0.7109 ± 0.0085
Binomial	0.7650 ± 0.0116	0.7592 ± 0.0102	$\mathbf{0.7728 \pm 0.0102}$	0.7073 ± 0.0083
Gamma	0.7800 ± 0.0073	0.8167 ± 0.0047	0.8038 ± 0.0059	$\mathbf{0.8349 \pm 0.0043}$
Cauchy	$\mathbf{0.6667 \pm 0.0115}$	0.6375 ± 0.0102	0.6400 ± 0.0119	0.6409 ± 0.0103

Figure 6: Normalized mutual information for noisy data (dimension 2), adaptive choice of q.

6 Proofs for Section 2.1

6.1 Intermediate results

The proof of Theorem 7, Theorem 10 and Theorem 11 make extensive use of the following technical lemmas. The first of them is a global upper bound on the radii $r_h(c)$, when c is in a compact subset of Ω.

Lemma 13. *Assume that ϕ is $\mathcal{C}^2$ and $F_0 = \overline{conv(supp(P))} \subset \mathring{\Omega}$. Then, for every $h \in (0, 1)$ and $K > 0$, there exists $r^+ < \infty$ such that*

$$\sup_{c \in F_0 \cap \overline{B}(0,K), s \leq h} r_s(c) \leq r^+.$$

As a consequence, if $\mathbf{c}$ is a codebook with a codepoint $c_{j_0} \in F_0$ satisfying $\|c_{j_0}\| \leq K$ and $s \leq h$,

$$r_s(\mathbf{c}) \leq r^+.$$

Proof of Lemma 13. Let K_+ be such that $P(B(0, K_+)) > h$. Thus, if $c \in \overline{B}(0, K)$, $P(B(c, K + K_+)) > h$. Since $B(c, K + K_+) \subset B(0, 2K + K_+)$, and ϕ is $\mathcal{C}^2$, according to the mean value theorem, there exists C_+ such that,

$$\forall x, y \in B(c, K + K_+) \cap F_0, \, d_\phi(x, y) \leq C_+ \|x - y\|.$$

Therefore, for every $c \in \overline{B}(0, K)$, $P\left(B_\phi\left(c, \sqrt{C_+(2K + K_+)}\right)\right) > h$.

Hence $r_s(c) \leq r_h(c) \leq \sqrt{C_+(2K + K_+)} = r^+$.

At last, if $\mathbf{c}$ is such that $c_{j_0} \in \overline{B}(0, K) \cap F_0$, then $B_\phi(c_{j_0}, r_h(c_{j_0})) \subset B_\phi(\mathbf{c}, r_h(c_{j_0}))$. Therefore $P(B_\phi(\mathbf{c}, r_h(c_{j_0}))) > s$, hence $r_s(\mathbf{c}) \leq r^+$. $\qquad\square$

Next, the following Lemma makes connections between the difference of Bregman divergences and distance between codebooks.

Lemma 14. *Assume that $F_0 \subset \mathring{\Omega}$ and ϕ is $\mathcal{C}^2$ on Ω. Then, for every $K > 0$, there exists $C_K > 0$ such that for every $\mathbf{c}$ and $\mathbf{c}'$ in $\overline{B}(0, K) \cap F_0$, and $x \in \Omega$,*

$$|d_\phi(x, \mathbf{c}) - d_\phi(x, \mathbf{c}')| \leq C_K D(\mathbf{c}, \mathbf{c}') \left(1 + \|x\|\right),$$

where we recall that $D(\mathbf{c}, \mathbf{c}') = \min_{\sigma \in \Sigma_k} \max_{j \in [\![1, k]\!]} |c_j - c'_{\sigma(j)}|$ with Σ_k the set of all permutations of $[\![1, k]\!]$.

Proof of Lemma 14. The set $F_0 \cap \overline{B}(0, K)$ is a convex compact subset of $\mathring{\Omega}$. Let $x \in \mathbb{R}^d$ and $\mathbf{c}, \mathbf{c}' \in F_0 \cap \overline{B}(0, K)$. Since ϕ and $x \mapsto \nabla \phi(x)$ are $\mathcal{C}^1$, the mean value theorem yields that

$$
\begin{aligned}
\left|d_\phi(x, c_j) - d_\phi(x, c'_j)\right| &\leq \left|\phi(c'_j) - \phi(c_j)\right| + \left|\left\langle x, \nabla_{c'_j} \phi - \nabla_{c_j} \phi \right\rangle\right| + \left|\left\langle \nabla_{c'_j} \phi, c'_j \right\rangle - \left\langle \nabla_{c_j} \phi, c_j \right\rangle\right| \\
&\leq C_K \|c_j - c'_j\|(1 + \|x\|),
\end{aligned}
$$

for some constant C_K. Thus,

$$|d_\phi(x, \mathbf{c}) - d_\phi(x, \mathbf{c}')| \leq C_K(1 + \|x\|) \max_j \|c_j - c'_j\|$$

and

$$|d_\phi(x, \mathbf{c}) - d_\phi(x, \mathbf{c}')| \leq C_K(1 + \|x\|)D(\mathbf{c}, \mathbf{c}').$$

$\qquad\square$

We will also need a continuity result on the function $(s, \mathbf{c}) \mapsto R_s(\mathbf{c})$, stated as the following Lemma.

Lemma 15. *Assume that $F_0 \subset \mathring{\Omega}$, $P\|u\| < \infty$ and ϕ is $\mathcal{C}^2$ on Ω. Then the map $(s, \mathbf{c}) \to R_s(\mathbf{c})$ is continuous. Moreover, for every $h \in (0, 1)$, $\epsilon > 0$ and $K > 0$, there is $s < h$ such that*

$$\sup_{\mathbf{c} \in \left(F_0 \cap \overline{B}(0, K)\right)^{(k)}} R_h(\mathbf{c}) - R_s(\mathbf{c}) \leq \epsilon.$$

Proof of Lemma 15. According to Lemma 2 and Lemma 14, for every $h \in (0, 1)$,

$$
\begin{aligned}
R_h(\mathbf{c}) - R_h(\mathbf{c}') &\leq h(P_{\mathbf{c}', h} d_\phi(u, \mathbf{c}) - P_{\mathbf{c}', h} d_\phi(u, \mathbf{c}')) \\
&\leq h P_{\mathbf{c}', h} |d_\phi(u, \mathbf{c}) - d_\phi(u, \mathbf{c}')| \\
&\leq C_K D(\mathbf{c}, \mathbf{c}')(1 + P\|u\|),
\end{aligned}
$$

for some $C_K > 0$. As a consequence, $|R_h(\mathbf{c}) - R_h(\mathbf{c}')| \to 0$ when $D(\mathbf{c}, \mathbf{c}') \to 0$. Now, note that for every $s < h$,

$$
\begin{aligned}
R_h(\mathbf{c}) - R_s(\mathbf{c}) &= P d_\phi(u, \mathbf{c}) \left(\mathbb{1}_{B_\phi(\mathbf{c}, r_h(\mathbf{c}))}(u) - \mathbb{1}_{B_\phi(\mathbf{c}, r_s(\mathbf{c}))}(u)\right)(u) \\
&\leq r_h(\mathbf{c})(h - s)
\end{aligned}
$$

