Robust Bregman Clustering

Claire Brécheteau, Aurélie Fischer, Clément Levrard

Robust Bregman Clustering

Brécheteau, Claire
claire.brecheteau@ec-nantes.fr
École Centrale de Nantes – LMJL

Fischer, Aurélie
aurelie.fischer@lpsm.paris
Université Paris Diderot – LPSM

Levrard, Clément
clement.levrard@lpsm.paris
Université Paris Diderot – LPSM

Abstract

Clustering with Bregman divergences encompasses a wide family of clustering procedures that are well-suited to mixtures of distributions from exponential families [3]. However these techniques are highly sensitive to noise. To address the issue of clustering data with possibly adversarial noise, we introduce a robustified version of Bregman clustering based on a trimming approach. We investigate its theoretical properties, showing for instance that our estimator converges at a sub-Gaussian rate $1/\sqrt{n}$, where n denotes the sample size, under mild tail assumptions. We also show that it is robust to a certain amount of noise, stated in terms of Breakdown Point. We also derive a Lloyd-type algorithm with a trimming parameter, along with a heuristic to select this parameter and the number of clusters from sample. Some numerical experiments assess the performance of our method on simulated and real datasets.

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 [18]. There are no labels known in advance, so clustering is an unsupervised learning task. To perform clustering, a distance-like function $d(\cdot, \cdot)$ is often needed to assess a notion of proximity between points and the separation between clusters.

Suppose that we know such a natural distance d, and assume that the points we want to cluster, say $X_1, \ldots, X_n$, are i.i.d., drawn from an unknown distribution P, and take values in $\mathbb{R}^d$. For $k \geq 1$, designing k meaningful classes with respect to d can be achieved via minimizing the so-called empirical distortion

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

over all possible cluster centers or codebooks $\mathbf{c} = (c_1, \ldots, c_k)$, with notation $[\![1, k]\!]$ for $\{1, 2, \ldots, k\}$. This results in a set of k codepoints. Clusters are then given by the sets of sample points that have the same closest codepoint.

A classical choice of d is the squared Euclidean distance, leading to the standard k-means clustering algorithm (see, e.g., [31]). However, some of the desirable properties of the Euclidean distance can be extended to a broader class of dissimilarity functions, namely Bregman divergences. These distance-like functions, denoted by d_ϕ in the sequel, are indexed by strictly convex functions ϕ. Introduced by [7], they are useful in a wide range of areas, among which statistical learning and data mining ([3], [11]), computational geometry [34], natural sciences, speech processing and information theory [24]. Squared Euclidean, Mahalanobis, Kullback-Leibler and L^2 distances are all particular cases of Bregman divergences.

A Bregman divergence is not necessarily 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 [7]. They also satisfy non-negativity and separation, convexity in the first argument and linearity (see [3], [34]). Moreover, Bregman divergences are closely related to exponential families [3]. In fact, they are a natural tool to measure proximity between observations arising from a mixture of such distributions. Consequently, clustering with Bregman divergences is particularly well-suited in this case.

Clustering with Bregman divergences allows to state the clustering problem within a contrast minimization framework. Namely, through minimizing $R_n(\mathbf{c}) = P_n d_\phi(u, \mathbf{c})$, where P_n denotes the empirical distribution associated with $\{X_1, \dots, X_n\}$ and $Qf(u)$ means integration of f with respect to the measure Q, we intend to find a codebook $\hat{\mathbf{c}}_n$ whose "real" distortion $Pd_\phi(u, \hat{\mathbf{c}}_n)$ is close to the optimal k-points distortion $R_k^* := \inf_{\mathbf{c}} Pd_\phi(u, \mathbf{c})$. The convergence properties of empirical distortion minimizers are now quite well understood when the source distribution P is assumed to have a finite support [30, 21], even in infinite-dimensional cases [4, 28]. 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 viewpoint, this corruption issue might be tackled by winsorizing or trimming classical estimators, or by introducing some new and robust estimators that adapt to heavy-tailed cases. Such estimators can be based on PAC-Bayesian or Median of Means techniques [10, 8, 27] for instance. In a nutshell, these estimators succeed in achieving sub-Gaussian deviation bounds under mild tail conditions such as bounded variances and expectations, and they are also provably robust to a certain amount of noise [27].

In the clustering framework, it is straightforward that the k-means procedure suffers from the same drawback as the empirical mean: only one adversarial datapoint is needed to drive both the empirically optimal codebook $\hat{\mathbf{c}}_n$ and its distortion arbitrarily far from the optimal. In fact we show that it is the case with every possible Bregman divergence. Up to our knowledge, the only theoretically grounded attempt to robustify clustering procedures is to be found in [16], where a trimmed k-means heuristic is introduced. See also [19] for trimmed clustering with Mahalanobis distances. In some sense, this paper extends this trimming approach to the general framework of clustering with Bregman divergences.

We introduce some notation, background and fundamental properties for trimmed clustering with Bregman divergences in Section 2. This will lead to the description of our robust clustering technique, based on the computation of a trimmed empirically optimal codebook $\hat{\mathbf{c}}_{n,h}$, for a fixed trim level h.

Theoretical properties of our trimmed empirical codebook $\hat{\mathbf{c}}_{n,h}$ are exposed in Section 3. To be more precise, we investigate convergence towards a trimmed optimal codebook $\mathbf{c}_h^*$ in terms of distance and distortion, showing for instance that the excess distortion achieves a sub-Gaussian convergence rate of $O(1/\sqrt{n})$ in terms of sample size, under a mild bounded variance assumption. This shows that our procedure can be thought of as robust whenever noisy situations are modeled as a signal corrupted with heavy-tailed additive noise. We also assess robustness of $\hat{\mathbf{c}}_{n,h}$ in terms of Finite-sample Breakdown Point (see, e.g., [32]), showing that our procedure can theoretically endure a positive proportion of adversarial noise. A precise bound on this proportion is given, that illustrates the possible confusion between too small clusters and noise.

Then, a modified Lloyd's type algorithm is proposed in Section 4, along with a heuristic to select both the trim level h and the number k of clusters from data. The numerical performances of our algorithm are then investigated. We compare our method to trimmed k-means [16], tclust [22], ToMATo [14], dbscan [25] and a trimmed version of k-median [9]. Our algorithm with the appropriate Bregman divergence outperforms other methods on samples generated from Gaussian, Poisson, Binomial, Gamma and Cauchy mixtures. We then investigate the performances of our method on real datasets. First, we consider daily rainfall measurements for January and September in Cayenne, from 2007 to 2017, and try to cluster data according to the month. As suggested by [15], our method with the divergence associated with Gamma distributions turns out to be the most accurate one. Second, we intend to cluster chunks of 5000 words picked from novels corresponding to 4 different authors, based on stylometric descriptors [38, Section 10], corrupted by noise. Following [20], we show that our method with Poisson divergence is particularly well adapted for this framework.

At last, proofs are gathered in Sections 5 and 6. Proofs of technical intermediate results are deferred to the Appendix, along with some additional figures and results for Section 4.

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*

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

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 [35, Theorem 25.1]). Note that by taking $\phi : x \mapsto \|x\|^2$, with $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$, one gets $d_\phi(x, y) = \|x - y\|^2$. Let us present a few other 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 $d_\phi(x, y) = \sum_{\ell=1}^d x_\ell \ln \frac{x_\ell}{y_\ell}$.

For any compact set $K \subset \Omega$, and $x \in \Omega$, we also define

$$d_\phi(K, x) = \min_{y \in K} d_\phi(y, x) \quad \text{and} \quad d_\phi(x, K) = \min_{y \in K} d_\phi(x, y).$$

For every codebook $\mathbf{c} = (c_1, c_2, \ldots, c_k) \in \Omega^{(k)}$, $d_\phi(x, \mathbf{c})$ is defined by $d_\phi(x, \mathbf{c}) = \min_{i \in [\![1,k]\!]} d_\phi(x, c_i)$. The main property of Bregman divergences is that means are always minimizers of Bregman inertias, as exposed below. For a distribution Q and a function f, we denote by $Qf(u)$ the integration of f with respect to Q.

Proposition 2. *[2, Theorem 1] Let P be a probability distribution, and let ϕ be a strictly convex $\mathcal{C}_1$ real-valued function defined on a convex set $\Omega \subset \mathbb{R}^d$. Then, for any $x \in \Omega$,*

$$Pd_\phi(u, x) = Pd_\phi(u, Pu) + d_\phi(Pu, x).$$

As mentioned in [3], this property allows to design iterative Bregman clustering algorithms that are similar to Lloyd's algorithm. 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

$$R(\mathbf{c}) = Pd_\phi(u, \mathbf{c}).$$

When only an i.i.d. sample $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ is available, we denote by $R_n(\mathbf{c})$ the corresponding empirical distortion (associated with P_n). When P is a mixture of distributions belonging to an exponential family, there exists a natural choice of Bregman divergence, as detailed in Section 4. Standard Bregman clustering intends to infer a minimizer of R via minimizing R_n, and works well in the bounded support case [21].