Moreover, according to Lemma 13, $\sup_{\mathbf{c} \in \left(F_0 \cap \overline{B}(0, K)\right)^{(k)}} r_h(\mathbf{c}) \leq r^+$ for some $r^+ < \infty$, hence the result. $\qquad\square$

6.2 Proof of Lemma 2

For $u \in [0,1]$, let $F_{\mathbf{c}}^{-1}(u) = r_u^2(\mathbf{c})$ denote the u-quantile of the random variable $d_\phi(X, \mathbf{c})$ for $X \sim P$. That is,

$$F_{\mathbf{c}}^{-1}(u) = \inf \{r \geq 0 \mid P(B_\phi(\mathbf{c}, r)) > u\}.$$

If $\tilde{F}_{\mathbf{c}}^{*\,-1}(u)$ denotes the u-quantile of $d_\phi(\tilde{X}^*, \mathbf{c})$, for $\tilde{X}^* \sim P_{\mathbf{c}} \in \mathcal{P}_h(\mathbf{c})$, it holds $\tilde{F}_{\mathbf{c}}^{*\,-1}(u) = F_{\mathbf{c}}^{-1}(hu)$. Let U be a uniform random variable on $[0,1]$, then $F_{\mathbf{c}}^{*\,-1}(U)$ and $d_\phi(\tilde{X}^*, \mathbf{c})$ have the same distribution. Thus, we may write:

$$R(\tilde{P}_{\mathbf{c}}, \mathbf{c}) = \mathbb{E}_{\tilde{X}^*} d_\phi(\tilde{X}^*, \mathbf{c})$$
$$= \int_{u=0}^1 F_{\mathbf{c}}^{-1}(hu) du$$

Let $\tilde{P} \in \mathcal{P}_h(P)$ be a Borel probability measure on Ω such that $h\tilde{P}$ is a sub-measure of P, and let $\tilde{F}_{\mathbf{c}}^{-1}(u)$ denote the u-quantile of $d_\phi(\tilde{X}, \mathbf{c})$ for $\tilde{X} \sim \tilde{P}$. Since $P(B_\phi(\mathbf{c}, r)) \geq h\tilde{P}(B_\phi(\mathbf{c}, r))$, it holds that $\tilde{F}_{\mathbf{c}}^{-1}(u) \geq F_{\mathbf{c}}^{-1}(hu)$. Thus, we may write

$$R(\tilde{P}, \mathbf{c}) = \int_{u=0}^1 \tilde{F}_{\mathbf{c}}^{-1}(u) du$$
$$\geq R(\tilde{P}_{\mathbf{c}}, \mathbf{c}).$$

Note that equality holds if and only if $\tilde{F}_{\mathbf{c}}^{-1}(u) = \tilde{F}_{\mathbf{c}}^{*\,-1}(u)$ for almost all $u \in [0,1]$, that is $\tilde{P} \in \mathcal{P}_h(\mathbf{c})$.

6.3 Proof of Lemma 4

Set $0 < h < h' < 1$, and recall that $F_{\mathbf{c}}^{-1}(u) = r_u^2(\mathbf{c})$ denote the u-quantile of the random variable $d_\phi(X, \mathbf{c})$ for $X \sim P$ and $u \in [0,1]$. Since $F_{\mathbf{c}}^{-1}$ is non-decreasing, we may write

$$\frac{R_h(\mathbf{c})}{h} = \int_{u=0}^1 F_{\mathbf{c}}^{-1}(hu) du$$
$$\leq \int_{u=0}^1 F_{\mathbf{c}}^{-1}(h'u) du$$
$$= \frac{R_{h'}(\mathbf{c})}{h'}.$$

Equality holds if and only if $F_{\mathbf{c}}^{-1}(hu) = F_{\mathbf{c}}^{-1}(h'u)$ for almost all $u \in [0,1]$. Since $F_{\mathbf{c}}^{-1}$ is non-decreasing and right-continuous, $F_{\mathbf{c}}^{-1}(hu) = F_{\mathbf{c}}^{-1}(h'u)$ entails $F_{\mathbf{c}}^{-1}(l) = F_{\mathbf{c}}^{-1}(0)$, for all $l < h'$, or, in other words, $r_l^2(\mathbf{c}) = r_0^2(\mathbf{c})$. From (2), it follows that $P(B_\phi(\mathbf{c}, r_{h'}(\mathbf{c}))) = 0$. Conversely, equality holds when $P(B_\phi(\mathbf{c}, r_{h'}(\mathbf{c}))) = 0$.

6.4 Proof of Theorem 7

In the following, we will use the more concise notation $R_{k,h}^*$ for the optimal risk $hV_{k,h} = \inf_{\mathbf{c} \in \Omega^{(k)}} R_h(\mathbf{c})$. For $s \in]0,1]$, let $T_s : \Omega^{(k)} \to \Omega^{(k)}$ denote the operator that maps a codebook $\mathbf{c}$ onto its local means, that is

$$T_s(\mathbf{c})_j = \frac{P_{\mathbf{c}_s} u \mathbb{1}_{W_j(\mathbf{c})}(u)}{P_{\mathbf{c},s}(W_j(\mathbf{c}))},$$

with $P_{\mathbf{c},s} \in \mathcal{P}_s(\mathbf{c})$ and adopting the convention that $T(\mathbf{c})_j = 0$ whenever $P_{\mathbf{c}}(W_j(\mathbf{c})) = 0$. The intuition behind the proof of Theorem 7 is that optimal codebooks can be found as iterations of T, and that T is a bounded operator whenever each Voronoi cell has enough weight. This idea is summarized by the following Lemma, that encompasses Theorem 7.

Lemma 16. *For every $k \geq 2$, if $R^*_{k-1,h} - R^*_{k,h} > 0$, then*

$$\alpha := \min_{j=2,\ldots,k} R^*_{j-1,h} - R^*_{j,h} > 0.$$

Moreover there exists $0 < h^- < h < h^+ < 1$ and C_{h^-,h^+} such that, for every $j \in [\![2,k]\!]$ and $s \in [h^-,h^+]$,

- $R^*_{j-1,h^-} - R^*_{j,h^+} \geq \frac{\alpha}{2}$.

- *For every $\frac{\alpha}{4}$-minimizer $\mathbf{c}^*_{j,s}$ of $R^*_{j,s}$, $\sup_{p \in [\![1,j]\!]} \|T_s(\mathbf{c}^*_{j,s})_p\| \leq C_{h^-,h^+}$.*

- *There is a minimizer $\mathbf{c}^*_{j,s}$ of $R^*_{j,s}$ such that $\forall p \in [\![1,j]\!]$, $\|c^*_{j,s,p}\| \leq C_{h^-,h^+}$.*

Proof of Lemma 16. First note that if there exists $j \leq k$ such that $R^*_{j-1,h} - R^*_{j,h} = 0$, then there exists a set A with $P(A) \geq h$ such that $P_{|A}$ is supported on $j-1$ points. Thus, $R^*_{k-1,h} = R^*_{k,h} = 0$. As a consequence, when $R^*_{k-1,h} - R^*_{k,h} > 0$, α is positive.

Note also that the third point follows on from the second point. Indeed, for every sequence $\mathbf{c}^{*\,(n)}_{k,s}$ of $\frac{\alpha}{4n}$ minimizers of $R^*_{k,s}$, for every $i \in [\![1,k]\!]$, $\|T_s(\mathbf{c}^{*\,(n)}_{k,s})_i\| \leq C_{h^-,h^+}$. Since $\left(\overline{B}(0,C_{h^-,h^+}) \cap F_0\right)^{(k)}$ is a compact set, the limit in $\left(\overline{B}(0,C_{h^-,h^+}) \cap F_0\right)^{(k)}$ of any converging subsequence of $(T_s(\mathbf{c}^{*\,(n)}_{k,s}))_n$ is a minimizer of $R^*_{k,s}$.

Now assume that $R^*_{k-1,h} - R^*_{k,h} > 0$, In order to prove the other points, we proceed recursively. Assume that $k = 2$. Since, for $s > 0$ and any 1-point codebooks c, $\|T_s(c)\| \leq P\|u\|/s$, optimal 1-point codebooks can be found in $B(0,C_1) \cap F_0$. From a compactness argument there exists an optimal 1-point codebook $\mathbf{c}^*_{1,s}$ satisfying $\|\mathbf{c}^*_{1,s}\| \leq P\|u\|/s$.

Denote by $\mathbf{c}^*_{1,h^-}$ a minimizer of R^*_{1,h^-}, and $\mathbf{c}^*_{2,h}$ an $\frac{\alpha}{8}$ minimizer of $R^*_{2,h}$. According to Lemma 15, for a fixed $\mathbf{c}$, $s \mapsto R_s(\mathbf{c})$ is continuous, thus we may choose h^+ such that $R_{h^+}(\mathbf{c}^*_{2,h}) \leq R_h(\mathbf{c}^*_{2,h}) + \frac{\alpha}{8}$. Then,

$$\begin{aligned}
R^*_{2,h^+} &\leq R_{h^+}(\mathbf{c}^*_{2,h}) \\
&\leq R_h(\mathbf{c}^*_{2,h}) + \frac{\alpha}{8} \\
&\leq R^*_{2,h} + \frac{\alpha}{4}.
\end{aligned}$$

On the other hand, set $h_1 = \frac{h}{2}$. Then $\sup_{s \geq h_1} \|\mathbf{c}^*_{1,s}\| \leq \frac{P\|u\|}{h_1} = C_{h_1}$. According to Lemma 15, there exists $h > h_2 \geq h_1$ such that $\sup_{\|c\| \leq C_{h_1}}(R_h(c) - R_{h_2}(c)) \leq \frac{\alpha}{4}$. For such an h_2, we may write

$$\begin{aligned}
R^*_{1,h_2} &= R_{h_2}(\mathbf{c}^*_{1,h_2}) \\
&\geq R_h(\mathbf{c}^*_{1,h_2}) - \frac{\alpha}{4} \\
&\geq R^*_{1,h} - \frac{\alpha}{4}.
\end{aligned}$$

Since $R^*_{k,h} - R^*_{k-1,h} \geq \alpha$, it comes that $R^*_{1,h_2} - R^*_{2,h^+} \geq \frac{\alpha}{2}$.

Now, if $\mathbf{c}$ is an $\alpha/4$-minimizer of $R^*_{2,s}$, for $s \geq h - (h - h_2)/2 := h^-$, its Bregman-Voronoï cells restricted to the s-trimming set, $V_{j,s}$, have weight not smaller than $h - h^-$. Indeed, suppose that $P(V_{1,s}) < h - h^-$. Then

$$\begin{aligned}
R^*_{2,h^+} &\geq R_s(\mathbf{c}) - \frac{\alpha}{4} \\
&\geq P d_\phi(u,c_2) \mathbb{1}_{V_{2,s}}(u) - \frac{\alpha}{4} \\
&\geq R^*_{1,h^-} - \frac{\alpha}{4} \\
&\geq R^*_{1,h_2} - \frac{\alpha}{4},
\end{aligned}$$

hence the contradiction. Choosing $s \geq h^-$ entails that $\|T_s(c_{j,s}^*)_p\| \leq \frac{P\|u\|}{h-h^-}$, this gives the result for $k = 2$.

Assume that the proposition is true for index $k - 1$, we will prove that it is also true for index k. Set $\alpha = \min_{j \in [\![2,k]\!]} R_{j-1,h}^* - R_{j,h}^* > 0$. Set h^{--} and h^{++} the elements h^- and h^+ associated with step $k - 1$. Set $\mathbf{c}_{k-1,h^-}^*$ a minimizer of R_{k-1,h^-}^* and $\mathbf{c}_{k,h}^*$ an $\frac{\alpha}{8}$-minimiser of $R_{k,h}^*$. According to Lemma 15, there exists $h < h^+ < h^{++}$ such that $R_{h^+}(\mathbf{c}_{k,h}^*) \leq R_h(\mathbf{c}_{k,h}^*) + \frac{\alpha}{8}$. Thus $R_{k,h^+}^* \leq R_{k,h}^* + \frac{\alpha}{4}$.

On the other hand, Lemma 15 provides $h > h_1 > h^{--}$ such that

$$\sup_{\mathbf{c} \in (\overline{B}(0, C_{h^{--},h^{++}}) \cap F_0)^{(k)}} (R_h(\mathbf{c}) - R_{h_1}(\mathbf{c})) \leq \frac{\alpha}{4}$$

Then, according to step $k - 1$, since $\|(\mathbf{c}_{k-1,h^-}^*)_j\| \leq C_{h^{--},h^{++}}$ for $j = 1, \ldots, k$, we may write

$$R_{k-1,h_1}^* = R_{h_1}(\mathbf{c}_{k-1,h_1}^*) \geq R_{k-1,h}^* - \frac{\alpha}{4}.$$