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 [16, 23]: for a trim level $h \in (0, 1]$, both a codebook and a subset of P-mass not smaller 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 $\mathcal{P}_{+h} = \cup_{s \geq h} \mathcal{P}_s$. By analogy with [16], optimal trimming sets and codebooks are designed to achieve the optimal h-trimmed k-variation,

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

where $R(\tilde{P}, \mathbf{c}) = \tilde{P} d_\phi(u, \mathbf{c})$. In other words, $V_{k,h}$ is the best possible k-point distortion based on a normalized sub-measure of P. Intuitively speaking, the h-trimmed k-variation may be thought of as the k-points optimal distortion of the best "denoised" version of P, with denoising level $1-h$. For instance, in a mixture setting, if $P = \gamma P_0 + (1-\gamma)N$, where P_0 is a signal supported by k points and N is a noise distribution, then, provided that $h \leq \gamma$, $V_{k,h} = 0$.

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

$$P(B_\phi(\mathbf{c}, r)) \leq h \leq P(\bar{B}_\phi(\mathbf{c}, r)). \tag{1}$$

We denote this radius by $r_{n,h}(\mathbf{c})$ when the distribution is P_n. Note that $r_{n,h}(\mathbf{c})^2$ is the Bregman divergence to the $\lceil nh \rceil$ d_ϕ-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 coincide with $\frac{P}{h}$ on $B_\phi(\mathbf{c}, r_h(\mathbf{c}))$, with support included in $\bar{B}_\phi(\mathbf{c}, r_h(\mathbf{c}))$, a straightforward result is the following.

Lemma 3. *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 [16, Lemma 2.1], [23] or [13]. A short proof is given in the Appendix, Section 7.1. 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 4. *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 choice of $\tilde{P}_\mathbf{c}$ 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 5. *Let $0 < h < h' < 1$ and $\mathbf{c} \in \Omega^{(k)}$. Then*

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

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

As well, this lemma generalizes previous results in [16, 23]. A proof can be found in Section 5.2. Lemma 7.1 and Lemma 5 ensure that for a given $\mathbf{c}$, optimal $\tilde{P}$ in $\mathcal{P}_{+h}$ for $R(\tilde{P}, \mathbf{c})$ can be found in $\mathcal{P}_h(\mathbf{c})$. Thus, the optimal h-trimmed k-variation may be achieved via minimizing the h-trimmed distortion.

Proposition 6. *For every positive integer k and $0 < h < 1$,*

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

This proposition is an extension of [16, Proposition 2.3]. In other words, Proposition 6 assesses the equivalence between minimization of our robustified distortion R_h, and the original robust clustering criterion depicted in [16] (extended to Bregman divergences). Thus, a good codebook in terms of trimmed k-variation can be found by minimizing R_h.

Definition 7. *An h-trimmed k-optimal codebook is any element $\mathbf{c}^*$ in $\arg\min_{\mathbf{c} \in \Omega^{(k)}} R_h(\mathbf{c})$.*

Under mild assumptions on P and ϕ, trimmed k-optimal codebooks exist.

Theorem 8. *Let $0 < h < 1$, assume that $P\|u\| < +\infty$, ϕ is $\mathcal{C}^2$ and strictly convex 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. Then, the set $\arg\min_{\mathbf{c}\in\Omega^{(k)}} R_h(\mathbf{c})$ is not empty.*

A proof of Theorem 8 is given in Section 5.3. Note that Theorem 8 only requires $P\|u\| < +\infty$. This can be compared with the standard squared Euclidean distance case, where $P\|u\|^2 < +\infty$ is required for $R : \mathbf{c} \mapsto P\|u - \mathbf{c}\|^2$ to have minimizers. From now on we denote by $\mathbf{c}_h^*$ a minimizer of R_h, and by $\hat{\mathbf{c}}_{n,h}$ 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 [34]. 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). Proposition 9 below extends the so-called centroid condition in the Euclidean case to our Bregman setting.

Proposition 9. *Let $\mathbf{c} \in \Omega^{(k)}$ and $\tilde{P}_\mathbf{c} \in \mathcal{P}_h(\mathbf{c})$. Assume that for all $i \in [\![1, k]\!]$, $\tilde{P}_\mathbf{c}(W_i(\mathbf{c})) > 0$, and denote by $\mathbf{m}$ the codebook of the local means of $\tilde{P}_\mathbf{c}$. In other words, $m_i = \tilde{P}_\mathbf{c}(u\mathbb{1}_{W_i(\mathbf{c})}(u))/\tilde{P}_\mathbf{c}(W_i(\mathbf{c}))$. 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 9 is a straightforward consequence of Proposition 2, that emphasizes the key property that Bregman divergences are minimized by expectations (this is not the case for the L_1 distance for instance). In addition, it can be proved that Bregman divergences are the only loss functions satisfying this property [2]. In line with [3] for the non-trimmed case, Proposition 9 provides an iterative scheme to minimize R_h, that is detailed in Section 4.

3 Theoretical results

3.1 Convergence of a trimmed empirical distortion minimizer

This section is devoted to investigate the convergence of a minimizer $\hat{\mathbf{c}}_{n,h}$ of the empirical trimmed distortion $R_{n,h}$. Throughout this section ϕ is assumed to be $\mathcal{C}^2$, and $F_0 = \overline{conv(supp(P))} \subset \mathring{\Omega}$. We begin with a generalization of [16, Theorem 3.4], assessing the almost sure convergence of optimal empirical trimmed codebooks.

Theorem 10. *Assume that P is absolutely continuous with respect to the Lebesgue measure and satisfies $P\|u\|^p < \infty$ for some $p > 2$, then there exists $\mathbf{c}_h^*$ an optimal codebook such that*

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

Moreover, up to extracting a subsequence, we have

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

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 $[\![1, k]\!]$. At last, if $\mathbf{c}_h^*$ is unique, then $\lim_{n\to+\infty} D(\hat{\mathbf{c}}_{n,h}, \mathbf{c}_h^*) = 0$ a.e. (without taking a subsequence).

Note that, contrary to [16, Theorem 3.4], uniqueness of trimmed optimal codebooks is not required in Theorem 10. A proof is given in Section 6.2. Interestingly, slightly milder conditions are required for the trimmed distortion of $\hat{\mathbf{c}}_{n,h}$ to converge towards the optimal at a parametric rate.

Theorem 11. *Assume that $P\|u\|^p < \infty$, where $p \geq 2$. Further, if $R^*_{k,h}$ denotes the h-trimmed optimal distortion with k points, assume that $R^*_{k-1,h} - R^*_{k,h} > 0$. Then, for n large enough, with probability larger than $1 - n^{-\frac{p}{2}} - 2e^{-x}$, we have*

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

The requirement $R^*_{k-1,h} - R^*_{k,h} > 0$ ensures that optimal codebooks will not have empty cells. Note that if $R^*_{k-1,h} - R^*_{k,h} = 0$, then there exists a subset A of $\mathbb{R}^d$ satisfying $P(A) \geq h$ and such that the restriction of P to A is supported by at most $k-1$ points, that allows optimal k-points codebooks with at least one empty cell. It is worth mentioning that Theorem 11 does not require a unique trimmed optimal codebook, and only requires an order 2 moment condition for $\hat{\mathbf{c}}_{n,h}$ to achieve a sub-Gaussian rate in terms of trimmed distortion. This condition is in line with the order 2 moment condition required in [8] for a robustified estimator of $\mathbf{c}^*$ to achieve similar guarantees, as well as the finite-variance condition required in [10] in a mean estimation framework. A proof of Theorem 11 is given in Section 6.3. To derive results in expectation, a technical additional condition is needed.

Corollary 12. *Assume that there exists a non-decreasing convex function ψ such that*

$$\sup_{c \in B(0,t) \cap F_0} \|\nabla_c \phi\| \leq \psi(t).$$

Assume that $P\|u\|^p < \infty$, with $p \geq 2$, and let $q = p/(p-1)$ be the harmonic conjugate of p. If $P\|u\|^q \psi^q\left(\frac{k\|u\|}{h}\right) < \infty$, then

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

A proof of Corollary 12 is given in Section 6.4. Note that such a function ψ exists in most of the classical cases. The requirement $P\|u\|^q \psi^q(k\|u\|/h)$ roughly ensures that $P d^q_\phi(u, \hat{\mathbf{c}}_{n,h})$ remains bounded whenever the event described in Theorem 11 does not occur. 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 k-means case $\phi(x) = \|x\|^2$ and $\Omega = \mathbb{R}^d$, we can choose $\psi(t) = 2t$. The condition of Corollary 12 is satisfied for $P\|u\|^3 < +\infty$.
2. For $\phi(x) = \exp(x)$, $\Omega = \mathbb{R}$, we may choose $\psi(t) = \exp(t)$. The condition of Corollary 12 may be written for $p = 2$ as $Pu^2 \exp\left(\frac{2k|u|}{h}\right) < +\infty$.

3.2 Robustness properties of trimmed empirical distortion minimizers

This section is devoted to discuss to what extent the trimming procedure we propose implies robustness of our estimator to adversarial contamination. First we choose to assess robustness via the so-called Finite-sample Breakdown Point [17], that seizes what proportion of adversarial noise can be added to a dataset without making estimators getting arbitrarily large. To be more precise, for an amount s of adversarial points $\{x_1, \ldots, x_s\}$, we denote by P_{n+s} the empirical distribution associated with $\mathbb{X}_n \cup \{x_1, \ldots, x_s\}$ and by $\hat{\mathbf{c}}_{n+s,h}$ a minimizer of $\mathbf{c} \mapsto R_{n+s,h}(\mathbf{c})$. We may then define the Finite-sample Breakdown Point (FSBP) as follows:

$$\widehat{BP}_{n,h} := \inf\left\{ \frac{s}{n+s} \mid \sup_{\{x_1,\ldots,x_s\}} \|\hat{\mathbf{c}}_{n+s,h}\| = +\infty \right\}.$$

To give an intuition, the standard mean (minimizer of the 1-trimmed empirical distortion for $k = 1$, $\phi(u) = \|u\|^2$) has breakdown point $\frac{1}{n+1}$, whereas h-trimmed means have breakdown point roughly $1 - h$ (see, e.g., [32, Section 3.2.5]). According to [40, Theorem 1], this is also the case whenever ϕ is strictly convex (but still $k = 1$). In the case $k > 1$, as noticed in [16] for trimmed k-means, the breakdown point may be much smaller than $1 - h$. Note that if an h-trimmed optimal codebook has a too small cluster, then adding an adversarial cluster with greater weight might switch the roles between noise and signal, resulting in an h-trimmed codebook that allocates one point to the adversarial cluster and trims the too small optimal cluster. To quantify this intuition, we introduce the following discernability factor B_h.

Definition 13. *Let $h \in]0, 1[$, and, for $b \leq h$, denote by $h_b^- = (h - b)/(1 - b)$, $h_b^+ = h/(1 - b)$. The discernability factor B_h is defined as*

$$B_h = \sup \left\{ b \geq 0 \mid b \leq h \wedge (1 - h) \text{ and } \min_{j \in [\![2,k]\!]} R^*_{j-1,h_b^-} - R^*_{j,h_b^+} > 0 \right\}.$$

In fact, $h - h^-_{B_h} = (1 - h)B_h/(1 - B_h)$ is the portion of mass in an optimal k-points h trimming set that may be considered as noise by an optimal $k - 1$-points $h^-_{B_h}$ trimming set. As exposed in the following proposition, B_h is related to the minimum cluster weight of optimal h-trimmed codebooks.

Proposition 14. *Assume that the requirements of Theorem 8 are satisfied. If $R^*_{k-1,h} - R^*_{k,h} > 0$, then $B_h > 0$.*

Moreover, for any $j \in [\![1, k]\!]$, if $\mathbf{c}^{*,(j)}$ is a j-points h-trimmed optimal codebook and $p_{j,h} = h \min_{p \in [\![1,j]\!]} \tilde{P}_{\mathbf{c}^{*,(j)}} \left(W_p(\mathbf{c}^{*,(j)}) \right)$, with $\tilde{P}_{\mathbf{c}^{*,(j)}} \in \mathcal{P}_h(\mathbf{c}^{*,(j)})$, then

$$B_h(1 - (h - p_{j,h})) \leq p_{j,h}.$$

A proof of Proposition 14 is given in Section 5.4. Theorem 15 below makes connection between this discernability factor and robustness properties of optimal k-points h-trimmed codebook, stated in terms of Bregman radius.

Theorem 15. *For $\ell \geq 1$, let $R^*_{\ell,h}$ denote the ℓ-points h-trimmed optimal distortion. Assume that $P\|u\|^p < +\infty$, for some $p \geq 2$. Moreover, assume that $R^*_{k-1,h} - R^*_{k,h} > 0$. Let $b < B_h$, and assume that $s/(n + s) \leq b$. Then, for n large enough, with probability larger than $1 - n^{-\frac{p}{2}}$,*

$$\max_{j \in [\![1,k]\!]} d_\phi \left(B(0, C_{P,b}), \hat{c}_{n+s,h,j} \right) \leq K_{P,b},$$

where $C_{P,b}$ and $K_{P,b}$ do not depend on n nor s.

A proof of Theorem 15 is given in Section 6.5. Theorem 15 guarantees that the proposed trimming procedure is robust in terms of Bregman divergence, that is, the corrupted empirical distortion minimizer belongs to some closed Bregman ball, provided the proportion of noise is smaller than the discernability factor introduced in Definition 13. Unfortunately Bregman balls might not be compact sets if $c \mapsto d_\phi(x, c)$ is not a proper map. For instance, with $\phi(x) = e^x$ and $\Omega = \mathbb{R}$, we have $]-\infty, 0] \subset \{c \mid d_\phi(0, c) \leq 1\}$. In the proper map case, Theorem 15 entails that the FSBP is larger than B_h, with high probability, for n large enough. In the other case, Corollary 16 below ensures that this breakdown point is positive, provided that $p > 2$.

Corollary 16. *Assume that $P\|u\|^p < +\infty$, for $p > 2$. Under the assumptions of Theorem 15, there exists $c > 0$ such that, almost surely, for n large enough, $\widehat{BP}_{n,h} \geq c$.*

In addition, if, for every $x \in \Omega$, $c \mapsto d_\phi(x, c)$ is a proper map, then almost surely, for n large enough $\widehat{BP}_{n,h} \geq B_h$.

A proof of Corollary 16 can be found in Section 6.6. Corollary 16 guarantees that our trimmed Bregman clustering procedure is asymptotically robust in the usual sense to a certain proportion of adversarial noise, contrary to plain Bregman clustering whose FSBP is $1/(n + 1)$. However this unknown authorized proportion depends on both the choice of Bregman divergence and the discernability factor B_h. In the proper map case,

the FSBP is larger than B_h. Note that for $x \in \Omega$, $c \mapsto d_\phi(x, c)$ is proper whenever ϕ is strictly convex, that is the case for trimmed k-means [16]. For this particular Bregman divergence, the result of Corollary 16 is provably tight.

Example 17. *Let $\phi_1 = \|.\|^2$, $\phi_2 = \exp(-.)$, $\Omega = \mathbb{R}$, $P = (1-p)\delta_{-1} + p\delta_1$, with $p \leq 1/2$. Then, for $\phi = \phi_j$, $j \in \{1, 2\}$, $k = 2$ and $h > (1-p)$, we have $B_h = \frac{h+p-1}{p} \wedge (1-h)$. Let $Q_{\gamma,N} = (1-\gamma)P + \gamma\delta_N$. The following holds.*

- *If $(1+p)h > 1$, $B_h = 1-h$, and for every $\gamma > 1-h$, any sequence of optimal 2-points h-trimmed codebook $\mathbf{c}_2^*(Q_{\gamma,N})$ for $Q_{\gamma,N}$ satisfies*

$$\lim_{N \to +\infty} \|\mathbf{c}_2^*(Q_{\gamma,N})\| = +\infty.$$

- *If $(1+p)h \leq 1$, then $B_h = \frac{h+p-1}{p}$, and, for $\gamma = B_h$, $(-1, N)$ is an optimal 2-points h-trimmed codebook for $Q_{\gamma,N}$.*

The calculations pertaining to Example 17 may be found in the Appendix, Section 8.1. Note that upper bounds on the FSBP when $n \to +\infty$ may be derived for Example 17 using standard deviation bounds. Example 17 illustrates the two situations that can be encountered when some adversarial noise is added, depending on the balance between trim level and smallest optimal cluster. If the trim level is high enough compared to the smallest mass of an optimal cluster (first case), then the breakdown point is simply $1-h$, that is the amount of points that can be trimmed. This corresponds to the breakdown point of the trimmed mean (see, e.g., [40]). When the trim level becomes small compared to the smallest mass of an optimal cluster (second case), optimal codebooks for the perturbed distribution can be codebooks that allocate one point to the noise and trim the small optimal cluster, leading to a breakdown point possibly smaller than $1-h$. This corresponds to the situation exposed in Proposition 14. In both cases, the breakdown point is smaller than B_h, thus, according to Corollary 16, it is equal to B_h.