As a consequence, since $R_{k-1,h}^* - R_{k,h}^* \geq \alpha$, we have $R_{k-1,h_1}^* - R_{k,h^+}^* \geq \frac{\alpha}{2}$. Now, let $\mathbf{c}$ be a $\frac{\alpha}{4}$-minimizer of $R_{k,s}^*$, for $s \geq h^- = \frac{h+h_1}{2}$, and assume that $P(V_{1,s}) < h - h_b^-$. Then

$$R_{k,h^+}^* \geq R_s(\mathbf{c}) - \frac{\alpha}{4}$$
$$\geq P \sum_{j=2}^{k} d_\phi(u, c_j) \mathbb{1}_{V_{j,s}}(u) - \frac{\alpha}{4}$$
$$\geq R_{k-1,h^-}^* - \frac{\alpha}{4}$$
$$\geq R_{k-1,h_1}^* - \frac{\alpha}{4},$$

this is a contradiction. Thus, for such a choice of h^- and $h^- \leq s \leq h^+$, $P(V_{p,s}) \geq h - h^-$, which entails $\|(T_s(\mathbf{c}))_p\| \leq P\|u\|/(h - h^-)$, for every $p \in [\![1,k]\!]$. □

7 Proofs for Section 4

7.1 Intermediate results

Theorem 10 and 11 require some additional probabilistic results that are gathered in this subsection. We begin with standard deviation bounds.

Lemma 17. *Let $\mathcal{C}$ denote the class of Bregman balls $B_\phi(x, r) = \{y \in \mathbb{R}^d | \sqrt{d_\phi(y, x)} < r\}$, $x \in \mathbb{R}^d$, $r \geq 0$. Then*

$$d_{VC}(\mathcal{C}) < d + 1,$$

where d_{VC} denotes the Vapnik-Chervonenkis dimension.

Proof of Lemma 17. Let $S = \{x_1, \ldots, x_{d+2}\}$ be shattered by $\mathcal{C}$. And let A_1, A_2 be a partition of S. Then we may write

$$A_1 = S \cap B_\phi(c_1, r_1) \cap B_\phi(c_2, r_2)^c$$
$$A_2 = S \cap B_\phi(c_2, r_2) \cap B_\phi(c_1, r_1)^c,$$

for $c_1, c_2 \in \mathbb{R}^d$ and $r_1, r_2 \geq 0$. Straightforward computation shows that, for any $x \in A_1$,

$$\ell_{1,2}(x) < 0,$$

where $\ell_{1,2}(x) = \phi(c_2) - \phi(c_1) + \langle x, \nabla_{c_2}\phi - \nabla_{c_1}\phi \rangle + \langle \nabla_{c_1}\phi, c_1 \rangle - \langle \nabla_{c_2}\phi, c_2 \rangle - r_1^2 + r_2^2$. Similarly we have that, for any $x \in A_2$ $\ell_{1,2}(x) > 0$. Thus S is shattered by affine hyperplanes (whose VC-dimension is $d + 1$), hence the contradiction. □

Proposition 18. *With probability larger than $1 - e^{-x}$, we have, for $k \geq 2$,*

$$\sup_{\mathbf{c} \in (\mathbb{R}^d)^k, r \geq 0} \left| (P - P_n) \mathbb{1}_{B_\phi(\mathbf{c}, r)} \right| \leq C \sqrt{\frac{(d+1)\log(k)}{n}} + \sqrt{\frac{2x}{n}}, \tag{3}$$

where C denotes a universal constant. Moreover, if r^+ and K are fixed, and $P\|x\|^2 \leq M_2^2 < \infty$, we have, with probability larger than $1 - e^{-x}$,

$$\sup_{\mathbf{c} \in B(0,K)^k \cap F_0, r \leq r^+} \left| (P - P_n) d_\phi(., \mathbf{c}) \mathbb{1}_{B_\phi(\mathbf{c}, r)} \right| \leq C_{K, r^+, M_2} r^+ \frac{\sqrt{kd}}{\sqrt{n}} + r^+ \sqrt{\frac{2x}{n}}, \tag{4}$$

where we recall that $\overline{conv(supp(P))} = F_0 \subset \overset{\circ}{\Omega}$.

Proof of Proposition 18. Let $\mathbf{c} \in \mathbb{R}^k$, since $B_\phi(\mathbf{c}, r) = \bigcup_{j=1}^k B_\phi(c_j, r)$, according to [van der Vaart and Wellner, 2009, Theorem 1.1], we may write

$$d_{VC}\left(\left\{ B_\phi(\mathbf{c}, r) | \mathbf{c} \in (\mathbb{R}^d)^k, r \geq 0 \right\} \right) \leq c_1(d+1)\log(c_2 k),$$

where c_1 and c_2 are universal constant. Thus, applying [Boucheron et al., 2005, Theorem 3.4] gives (3).

The second inequality (4) is more involved. Denote by Z its left-hand side. A bounded difference inequality (see, e.g., [Boucheron et al., 2013, Theorem 6.2]) yields

$$\mathbb{P}\left(Z \geq \mathbb{E}Z + r^+ \sqrt{\frac{2x}{n}} \right) \leq e^{-x}.$$

It remains to bound

$$\mathbb{E}Z \leq 2\mathbb{E}_X \mathbb{E}_\sigma \frac{1}{n} \sup_{\mathbf{c} \in B(0,K)^k \cap F_0, r \leq r^+} \sum_{i=1}^n \sigma_i d_\phi(X_i, \mathbf{c}) \mathbb{1}_{B_\phi(\mathbf{c}, r)}(X_i),$$

according to the symmetrization principle (see, e.g., [Boucheron et al., 2013, Lemma 11.4]), where for short $\mathbb{E}_Y$ denotes expectation with respect to the random variable Y. Consider $X_1, \ldots, X_n$ as fixed, and denote, for any set of real-valued functions $\mathcal{F}$, by $\mathcal{N}(\mathcal{F}, \varepsilon, L_2(P_n))$ its ε covering number with respect to the $L_2(P_n)$ norm. In order to apply Dudley's entropy integral, we will make use of the following result.

Theorem 19. *[Mendelson and Vershynin, 2003, Theorem 1] If $\mathcal{F}$ is a set of functions taking values in $[-1, 1]$. Then, for all S and $\varepsilon \leq 1$*

$$\mathcal{N}(\mathcal{F}, \varepsilon, L_2(S)) \leq \left(\frac{2}{\varepsilon} \right)^{\kappa d_{VC}(\mathcal{F})},$$

where κ denotes a universal constant and with a slight abuse of notation $d_{VC}(\mathcal{F})$ denotes the pseudo-dimension of $\mathcal{F}$.