As mentioned in [16] for the trimmed k-means, in practice, breakdown point and choice of the correct number of clusters are closely related questions. This point is illustrated in Section 4.6, where the correct number of clusters depends on what is considered as noise. From a theoretical viewpoint, this question is tackled by Corollary 16 and Example 17, in the proper map case.

4 Numerical experiments

4.1 Description of the algorithm

The algorithm we introduce is inspired by the trimmed version of Lloyd's algorithm [16], and is also a generalization of the Bregman clustering algorithm [3, 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,... 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 $j \in [\![1, k]\!]$, $C_j^{(t)} \leftarrow W_j(\mathbf{c}^{(t-1)}) \cap NN_q^{(t)}$.*
  - *For $j \in [\![1, k]\!]$, $c_j^{(t)} \leftarrow \left(\sum_{x \in C_j^{(t)}} x\right) / \left|C_j^{(t)}\right|$.*

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

As for every EM-type algorithm, initialization may be crucial. This point will not be theoretically investigated in this paper. In practice, several random starts will be proceeded. More sophisticated strategies, such as k-`means ++` [1], could be an efficient way to address the initialization issue. An easy consequence of Proposition 9 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 18. *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, suitable clusterability assumptions as in [26, 37, 29] might lead to further guarantees on such an output.

4.2 Exponential Mixture Models

In this section we describe the generative models onto which Algorithm 1 will be applied. Namely, we consider mixtures of distributions belonging to some exponential family. As presented in [3], a distribution from an exponential family may be associated to a Bregman divergence via Legendre duality of convex functions. For a particular distribution, the corresponding Bregman divergence is more adapted for the clustering than other divergences [3].

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

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

The function ψ is called the cumulant function and θ is the natural parameter. For this model, the expectation of $P_{\psi,\theta}$ may be expressed as $\mu(\theta) = \nabla_\theta \psi$. We define

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

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. Table 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	$-k+k\log\left(\frac{k}{\mu}\right)$	$\frac{k}{\mu}\left(\mu\log\left(\frac{\mu}{x}\right)+x-\mu\right)$

Table 1: Exponential family distributions and associated Bregman divergences.

As emphasized in [3], clustering with Bregman divergence may be thought of as a hard-threshold model-based clustering scheme, where components of the model are assumed to belong to some exponential family. The following Remark 19 gives an illustration of this connection in a simple case.

Remark 19. *We let $k=2$, $\theta_1\neq\theta_2$, $z_1^*,\dots,z_n^*$ be hidden labels in $\{1,2\}$, and $X_1,\dots,X_n$ be an independent sample such that X_i has density*

$$\mathbb{1}_{z_i^*=1}p_{\psi,\theta_1}(x)+\mathbb{1}_{z_i^*=2}p_{\psi,\theta_2}(x),$$

where $p_{\psi,\theta_j}(x)=\exp(-d_\phi(x,\mu_j)+\phi(x))$, for $j\in\{1,2\}$. The parameters of this model are $(z_i^*)_{i\in[\![1,n]\!]},\theta_1,\theta_2$. This model slightly differs from a classical mixture model since the labels are not assumed to be drawn at random.

Let $z_{i,j}$, $i\in[\![1,n]\!]$, $j\in\{1,2\}$, denote assignment variables, that is such that $z_{i,j}=1$ if X_i is assigned to class j and 0 otherwise. Also denote by $m=\sum_{i=1}^n z_{i,1}$, $n-m=\sum_{i=1}^n z_{i,2}$, $\bar{X}_1=\sum_{i=1}^n X_i z_{i,1}/m$, $\bar{X}_2=\sum_{i=1}^n X_i z_{i,2}/(n-m)$. Maximizing the log-likelihood of the observations boils down to maximizing in $(z_{i,j})_{i,j}$:

$$\ln\prod_{i=1}^n\exp\left[-z_{i,1}d_\phi\left(X_i,\bar{X}_1\right)-z_{i,2}d_\phi\left(X_i,\bar{X}_2\right)+\phi(X_i)\right]$$
$$=-\sum_{i=1}^n z_{i,1}d_\phi\left(X_i,\bar{X}_1\right)-\sum_{i=1}^n z_{i,2}d_\phi\left(X_i,\bar{X}_2\right)+\sum_{i=1}^n\phi(X_i).$$

On the other hand, since optimal codebooks are local means of their Bregman-Voronoi cells (Proposition 9), minimizing $P_n d_\phi(.,\mathbf{c})$ is equivalent to minimizing $\sum_{i=1}^n z_{i,1}d_\phi\left(X_i,\bar{X}_1\right)+\sum_{i=1}^n z_{i,2}d_\phi\left(X_i,\bar{X}_2\right)$. Thus, clustering with Bregman divergences is the same as maximum likelihood clustering based on this model. Further, if we assume that μ_1 and μ_2 are known, then the Bregman assignment rule $x\mapsto\arg\min_{j\in\{1,2\}}d_\phi(x,\mu_j)$ is the Bayes rule.

4.3 Calibration of trimming parameter and number of clusters

When the number of clusters k is known beforehand, we propose the following 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. 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

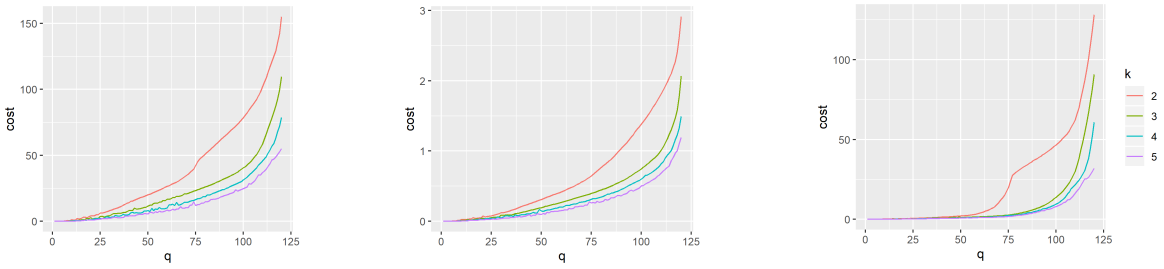
the parameter q gets large enough, it is likely that the procedure begins to assign outliers to clusters, which dramatically deprecates the empirical distortion.

Whenever both k (number of clusters) and q are unknown, we propose to select these two parameters following the same principle as the algorithm `tclust` [22]. First we draw, for different values of k, the cost curves $q \mapsto cost_k[q]$, for $1 \leq q \leq n$. For each curve, the q's for which there is an abrupt slope increase can correspond to cases where outliers are assigned to clusters, or where some small clusters are included in the set of signal points (if k is chosen too small). In the sequel, we split $[\![1, n]\!]$ into several bins $[\![q_j, q_{j+1}]\!]$. On every such bin, we select a k that provides a significant cost decrease, as well as the q yielding a slope jump. Note that this heuristic may result in several possible pairs (k, q), corresponding to different point of views, depending on what data point are considered as outliers or not. An illustration of this fact is given in Section 4.6, where outliers consist in small additional clusters.

4.4 Comparative performances of Bregman clustering for mixtures with noise

To assess the good behavior of our procedure with respect to outliers, we replicate some experiments in [3], with additional noise. We consider mixture models of Gaussian, Poisson, Binomial, Cauchy and Gamma distributions in $\mathbb{R}^2$. Namely, we sample 100 points from $X = (X^1, X^2)$, where X^1 and X^2 are independent, distributed according to a mixture distribution with 3 components. In each case, the means of the components are set to $10, 20, 40$. The weights of the components are $(1/3, 1/3, 1/3)$. We also consider a mixture of 3 different components in $\mathbb{R}^2$: Gamma, Gaussian and Binomial, with respective means 10, 20 and 40. In the Gaussian case, the standard deviations of the components are set to 5, in the Binomial case, the number of trials are set to 100 and in the Gamma case the shape parameters are set to 40. Since Gaussian and Cauchy distributions take negative values, we force the points from each components to lie respectively in the squares $[0, 20]^2$, $[0, 40]^2$ and $[0, 80]^2$. 20 outliers are added, uniformly sampled on $[0, 60]^2$.

First, we use Algorithm 1 with 20 random starts for each of these noisy mixture distributions, using the corresponding divergence, and also make the same experiment for the Cauchy distribution with squared Euclidean distance. For these procedures, we select k and q following the heuristic exposed in Section 4.3. According to Figure 1, this leads to the choice $k = 3$, $q = 104$ for the Gaussian mixture and $q = 110$ for the other mixtures. The resulting partitions for the selected parameters are depicted in Figure 2.



Gaussian mixture Poisson mixture Cauchy mixture

Figure 1: Cost curves for selection of k and q.

Then we compare the proposed method, in every case, to clustering with other Bregman divergences (including trimmed k-means [16]), trimmed k-median [9], `tclust` [22], and density/distance functions-based clustering schemes such as a robustified version of the classical single linkage procedure, the ToMATo algorithm [14] with the inverse of the distance-to-measure function [12] and `dbscan` [25]. Details concerning these methods are available in the Appendix, Section 11.1.1. Quality of partitions is assessed via the normalized mutual information (NMI, [36]) with respect to the ground truth clustering, where the "noise" points are assigned to one same cluster.

This experiment is repeated 1000 times, the results in terms of NMI's are exposed in Figure 3: Algorithm 1 refers to our method with $q = 110$ and $k = 3$. For the Cauchy and heterogeneous distributions, the Gaussian

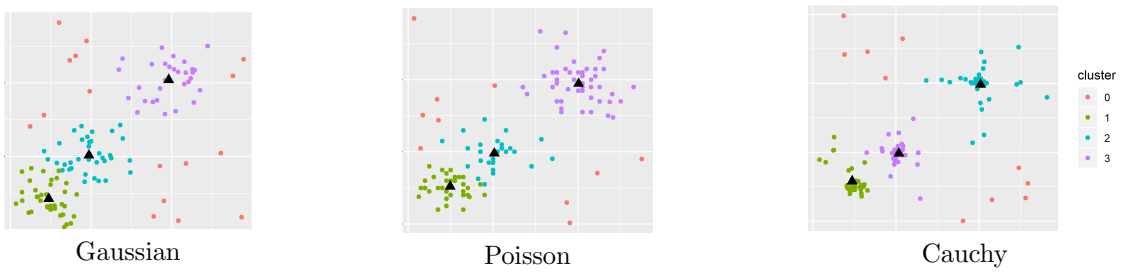


Figure 2: Clustering associated to the selected parameters k and q, where cluster 0 refers to noise.

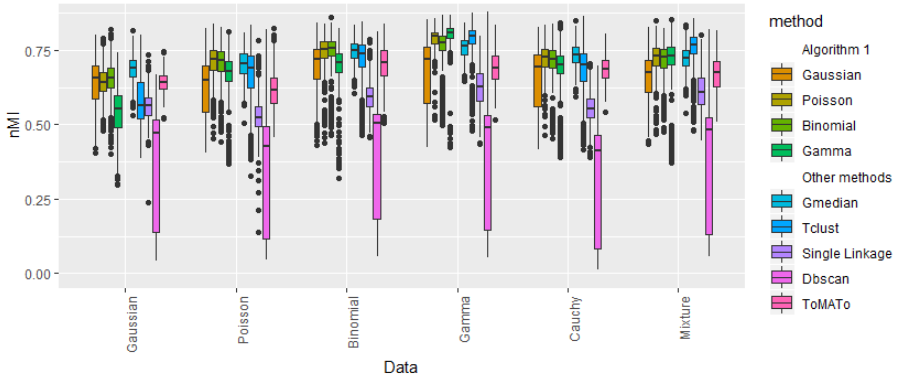


Figure 3: Comparison of robust clustering methods, for mixtures of Binomial, Gamma, Gaussian, Poisson, Cauchy, and heterogeneous distributions.

divergence is less efficient, since the 3 clusters have increasing variances. This divergence is well suited for clusters with the same variance, the Gamma and Poisson divergence for data with increasing variance, the Binomial divergence for data with increasing and then decreasing variances, for the proper parameter N. It is also possible to choose a different Bregman divergence for the different coordinates. Further explanations and numerical illustrations are available in Section 11.2. Choosing between a Gamma or a Poisson divergence depends on the knowledge on the data, as illustrated in the following section with two different real datasets. Note that Algorithm 1 with the proper Bregman divergence (almost) systematically outperforms other clustering schemes. This point is confirmed in Section 11.1.2 for large datasets ($n = 12000$).

4.5 Daily waterfall data

We consider the daily rainfalls (expressed in mm) for january (241 data points) and september (88 data points), from 2007 to 2017, in Cayenne/Rochambeau. Datapoints are defined as the amount of rain within a rainy day. According to [15], the positive daily rainfalls within one month are often modeled as Gamma distribution with parameters depending on the month. We experiment Algorithm 1 with the Gamma and the Gaussian Bregman divergences (the latter is plain trimmed k-means). The NMI's between the true labels (i.e. the month from which the datapoint was extracted) and the labels returned by the algorithm for different trimming parameters q are depicted in the right panel of Figure 4. When q is small, the Gaussian divergence yields better NMI's than Gamma. In this case, outliers are considered as a significant cluster in the computation of the NMI. Thus, the "outlier" cluster associated with Gaussian divergence seems closer to a real cluster than the Gamma one. When q is large enough (small amount of outliers), the clustering associated with Gamma divergence outperforms the Gaussian clustering. The left panel of Figure 4 depicts

the associated clustering, for $q = 300$. Of course we cannot expect a perfect clustering since the true clusters are not well-separated. However, it seems that the Gamma divergence clustering allows to consider small precipitations as outliers, contrary to the Gaussian case. This point can be further exploited to choose in practice an appropriate Bregman divergence for to the data to be clustered. For instance, in the case of positive data points, if noise points are expected close to zero, then a Poisson or Gamma divergence might be more suitable than a Gaussian one. Again, the choice of an appropriate Bregman divergence depends on prior knowledge on the structure of data and noise.

Figure 4: From left to right: Clustering with Gaussian divergence and $q = 300$. Clustering with Gamma divergence and $q = 300$. NMI as a function of the trimming parameter q.

4.6 Authors stylometric clustering

In this Section we perform clustering on texts based on stylometric descriptors exposed in [38, Section 10]. To be more precise, raw data consist in 26 annotated texts from 4 authors (Mark Twain, Sir Arthur Conan Doyle, Nathaniel Hawthorne and Charles Dickens). These texts are available as supplementary material for [38], and are framed as a sequence of lemmatized string characters (for instance "be" and "is" are instances of the same lemma "be"). Following [38], we base our stylometric comparison on lemmas corresponding to nouns, verbs and adverbs, and split every original text in chunks of size 5000 of such lemmas that will be considered as data points. Then the 50 overall most frequent lemmas are chosen, and every chunk is described as the vector of counts of these lemmas within it. Thus, signal points consists of 189 count vectors with dimension 50, originating from 4 different authors.

The signal points are corrupted using the same process for the 8 State of the Union Addresses given by Barack Obama (available in `obama` dataset from package CleanNLP in R), resulting in 5 additional points, and for the King James Version of the Bible (available on Project Gutenberg) that we preliminary lemmatize using the CleanNLP package, resulting in 15 more additional points. Our final dataset consists of the 189 signal points and the 20 outlier points described above. Slightly anticipating, these 20 outliers might also be thought of as two additional small clusters with size 5 and 15.

Since every individual lemma count can be modeled as a Poisson random variable in the random character sequence model [20], the appropriate Bregman divergence for this dataset is likely to be the Poisson divergence. In the following, we compare our method with Poisson divergence to trimmed k-means, trimmed k-medians, and t-clust.

In Figure 5, we draw the cost of our method as a function of q, for different cluster numbers k. According to this figure, several choices of k and q are possible. For values of q up to 175, the significant jumps in the risk function are for $k = 3$ and $k = 6$. For $k = 3$, the slope heuristic yields $q = 175$, whereas for $k = 6$ the slope heuristic suggests that no data points might be considered as outliers. When q ranges between 175 and 193, the significant distortion jumps are for $k = 4$ and $k = 6$, another possible choice is then $k = 4$ and $q = 188$. When q is larger than 193, the only significant jump is for $k = 6$. To summarize, the pairs $(k = 3, q = 175)$, $(k = 4, q = 188)$, $(k = 6, q = n = 209)$ seem reasonable. These three solutions correspond to the 3 natural trimmed partitions: clustering only 3 authors writings (Twain writings being considered as outliers), clustering the 4 authors writings and removing the outliers from the Bible and B. Obama addresses, and at last clustering the six sources of writings (none of them being considered as noise). The two latter situations are depicted in Figure 6, in the 2-dimensional basis given by a linear discriminant analysis of the proposed clustering.

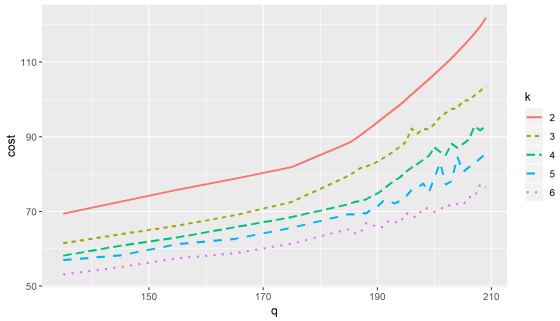


Figure 5: Cost curves for authors clustering with Poisson divergence.