Now let Γ_0 denote the set of functions $\left\{ \frac{d_\phi(., \mathbf{c})}{r^+} \mathbb{1}_{B_\phi(\mathbf{c}, r)} | \mathbf{c} \in B(0, K)^k, r \leq r^+ \right\}$. It is immediate that

$$\mathcal{N}(\Gamma_0, \varepsilon, L_2(P_n)) \leq \mathcal{N}(\Gamma_1, \varepsilon/2, L_2(P_n)) \times \mathcal{N}(\Gamma_2, \varepsilon/2, L_2(P_n)),$$

where $\Gamma_1 = \left\{ \frac{d_\phi(., \mathbf{c})}{r^+} \wedge 1 \right\}$ and $\Gamma_2 = \left\{ \mathbb{1}_{B_\phi(\mathbf{c}, r)} \right\}$. On one hand, we have

$$\mathcal{N}(\Gamma_2, u, L_2(P_n)) = \mathcal{N}(1 - \Gamma_2, u, L_2(P_n))$$

$$= \mathcal{N}\left(\left\{ \prod_{j=1}^k \mathbb{1}_{B_\phi(c_j, r)^c} \mid \mathbf{c}, r \right\}, u, L_2(P_n) \right)$$

$$\leq \mathcal{N}\left(\left\{ \mathbb{1}_{B_\phi(c, r)} | c \in \mathbb{R}^d, r \geq 0 \right\}, u/k, L_2(P_n) \right)^k$$

$$\leq \left(\frac{2k}{u} \right)^{\kappa(d+1)},$$

according to Theorem 19.

Now turn to Γ_1. According to Lemma 14, we may write

$$\mathcal{N}(\Gamma_1, u, L_2(P_n)) \leq \mathcal{N}\left(\left\{\frac{d_\phi(., \mathbf{c})}{r^+} | \mathbf{c} \in (F_0 \cap B(0, K))^k\right\}, u, L_2(P_n)\right)$$
$$\leq \mathcal{N}\left(B(0, K)^k, \frac{r^+ u}{C_K(1 + \|x\|_{L_2(P_n)})}, d_H\right).$$

Since $\mathcal{N}(B(0,1), u, \|\|) \leq \left(\frac{3}{u}\right)^d$, it follows that

$$\mathcal{N}(\Gamma_1, u, L_2(P_n)) \leq \left(\frac{3KC_K(1 + \|x\|_{L_2(P_n)})}{r^+ u}\right)^{kd},$$

hence

$$\mathcal{N}(\Gamma_0, \varepsilon, L_2(P_n)) \leq \left(\frac{6KC_K(1 + \|x\|_{L_2(P_n)})}{r^+ \varepsilon}\right)^{kd} \times \left(\frac{4k}{\varepsilon}\right)^{\kappa(d+1)}.$$

Using Dudley's entropy integral (see, e.g., [Boucheron et al., 2013, Corollary 13.2]) yields, for $k \geq 2$,

$$\mathbb{E}_\sigma \frac{1}{n} \sup_{\mathbf{c} \in B(0,K)^k \cap F_0, r \leq r^+} \sum_{i=1}^n \sigma_i \frac{d_\phi(X_i, \mathbf{c})}{r^+} \mathbb{1}_{B_\phi(\mathbf{c}, r)}(X_i)$$
$$\leq C \frac{r^+}{\sqrt{n}} \sqrt{kd \log(\frac{C_K(1 + \|x\|_{L_2(P_n)})}{r^+}) + \kappa(d+1)\log(4k)}.$$

Thus, applying Jensen's inequality leads to

$$\mathbb{E}_X \mathbb{E}_\sigma \mathbb{E}_\sigma \frac{1}{n} \sup_{\mathbf{c} \in B(0,K)^k \cap F_0, r \leq r^+} \sum_{i=1}^n \sigma_i \frac{d_\phi(X_i, \mathbf{c})}{r^+} \mathbb{1}_{B_\phi(\mathbf{c}, r)}(X_i) \leq r^+ C_{K, r^+, M_2} \sqrt{\frac{kd}{n}}$$

and to (4). $\qquad\square$

To prove Theorem 10, a more involved version of Markov's inequality is needed, stated below.

Lemma 20. *If $P\|u\|^p < \infty$ for some $p \geq 2$, then, there exists some positive constant C such that with probability larger than $1 - n^{-\frac{p}{2}}$,*
$$P_n\|u\| \leq C.$$

Proof. According to the Markov inequality, we may write

$$\mathbb{P}\left(P_n\|u\| - P\|u\| \geq \epsilon\right) \leq \frac{\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \|X_i\| - P\|u\|\right|^p\right]}{\epsilon^p}.$$

That leads to

$$\mathbb{P}\left(P_n\|u\| - P\|u\| \geq \epsilon\right) \leq \frac{\mathbb{E}\left[\left|\sum_{i=1}^n \|X_i\| - P\|u\|\right|^p\right]}{n^p \epsilon^p}.$$

From the Marcinkiewicz-Zygmund inequality applied to the real-valued centred random variables $Y_i = \|X_i\| -$

$P\|u\|$ and the Minkowski inequality, it follows that

$$\mathbb{E}\left[\left|\sum_{i=1}^{n}\|X_i\| - P\|u\|\right|^p\right] = \mathbb{E}\left[\left|\sum_{i=1}^{n}Y_i\right|^p\right]$$

$$\leq B_p\mathbb{E}\left[\left(\sum_{i=1}^{n}Y_i^2\right)^{\frac{p}{2}}\right]$$

$$\leq B_p\left(\sum_{i=1}^{n}(\mathbb{E}|Y_i|^p)^{\frac{2}{p}}\right)^{\frac{p}{2}}$$

$$= B_p n^{\frac{p}{2}}\mathbb{E}\left[|Y|^p\right]$$

$$= B_p n^{\frac{p}{2}}P\left(|\|u\| - P\|u\||^p\right),$$

for some positive constant B_p. Since $P\left(|\|u\| - P\|u\||^p\right) \leq P\|u\|^p + (P\|u\|)^p \leq 2P\|u\|^p$, according to Jensen inequality, the result derives from a suitable choice of ϵ. $\qquad\square$

A key intermediate result shows that, on the probability events defined above, empirical risk minimizers must have bounded codepoints.

Proposition 21. *Assume that $P\|u\|^p < +\infty$ for some $p \geq 2$. Let h^- and h^+ denote the quantities such that $\min_{j=2,\dots,k} R^*_{j-1,h^-} - R^*_{j,h^+} \geq \frac{\alpha}{2}$ with $h^- < h < h^+$, as in Lemma 16. Denote by $\eta = \frac{h-h^-}{k-1}$. Then there exists C_P such that, for n large enough, with probability larger than $1 - n^{-\frac{p}{2}}$, we have, for all $j = 2,\dots,k$, and $i = 1,\dots,j$,*

$$\sup_{h-(k-j)\eta \leq s \leq h}\|\hat{c}_{j,s,i}\| \leq C_P,$$

where $\hat{\mathbf{c}}_{j,s}$ denotes a j-codepoints empirical risk minimizer with trimming level s.

Proof of Proposition 21. We let C_P denote a constant to be fixed later. Similarly to the proof of Theorem 7, we denote by $\hat{T}_s$ the operator that maps $\hat{\mathbf{c}}$ to the empirical means of its Bregman-Voronoi cells. Let h^- and h^+ be as in Lemma 16. Then, according to Proposition 18, for n large enough we have that,

$$\sup_{\mathbf{c}\in(\mathbb{R}^d)^{(k)}, r\geq 0}|(P-P_n)B_\phi(\mathbf{c},r)| \leq \min\left(\frac{\eta}{2}, h^+ - h\right). \tag{5}$$

with probability larger than $1 - \frac{1}{8n^{\frac{p}{2}}}$. On this probability event, from Lemma 13 and the fact that $s \mapsto r_s(\mathbf{c})$ is non-decreasing, we deduce that for some $r^+ > 0$,

$$\sup_{c\in B(0,C_P)\cap F_0, s\leq h^+} r_{n,s}(c) \vee r_s(c) \leq r^+.$$

Since $P\|u\|^p < +\infty$, Lemma 20 yields that $P_n\|u\| \leq C_1$, for C_1 large enough, with probability larger than $1 - \frac{1}{8n^{\frac{p}{2}}}$. Besides, choosing $x = \log(8n^{\frac{p}{2}})$ in Proposition 18, we also have, with probability larger than $1 - \frac{1}{8n^{\frac{p}{2}}}$,

$$\sup_{\mathbf{c}\in(B(0,C_P)\cap F_0)^k, r\leq r^+}\left|(P-P_n)d_\phi(u,\mathbf{c})\mathbb{1}_{B_\phi(\mathbf{c},r)}(u)\right| \leq \alpha_n, \tag{6}$$

where $\alpha_n = O(\sqrt{\log(n)/n})$. We then work on the global probability event on which all these deviation inequalities are satisfied, that have probability larger than $1 - \frac{1}{n^{\frac{p}{2}}}$, and proceed recursively on j.

For $j = 1$ and $s \geq h^-$, according to Proposition 8, $\hat{T}_s(\hat{\mathbf{c}}_{1,s}) = \hat{\mathbf{c}}_{1,s}$, hence

$$\|\hat{\mathbf{c}}_{1,s}\| \leq \frac{P_n\|u\|}{h^-} \leq C_P.$$

Now assume that the statement of Proposition 21 holds up to order $j-1$. We choose $C_P = \frac{C_1}{h^-} \vee C_{h^-,h^+} \vee C_P'$, where C_P' corresponds to the constant C_P for the step $j-1$ and C_{h^-,h^+} is given by Lemma 16. Let $\hat{\mathbf{c}}_{j,s}$ be a j-points empirically optimal codebook with trimming level $s \geq h - (k-j)\eta$, and assume that there exists one cell (say V_1) such that $P_n(V_1(\hat{\mathbf{c}}_{j,s}) \cap B_\phi(\hat{\mathbf{c}}_{j,s}, r_{n,s}(\hat{\mathbf{c}}_{j,s}))) \leq \frac{\eta}{2}$. On the one hand, we may write

$$\hat{R}_s(\hat{\mathbf{c}}_{j,s}) \leq \hat{R}_s(\mathbf{c}_{j,h^+}^*)$$
$$\leq P_n d_\phi(u, \mathbf{c}_{j,h^+}^*) \mathbb{1}_{B_\phi(\mathbf{c}_{j,h^+}^*, r_{h^+}(\mathbf{c}_{j,h^+}^*))}(u)$$
$$\leq R_{j,h^+}^* + \alpha_n,$$

where $\mathbf{c}_{j,h^+}^*$ is a R_{j,h^+}^* minimizer provided by Theorem 7.

On the other hand, letting $h' = h - \left(k - j + \frac{1}{2}\right)\eta \geq h^-$, we have

$$\hat{R}_s(\hat{\mathbf{c}}_{j,s}) \geq \sum_{p=2}^{j} P_n d_\phi(u, \hat{c}_{j,s,p}) \mathbb{1}_{V_p(\hat{\mathbf{c}}_{j,s}) \cap B_\phi(\hat{\mathbf{c}}_{j,s}, r_{n,s}(\hat{\mathbf{c}}_{j,s}))}(u)$$
$$\geq \hat{R}_{h'}(\hat{\mathbf{c}}_{j-1,h'})$$
$$\geq P d_\phi(u, \hat{\mathbf{c}}_{j-1,h'}) \mathbb{1}_{B_\phi(\hat{\mathbf{c}}_{j-1,h'}, r_{h'})}(u) - \alpha_n,$$

according to the recursion assumption and (6). Thus,

$$\hat{R}_s(\hat{\mathbf{c}}_{j,s}) \geq R_{j-1,h'}^* - \alpha_n,$$

which is impossible for $\alpha_n < \frac{\alpha}{4}$. Thus, for every $p \in [\![1, j]\!]$,

$$P_n(V_p(\hat{\mathbf{c}}_{j,s}) \cap B_\phi(\hat{\mathbf{c}}_{j,s}, r_{n,s}(\hat{\mathbf{c}}_{j,s}))) \geq \frac{\eta}{2}.$$

According to Proposition 8, equality $\hat{T}(\hat{\mathbf{c}}_{j,s}) = \hat{\mathbf{c}}_{j,s}$ holds and entails

$$\|\hat{c}_{j,s,p}\| \leq \frac{P_n\|u\|}{h - (k-j)\eta} \leq \frac{P_n\|u\|}{h^-} \leq C_P.$$

$\square$

7.2 Proof of Theorem 10

Before proving Theorem 10, first note that since ϕ is strictly convex and continuous, $\psi : x \mapsto \phi(x) - \langle x, a \rangle + b$ is also strictly convex and continuous, for every a and b, thus $\psi^{-1}(0)$ is a closed set. This proves that $\mathbb{1}_{\psi^{-1}(0)}$ is measurable. Moreover, since ψ is strictly convex, there are at most two points of $\psi^{-1}(0)$ in any line containing the point 0. Thus, the Lebesgue measure of $\psi^{-1}(0)$ is 0. Since P is absolutely continuous with respect to the Lebesgue measure, it follows that boundaries of Bregman balls have P-mass equal to 0.