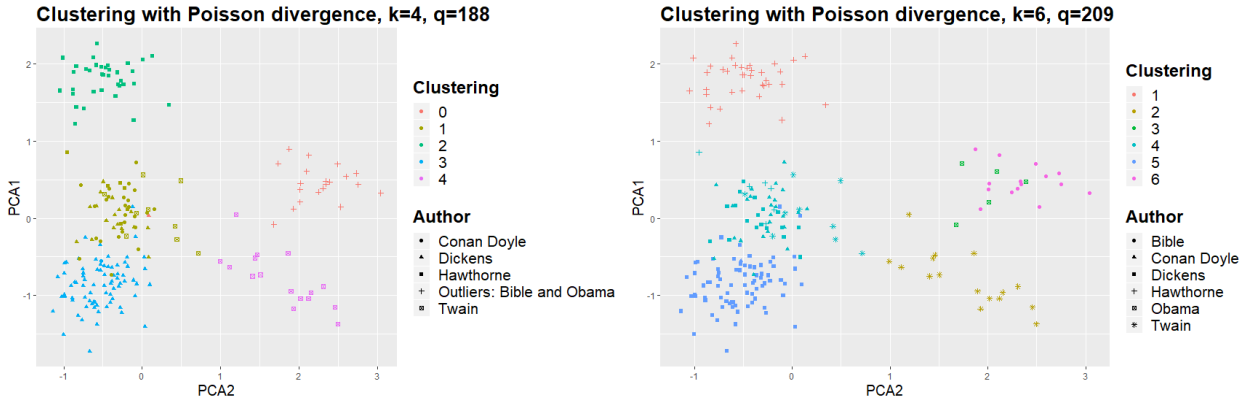


Figure 6: Author stylometric clustering with Poisson divergence.

For $k = 6$ and $q = 209$, our clustering globally retrieves the corresponding author. When $k = 4, q = 188$ is chosen, outliers are correctly identified and only one sample text from C. Dickens is labeled as outlier. The sample points seem on the whole well classified, that is assessed by a NMI of 0.7347. This performance is compared with the other clustering algorithms in Table 2. Note that values of q have been chosen to minimize the NMI, leading to $q = 190$ for trimmed k-means, $q = 202$ for trimmed k-medians, and $q = 184$ for `tclust`. The NMI curves may be found in the Appendix, Section 11.

Method	trimmed 4-means	trimmed 4-medians	tclust	Poisson
NMI	0.5336	0.4334	0.4913	0.7347

Table 2: Comparison of robust clustering methods for Author retrieving.

The associated partitions for k-median and `tclust` are depicted in Figure 7, showing that these two methods fail in correctly identifying outliers.

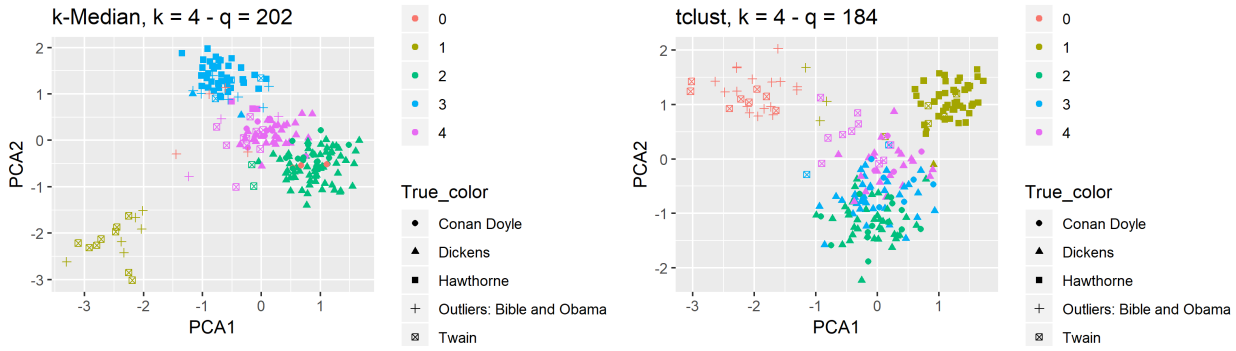


Figure 7: Author stylometric clustering with trimmed k-median and tclust.

5 Proofs for Section 2

5.1 Intermediate results

The proofs of Theorem 8, Theorem 10 and Theorem 11 make extensive use of the following lemmas, whose proofs are deferred to the Appendix, Section 9. 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 20. *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*

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

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 $s \leq h$, then $r_s(\mathbf{c}) \leq r^+$.

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

Lemma 21. *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 $\bar{B}(0,K) \cap F_0$, 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 $D(\mathbf{c}, \mathbf{c}') = \min_{\sigma \in \Sigma_k} \max_{j \in [\![1,k]\!]} |c_j - c'_{\sigma(j)}|$ (cf. Theorem 10).

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

Lemma 22. *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 $s_0 < h$ such that*

$$\forall s_0 < s < h, \quad \sup_{\mathbf{c} \in \left(F_0 \cap \bar{B}(0,K)\right)^{(k)}} R_h(\mathbf{c}) - R_s(\mathbf{c}) \leq \epsilon.$$

5.2 Proof of Lemma 5

Set $0 < h < h' < 1$, and recall that $F_{\mathbf{c}}^{-1}(u) = r_u^2(\mathbf{c})$ denotes 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

$$\frac{R_h(\mathbf{c})}{h} = \int_0^1 F_{\mathbf{c}}^{-1}(hu) du \leq \int_0^1 F_{\mathbf{c}}^{-1}(h'u) du = \frac{R_{h'}(\mathbf{c})}{h'}.$$

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, $L_{\mathbf{c}} := \lim_{l' \to 0} F_{\mathbf{c}}^{-1}(l')$ exists. Moreover, for $l < h'$, $F_{\mathbf{c}}^{-1}(hu) = F_{\mathbf{c}}^{-1}(h'u)$ a.s., and $F_{\mathbf{c}}^{-1}(l) = \lim_{n\to\infty} F_{\mathbf{c}}^{-1}\left((h/h')^n l\right) = L_{\mathbf{c}}$, that is, $r_l^2(\mathbf{c}) = \lim_{l' \to 0} r_{l'}^2(\mathbf{c})$. From (1), it follows that $P(B_\phi(\mathbf{c}, r_{h'}(\mathbf{c}))) = 0$. Conversely, equality holds when $P(B_\phi(\mathbf{c}, r_{h'}(\mathbf{c}))) = 0$. □

5.3 Proof of Theorem 8

The intuition behind the proof of Theorem 8 is that optimal codebooks satisfy a so-called centroid condition, namely their code points are means of their trimmed Bregman-Voronoi cells. Thus, provided that optimal Bregman-Voronoi cells have enough weight, the assumption $P\|u\| < +\infty$ leads to a bound on the norm of these code points. This idea is summarized by the following lemma, that is also a key ingredient in the proofs of the results of Section 3.

Lemma 23. *Assume that the requirements of Theorem 8 are satisfied. For every $k \geq 2$, if $R^*_{k-1,h} - R^*_{k,h} > 0$, then*

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

Moreover there exist $h^-, h^+ \in (0,1)$ with $h \in (h^-, h^+)$ such that, for every $j \in [\![2,k]\!]$, $R^*_{j-1,h^-} - R^*_{j,h^+} \geq \frac{\alpha}{2}$.

For every $b \in (0, h \wedge (1-h)]$, set $h_b^- = (h-b)/(1-b)$ and $h_b^+ = h/(1-b)$. Let b be such that $\min_{j\in[\![2,k]\!]} R^*_{j-1,h_b^-} - R^*_{j,h_b^+} > 0$, and, for κ_1 in $(0,1)$, set $b_1 = \kappa_1 b$. Then, for every $s \in [h_{b_1}^-, h_{b_1}^+]$ and $j \in [\![1,k]\!]$, there exists a minimizer $\mathbf{c}^*_{j,s}$ of $R_{j,s}$ satisfying

$$\forall p \in [\![1,j]\!], \quad \|c^*_{j,s,p}\| \leq \frac{P\|u\|}{b(1-h)(1-\kappa_1)}.$$

The proof of Lemma 23 is deferred to the Appendix, Section 9. When $R^*_{k-1,h} - R^*_{k,h} > 0$, Theorem 8 follows from Lemma 23. In the case where $R^*_{k-1,h} - R^*_{k,h} = 0$, there exists a set A with $P(A) \geq h$ such that the restriction of P to A is supported by at most $k-1$ points. These $k-1$ points provide an optimal k-points codebook. Hence the result of Theorem 8. □

5.4 Proof of Proposition 14

The first part of Proposition 14 follows from Lemma 23. Indeed, if h^- and h^+ are such that $\min_{j\in[\![2,k]\!]} R^*_{j-1,h^-} - R^*_{j,h^+} > 0$, then for b small enough so that $h^- \leq h_b^- < h < h_b^+ \leq h^+$, we have $\min_{j\in[\![2,k]\!]} R^*_{j-1,h_b^-} - R^*_{j,h_b^+} \geq \min_{j\in[\![2,k]\!]} R^*_{j-1,h^-} - R^*_{j,h^+} > 0$.