The proof of Theorem 10 is based on the following lemma:

Lemma 22. *Let $(P_n)_{n \in \mathbb{N}}$ be a sequence of probabilities that converges weakly to a distribution P. Assume that $supp(P_n) \subset supp(P) \subset \mathbb{R}^d$, $F_0 = \overline{conv(supp(P))} \subset \mathring{\Omega}$ and ϕ is $\mathcal{C}_2$ on Ω. Then, for every $h \in (0,1)$ and $K > 0$, there exists $K_+ > 0$ such that for every $\mathbf{c} \in \Omega^{(k)}$ satisfying $|c_i| \leq K$ for some $i \in [\![1, k]\!]$ and every $n \in \mathbb{N}$,*

$$r_{n,h}(\mathbf{c}) \leq r_+ = \sqrt{4(2K + K_+) \sup_{c \in F_0 \cap \overline{B}(0, 2K + K_+)} \|\nabla_c \phi\|}.$$

Proof of Lemma 22. Set $c \in \overline{B}(0, K) \cap F_0$. Since P_n converges weakly to P, according to the Prokhorov theorem, $(P_n)_n$ is tight. Thus, there is $K_+ > 0$ such that $P_n(B(0, K_+)) > h$ for all $n \in \mathbb{N}$ and $P(B(0, K_+)) >$

h. It comes that $P_n(B(c, K + K_+)) > h$. Moreover, for every x, y in $F_0 \cap \overline{B}(0, 2K + K_+)$, the mean value theorem yields

$$d_\phi(x, y) \leq 2 \sup_{c \in F_0 \cap \overline{B}(0, 2K + K_+)} \|\nabla_c \phi\| \|x - y\|$$
$$\leq 4(2K + K_+)C_+ = (r^+)^2,$$

for $C_+ = \sup_{c \in F_0 \cap \overline{B}(0, 2K + K_+)} \|\nabla_c \phi\|$ that is finite since $F_0 \cap \overline{B}(0, 2K + K_+)$ is compact. Thus, it follows that

$$B(c, K + K_+) \subset B_\phi(c, r_+). \tag{7}$$

As a consequence, $P_n(B_\phi(c, r_+)) > h$ and $P_n(B_\phi(\mathbf{c}, r_+)) > h$ if $c \in \mathbf{c}$ and $r_{n,\phi,h}(\mathbf{c}) \leq r_+$. $\qquad \square$

The proof of Theorem 10 is an adaptation of the proof of Theorem 3.4 in Cuesta-Albertos et al. [1997].

Let $(X'_n)_{n \in \mathbb{N}}$ be a sequence of independent random variables from P. We can define $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X'_i}$ the empirical distribution associated with the n first realisations. Note that P_n is random, and P_n converges weakly to P a.s..

According to Proposition 21, provided that $P\|u\|^p < +\infty$, for some $p > 2$, there exists $C_P > 0$ such that for some $N \in \mathbb{N}$, $\sum_{n \geq N} P(\max_{i \in [\![1,k]\!]} |\hat{c}_{n,i}| > C_P) < \infty$. Thus, according to the Borel-Cantelli Lemma, a.s. for n large enough, for every $i \in [\![1, k]\!]$, $|\hat{c}_{n,i}| \leq C_P$.

Now $C_P > 0$ and $N \in \mathbb{N}$ such that $\forall n \geq N$, $\forall i \in [\![1, k]\!]$, $|\hat{c}_{n,i}| \leq C_P$. As aforementioned, this occurs with probability 1.

According to the Skorokhod's representation theorem in the Polish space $\mathbb{R}^d$, it is possible to construct a measured space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ and a sequence of random variables $(X_n)_{n \in \mathbb{N}}$ and a random variable X on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ such that $X_n \sim P_n$, $X \sim P$ and X_n converges to X $\tilde{P}$-a.s.

Denote by $\mathbf{c}^*$ the unique minimiser of $\mathbf{c} \mapsto \mathrm{V}^P_{\phi,h}(\mathbf{c})$, $r'_n = r_{n,h}(\mathbf{c}^*)$ and τ'_n a $[0,1]$-valued measurable function such that $hP_{n,\mathbf{c}^*,h} = P_n\tau'_n$, that is, such that $P_n\tau'_n(u) = h$ and

$$\mathbb{1}_{B_\phi(\mathbf{c}^*, r'_n)} \leq \tau'_n \leq \mathbb{1}_{\overline{B}_\phi(\mathbf{c}^*, r'_n)}.$$

According to Lemma 22 with $K = |c_1^*|$ for instance, it comes $r'_n \leq r^+$, for some finite r^+. Thus, up to a subsequence, we may assume that $r'_n \to r'_0$ for some $r'_0 \leq r_+$.

Moreover,

$$|d_\phi(X_n, \mathbf{c}^*) - d_\phi(X, \mathbf{c}^*)| \leq |\phi(X_n) - \phi(X)| + \max_{j \in [\![1,k]\!]} \|\nabla_{c_j^*} \phi\| |X_n - X| \to 0, \tag{8}$$

when $n \to \infty$. As a consequence, $\tau'_n(X_n) \to \mathbb{1}_{B_\phi(\mathbf{c}^*, r'_0)}(X)$ $\tilde{P}$-a.e. The dominate convergence theorem yields $h = P_n\tau'_n(u) \to P(B_\phi(\mathbf{c}^*, r'_0))$. Thus, $\mathbb{1}_{B_\phi(\mathbf{c}^*, r'_0)} = \tau_0$ P-a.e where τ_0 denotes the trimming set associated with $\mathbf{c}^*$ and P. Moreover, since $\tau'_n(X_n)d_\phi(X_n, \mathbf{c}^*)$ is bounded by r_+ and converges to $\tau_0(X)d_\phi(X, \mathbf{c}^*)$ a.e., the dominated convergence theorem entails

$$R_{n,h}(\hat{\mathbf{c}}_n) \leq R_{n,h}(\mathbf{c}^*) \leq \mathbb{E}\left[\tau'_n(X_n)d_\phi(X_n, \mathbf{c}^*)\right] \to \mathbb{E}\left[\tau_0(X)d_\phi(X, \mathbf{c}^*)\right].$$

Thus, up to a subsequence,

$$\limsup_{n \to \infty} R_{n,h}(\hat{\mathbf{c}}_n) \leq R^*_{k,h}.$$

Since, for $n \geq N$ and every $i \in [\![1, k]\!]$, $|\hat{c}_{n,i}| \leq C_P$, we have $\hat{c}_{n,i} \to c_i$ for some $c_i \in F_0 \cap \overline{B}(0, C_P)$, up to a subsequence. Set $\mathbf{c} = (c_1, c_2, \ldots, c_k)$. Again, according to Lemma 22 with $K = C_P$, it comes that up to a subsequence, $r_{n,h}(\hat{\mathbf{c}}_n) \to r$ for some $r \geq 0$. As a consequence, from (8), Lemma 14 and the continuity of P,

$$\lim_{n \to \infty} \tau_n(X_n) = \mathbb{1}_{B_\phi(\mathbf{c}, r)}(X) a.e.$$

According to the dominated convergence theorem, we have $h = P(B_\phi(\mathbf{c}, r)) = P_n(\tau_n(u))$.

Again, the dominated convergence theorem entails that up to a subsequence,

$$\liminf_{n \to \infty} R_{n,h}(\hat{\mathbf{c}}_n) \geq \int \mathbb{1}_{B_\phi(\mathbf{c},r)}(u)d_\phi(u,\mathbf{c}) \geq R_{k,h}^*.$$

As a consequence, $\lim_{n\to\infty} R_{n,h}(\hat{\mathbf{c}}_n) = R_{k,h}^*$ and $\mathbf{c} = \mathbf{c}^*$. Since every subsequence of $(\hat{\mathbf{c}}_n)_{n \in \mathbb{N}}$ has a converging subsequence to $\mathbf{c}^*$, the sequence $(\hat{\mathbf{c}}_n)_{n \in \mathbb{N}}$ converges to $\mathbf{c}^*$.