We turn to the second part of Proposition 14. Let $\mathbf{c}^{*,(j)}$ be a j-points h-trimmed optimal codebook, and $p_{j,h} = h \min_{l\in[\![1,j]\!]} \tilde{P}_{\mathbf{c}^{*,(j)}}\left(W_l(\mathbf{c}^{*,(j)})\right)$, where $\tilde{P}_{\mathbf{c}^{*,(j)}} \in \mathcal{P}_h(\mathbf{c}^{*,(j)})$. Let $\tau_{j,h}$ denote the $[0,1]$-valued function such that $h\tilde{P}_{\mathbf{c}^{*,(j)}} = P\tau_{j,h}$. Assume that $p_{j,h} = P\tau_{j,h}(u)\mathbb{1}_{W_1(\mathbf{c}^{*,(j)})}(u)$, without loss of generality. Then we have

$$R^*_{j,h} \geq \sum_{l=2}^{k} Pd_\phi(u, c_l^{*,(j)})(u)\tau_{j,h}(u)\mathbb{1}_{W_l(\mathbf{c}^{*,(j)})}(u) \geq R^*_{j-1,h-p_{j,h}}.$$

Thus, $(h - B_h)/(1-B_h) \geq h - p_{j,h}$, that entails $(1-(h-p_{j,h}))B_h \leq p_{j,h}$.

6 Proofs for Section 3

6.1 Intermediate results

Theorem 10 and 11 require some additional probabilistic results that are gathered in this subsection. Some of them are applications of standard techniques, their proofs are thus deferred to the Appendix, Section 10, for the sake of completeness. We begin with deviation bounds.

Proposition 24. *With probability larger than $1-e^{-x}$, we have, for $k \geq 2$,*

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

where $\partial B_\phi(\mathbf{c},r)$ denotes $\left\{x \mid d_\phi(x,\mathbf{c}) = r^2\right\}$ and C denotes a universal constant. Moreover, if r^+ and K are fixed, and $P\|u\|^2 \leq M_2^2 < \infty$, we have, with probability larger than $1-e^{-x}$,

$$\sup_{\mathbf{c}\in\left(\bar{B}(0,K)\cap F_0\right)^{(k)},r\leq r^+} \left|(P-P_n)d_\phi(.,\mathbf{c})\mathbb{1}_{B_\phi(\mathbf{c},r)}\right| \leq (r^+)^2\left[C_{K,r^+,M_2}\frac{\sqrt{kd\log(k)}}{\sqrt{n}} + \sqrt{\frac{2x}{n}}\right], \tag{3}$$
$$\sup_{\mathbf{c}\in\left(\bar{B}(0,K)\cap F_0\right)^{(k)},r\leq r^+} \left|(P-P_n)d_\phi(.,\mathbf{c})\mathbb{1}_{\partial B_\phi(\mathbf{c},r)}\right| \leq (r^+)^2\left[C_{K,r^+,M_2}\frac{\sqrt{kd\log(k)}}{\sqrt{n}} + \sqrt{\frac{2x}{n}}\right],$$

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

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

Proposition 25. *Assume that $P\|u\|^p < +\infty$ for some $p \geq 2$, and let $b > 0$ be such that $\min_{j\in[\![2,k]\!]} R^*_{j-1,h_b^-} - R^*_{j,h_b^+} > 0$, where $h_b^- = (h-b)/(1-b)$, $h_b^+ = h/(1-b)$, as in Lemma 23. Let $\kappa_2 < 1$, and denote by $b_2 = \kappa_2 b$. Then there exists $C_{P,h,k,\kappa_2,b}$ such that, for n large enough, with probability larger than $1-n^{-\frac{p}{2}}$, we have, for all $j \in [\![2,k]\!]$, and $i \in [\![1,j]\!]$,*

$$\sup_{h_{b_2^-}\leq s\leq h} \|\hat{c}_{j,s,i}\| \leq C_{P,h,k,\kappa_2,b},$$

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

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

Lemma 26. *If $P\|u\|^p < \infty$ for some $p \geq 2$, then there exists some positive constant C such that with probability larger than $1-n^{-\frac{p}{2}}$, $P_n\|u\| \leq C$.*

At last, a technical lemma on empirical quantiles of Bregman divergences will be needed.

Lemma 27. *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 $n \in \mathbb{N}$,*

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

6.2 Proof of Theorem 10

The proof of Theorem 10 is an adaptation of the proof of [16, Theorem 3.4]. 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, b. Thus $\psi^{-1}(\{0\})$ is a closed set. Moreover, since ψ is strictly convex, any line that contains 0 contains at most two points of $\psi^{-1}(\{0\})$. Thus, the Lebesgue measure of $\psi^{-1}(\{0\})$ is 0. Since P is absolutely continuous, it follows that boundaries of Bregman balls have P-mass equal to 0.

According to Proposition 25, provided that $P\|u\|^p < +\infty$, for some $p > 2$, there exists $C_P > 0$ such that for some $N \in \mathbb{N}$, $\sum_{n \geq N} P(\max_{i \in [\![1,k]\!]} \|\hat{c}_{n,h,i}\| > C_P) < \infty$. Thus, the Borel-Cantelli Lemma ensures that, a.s. for n large enough, for every $i \in [\![1,k]\!]$, $\|\hat{c}_{n,h,i}\| \leq C_P$. According to the Skorokhod's representation theorem in the Polish space $\mathbb{R}^d$, there exists a measured space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ and a sequence of random variables $(X_n)_{n \in \mathbb{N}}$ along with 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 $\mathbf{c}^*$ a minimizer 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

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

According to Lemma 27, with $K = \|c_1^*\|$ for instance, it comes $r'_n \leq r^+$, for some finite r^+. Thus, up to extracting a subsequence, we may assume that $r'_n \to r'_0$ for some $r'_0 \leq r_+$. Moreover, it holds

$$|d_\phi(X_n, \mathbf{c}^*) - d_\phi(X, \mathbf{c}^*)| \leq |\phi(X_n) - \phi(X)| + \max_{j \in [\![1,k]\!]} \|\nabla_{c_j^*} \phi\| |X_n - X|. \quad (4)$$

Thus, $d_\phi(X_n, \mathbf{c}^*) \to d_\phi(X, \mathbf{c}^*)$ a.e. 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 dominated 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, $\limsup_{n \to \infty} R_{n,h}(\hat{\mathbf{c}}_n) \leq R^*_{k,h}$.

Since, for $n \geq N$ and every $i \in [\![1,k]\!]$, $\|\hat{c}_{n,i}\| \leq C_P$, we have $\hat{c}_{u(n),i} \to c_i$ for some $c_i \in F_0 \cap \bar{B}(0, C_P)$, where $\hat{\mathbf{c}}_{u(n)}$ is a subsequence. Set $\mathbf{c} = (c_1, c_2, \ldots, c_k)$. Again, according to Lemma 27 with $K = C_P$, it comes that $r_{u(n),h}(\hat{\mathbf{c}}_{u(n)}) \to r$ for some $r \geq 0$. Therefore, from (4), Lemma 21 and the continuity of P,

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

where $\tau_{u(n)} = \mathbb{1}_{B_\phi(\hat{\mathbf{c}}_{u(n)}, r_{u(n),h}(\hat{\mathbf{c}}_{u(n)}))}$. According to the dominated convergence theorem, we have $h = P(B_\phi(\mathbf{c}, r)) = P_{u(n)}(\tau_{u(n)}(u))$. Again, the dominated convergence theorem implies that

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

As a consequence, $\lim_{n \to \infty} R_{u(n),h}(\hat{\mathbf{c}}_{u(n)}) = R_h(\mathbf{c}) = R^*_{k,h}$ and $\mathbf{c}$ is an optimal trimmed codebook. Since, given a subsequence of $(\hat{\mathbf{c}}_n)_{n \in \mathbb{N}}$, we may find a subsequence of indices $u(n)$ such that $\lim_{n \to \infty} R_{u(n),h}(\hat{\mathbf{c}}_{u(n)}) = R^*_{k,h}$, we deduce that $\lim_{n \to +\infty} R_{n,h}(\hat{\mathbf{c}}_n) = R^*_{k,h}$.