7.3   Proof of Theorem 11

We assume that all the probability events described in the proof of Proposition 21 hold. This occurs with probability larger than $1 - n^{-\frac{p}{2}}$. On this probability event, recall that for all j $\|\hat{\mathbf{c}}_{n,j}\| \leq C_P$, $\sup_{\mathbf{c} \in (F_0 \cap \overline{B}(0,C_P))^{(k)}} r_{n,h}(\mathbf{c}) \vee r_h(\mathbf{c}) \leq r^+$. Next we further assume that the deviation bounds of Proposition 18 hold, with parameter C_P and r_+, to define a global probability event with mass large than $1 - n^{-\frac{p}{2}} - 2e^{-x}$. On this event, we have

$$
\begin{aligned}
R_h(\hat{\mathbf{c}}_n) - R_{k,h}^* &= Pd_\phi(u,\hat{\mathbf{c}}_n)\mathbb{1}_{B_\phi(\hat{\mathbf{c}}_n,r_h(\hat{\mathbf{c}}_n))}(u) - R_{k,h}^* \\
&\leq Pd_\phi(u,\hat{\mathbf{c}}_n)\mathbb{1}_{B_\phi(\hat{\mathbf{c}}_n,r_{n,h}(\hat{\mathbf{c}}_n))}(u) - Pd_\phi(u,\mathbf{c}^*)\mathbb{1}_{B_\phi(\mathbf{c}^*,r_{n,h}(\mathbf{c}^*))}(u) \\
&\quad + Pd_\phi(u,\hat{\mathbf{c}}_n)(\mathbb{1}_{B_\phi(\hat{\mathbf{c}}_n,r_h(\hat{\mathbf{c}}_n))}(u) - \mathbb{1}_{B_\phi(\hat{\mathbf{c}}_n,r_{n,h}(\hat{\mathbf{c}}_n))}(u)) \\
&\quad + Pd_\phi(u,\mathbf{c}^*)(\mathbb{1}_{B_\phi(\mathbf{c}^*,r_{n,h}(\mathbf{c}^*))}(u) - \mathbb{1}_{B_\phi(\mathbf{c}^*,r_h(\mathbf{c}^*))}(u)) \\
&\leq 2 \sup_{\mathbf{c} \in (F_0 \cap \overline{B}(0,C_P))^{(k)}, r \leq r^+} |(P - P_n)d_\phi(u,\mathbf{c})\mathbb{1}_{B_\phi(\mathbf{c},r)}(u)| \\
&\quad + 2r^+ \sup_{\mathbf{c} \in \Omega^{(k)}, r \geq 0} |(P - P_n)B_\phi(\mathbf{c},r)| \\
&\leq \frac{C_P}{\sqrt{n}}(1 + \sqrt{x}),
\end{aligned}
$$

for some constant C_P.

References

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

A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51, 2005a.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005b.

G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory*, 54(2):781–790, 2008. ISSN 0018-9448. doi: 10.1109/TIT.2007.913516. URL https://doi.org/10.1109/TIT.2007.913516.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005. ISSN 1292-8100. doi: 10.1051/ps:2005018. URL http://dx.doi.org/10.1051/ps:2005018.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL http://dx.doi.org/10.1093/acprof:oso/9780199535255.001.0001. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *Ann. Statist.*, 43(6):2507–2536, 2015. ISSN 0090-5364. doi: 10.1214/15-AOS1350. URL https://doi.org/10.1214/15-AOS1350.

O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector. *ArXiv e-prints*, Feb. 2018.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.

F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics archive*, 11:733–751, 2011.

J. A. Cuesta-Albertos, A. Gordaliza, and C. Matrán. Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.*, 25(2):553–576, 04 1997. doi: 10.1214/aos/1031833664. URL http://dx.doi.org/10.1214/aos/1031833664.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, New York, 2000.

A. Fischer. Quantization and clustering with Bregman divergences. *J. Multivariate Anal.*, 101(9):2207–2221, 2010. ISSN 0047-259X. doi: 10.1016/j.jmva.2010.05.008. URL https://doi.org/10.1016/j.jmva.2010.05.008.

L. A. Garcí a Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A general trimming approach to robust cluster analysis. *Ann. Statist.*, 36(3):1324–1345, 2008. ISSN 0090-5364. doi: 10.1214/07-AOS515. URL https://doi.org/10.1214/07-AOS515.

A. Gordaliza. Best approximations to random variables based on trimming procedures. *J. Approx. Theory*, 64(2):162–180, 1991. ISSN 0021-9045. doi: 10.1016/0021-9045(91)90072-I. URL https://doi.org/10.1016/0021-9045(91)90072-I.

R. M. Gray, A. Buzo, A. H. Gray, and Y. Matsuyama. Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:367–376, 1980.

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

G. Lecué and M. Lerasle. Robust machine learning by median-of-means : theory and practice. *ArXiv e-prints*, Nov. 2017.

C. Levrard. Nonasymptotic bounds for vector quantization in hilbert spaces. *Ann. Statist.*, 43(2):592–619, 04 2015. doi: 10.1214/14-AOS1293. URL https://doi.org/10.1214/14-AOS1293.

C. Levrard. Quantization/clustering: when and why does k-means work? *JSFdS*, 159(1):1–26, 2018.

T. Linder. Learning-theoretic methods in vector quantization. In *Principles of nonparametric learning (Udine, 2001)*, volume 434 of *CISM Courses and Lect.*, pages 163–210. Springer, Vienna, 2002.

S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Invent. Math.*, 152(1):37–55, 2003. ISSN 0020-9910. doi: 10.1007/s00222-002-0266-3. URL `http://dx.doi.org/10.1007/s00222-002-0266-3`.

F. Nielsen, J. Boissonnat, and R. Nock. Bregman Voronoi diagrams: properties, algorithms and applications. Technical Report 6154, INRIA, 2007.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.

A. Strehl and J. Ghosh. Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

C. Tang and C. Monteleoni. On lloyd's algorithm: New theoretical insights for clustering in practice. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1280–1289, Cadiz, Spain, 09–11 May 2016. PMLR.

A. van der Vaart and J. A. Wellner. A note on bounds for VC dimensions. In *High dimensional probability V: the Luminy volume*, volume 5 of *Inst. Math. Stat. (IMS) Collect.*, pages 103–107. Inst. Math. Statist., Beachwood, OH, 2009. doi: 10.1214/09-IMSCOLL508. URL `https://doi.org/10.1214/09-IMSCOLL508`.