Now assume that $\mathbf{c}^*_h$ is unique. Then, for every subsequence of $(\hat{\mathbf{c}}_n)_{n \in \mathbb{N}}$, there exists $u(n)$ such that $\hat{\mathbf{c}}_{u(n)} \to \mathbf{c} = \mathbf{c}^*_h$. Thus, a.e., $\hat{\mathbf{c}}_n \to \mathbf{c}^*_h$. □

6.3 Proof of Theorem 11

For $h > 0$ and a codebook $\mathbf{c}$, we denote by $\tau_h(\mathbf{c})$ the trimming function $\mathbb{1}_{B_\phi(\mathbf{c}, r_h(\mathbf{c}))} + \delta_h(\mathbf{c}) \mathbb{1}_{\partial B_\phi(\mathbf{c}, r_h(\mathbf{c}))}$, so that $P\tau_h(\mathbf{c})/h \in \mathcal{P}_h(\mathbf{c})$. We also denote by $\hat{\tau}_h(\mathbf{c})$ its empirical counterpart. Note that $\delta_h(\mathbf{c})$ and $\hat{\delta}_h(\mathbf{c})$ are smaller than 1. It follows that

$$\begin{aligned}
|P\tau_h(\mathbf{c}) - P\hat{\tau}_h(\mathbf{c})| &= |(P - P_n)\hat{\tau}_h(\mathbf{c})| \\
&\leq |(P - P_n) B_\phi(\mathbf{c}, r_{n,h}(\mathbf{c}))| + \hat{\delta}(\mathbf{c}) |(P - P_n) \partial B_\phi(\mathbf{c}, r_{n,h}(\mathbf{c}))| \\
&\leq |(P - P_n) B_\phi(\mathbf{c}, r_{n,h}(\mathbf{c}))| + |(P - P_n) \partial B_\phi(\mathbf{c}, r_{n,h}(\mathbf{c}))|.
\end{aligned}$$

As well, we bound $|P_n \tau_h(\mathbf{c}) - P_n \hat{\tau}_h(\mathbf{c})|$ the same way. Combining Lemma 23 and Proposition 25, we consider a probability event onto which, for all j, $\|\hat{\mathbf{c}}_{n,j}\| \leq C_P$ and $\sup_{\mathbf{c} \in (F_0 \cap \bar{B}(0, C_P))^{(k)}} r_{n,h}(\mathbf{c}) \vee r_h(\mathbf{c}) \leq r^+$. This

occurs with probability at least $1 - n^{-p/2}$ for C_P and r^+ large enough (more details are given in Appendix, Section 10.2). We also assume that the deviation bounds of Proposition 24 hold, with parameter C_P and r_+, to define a global probability event with mass larger than $1 - n^{-p/2} - 2e^{-x}$. On this event, we have

$$\begin{aligned}
R_h(\hat{\mathbf{c}}_n) - R^*_{k,h} = \; & Pd_\phi(u, \hat{\mathbf{c}}_n)\hat{\tau}_h(\hat{\mathbf{c}}_n) - Pd_\phi(u, \mathbf{c}^*)\hat{\tau}_h(\mathbf{c}^*) \\
& + Pd_\phi(u, \hat{\mathbf{c}}_n)(\tau_h(\hat{\mathbf{c}}_n) - \hat{\tau}_h(\hat{\mathbf{c}}_n) - (Pd_\phi(u, \mathbf{c}^*)(\tau_h(\mathbf{c}^*) - \hat{\tau}_h(\mathbf{c}^*)) \\
\leq \; & 2 \sup_{\mathbf{c} \in (F_0 \cap \bar{B}(0, C_P))^{(k)}, r \leq r^+} |(P - P_n) d_\phi(u, \mathbf{c}) \mathbb{1}_{B_\phi(\mathbf{c}, r)}(u)| \\
& + 2(r^+)^2 \sup_{\mathbf{c} \in \Omega^{(k)}, r \geq 0} |(P - P_n) B_\phi(\mathbf{c}, r)| \\
& + 2 \sup_{\mathbf{c} \in (F_0 \cap \bar{B}(0, C_P))^{(k)}, r \leq r^+} |(P - P_n) d_\phi(u, \mathbf{c}) \mathbb{1}_{\partial B_\phi(\mathbf{c}, r)}(u)| \\
& + 2(r^+)^2 \sup_{\mathbf{c} \in \Omega^{(k)}, r \geq 0} |(P - P_n) \partial B_\phi(\mathbf{c}, r)|.
\end{aligned}$$

Therefore, $R_h(\hat{\mathbf{c}}_n) - R^*_{k,h} \leq C_P(1 + \sqrt{x})/\sqrt{n}$, for some constant C_P. □

6.4 Proof of Corollary 12

Denote by A the intersection of the probability events described in Proposition 25, that has probability larger than $1 - n^{-\frac{p}{2}}$. Decomposing the excess risk as in the proof of Theorem 11 yields

$$R_h(\hat{\mathbf{c}}_n) - R^*_{k,h} = (R_h(\hat{\mathbf{c}}_n) - R^*_{k,h})\mathbb{1}_A + (R_h(\hat{\mathbf{c}}_n) - R^*_{k,h})\mathbb{1}_{A^c}.$$

According to Proposition 24, we have $\mathbb{E}((R_h(\hat{\mathbf{c}}_n) - R^*_{k,h})\mathbb{1}_A) \leq C_P/\sqrt{n}$. It only remains to bound the expectation of the second term. This is the aim of the following Lemma, whose proof is deferred to the Appendix, Section 10.5.

Lemma 28. *Assume that $P\|u\|^q \psi^q(k\|u\|/h) < \infty$. Then there exists a constant C_q such that $\mathbb{E}R^q_h(\hat{\mathbf{c}}_n) \leq C^q_P$.*

Equipped with Lemma 28, we may bound $\mathbb{E}(R_h(\hat{\mathbf{c}}_n)\mathbb{1}_{A^c})$ as follows, using Hölder's inequality,

$$\mathbb{E}(R_h(\hat{\mathbf{c}}_n)\mathbb{1}_{A^c}) \leq (\mathbb{P}(A^c))^{\frac{1}{p}} \left(\mathbb{E}R^q_h(\hat{\mathbf{c}}_n)\right)^{\frac{1}{q}} \leq C_P/\sqrt{n}.$$

6.5 Proof of Theorem 15

A key ingredient of the proof of Theorem 15 is the following lemma, ensuring that every cell of a trimmed and corrupted empirical distortion minimizer contains a minimal portion of signal points. In what follows, the $\hat{\tau}$'s are the trimming function with respect to P_n (uncorrupted sample), as defined in Section 6.3.

Lemma 29. *Assume that $B_h > 0$ (see Definition 13), let $b < B_h$ and $b < b_1 < B_h$ such that $b = \kappa_1 b_1$, with $\kappa_1 < 1$. Denote by $\beta_1 = (1 - \kappa_1)b_1 [h \wedge (1 - h)]/2$. Assume that $s/(n + s) \leq b$. Then, for n large enough, with probability larger than $1 - n^{-\frac{p}{2}}$, we have, for all $j \in [\![1, k]\!]$,*

$$P_n\left(\hat{\tau}_{h_b^-}(\hat{\mathbf{c}}_{n+s,h})\mathbb{1}_{W_j(\hat{\mathbf{c}}_{n+s,h})}\right) \geq \beta_1.$$

The proof of Lemma 29 is postponed to the Appendix, Section 10.6. We are now in a position to prove Theorem 15.

Proof of Theorem 15. We adopt the same notation and assumptions as in the proof of Lemma 29. We may write

$$(n + s)\hat{R}_{n+s,h}(\hat{\mathbf{c}}_{n+s,h}) \leq n\left(R^*_{h_b^+ + \beta_n} + \alpha_n\right).$$

On the other hand, recall that for all $j \in [\![1,k]\!]$, $P_n\left(\hat{\tau}_{h_b^-}(\hat{\mathbf{c}}_{n+s,h})\mathbb{1}_{W_j(\hat{\mathbf{c}}_{n+s,h})}\right) \geq \beta_1$, and denote by $\mathbf{m} = (m_1,\ldots,m_k)$ the codebook such that

$$m_j = \left[P_n(u\hat{\tau}_{h_b^-}(\hat{\mathbf{c}}_{n+s,h})\mathbb{1}_{W_j(\hat{\mathbf{c}}_{n+s,h})})\right] / \left[P_n\left(\hat{\tau}_{h_b^-}(\hat{\mathbf{c}}_{n+s,h})\mathbb{1}_{W_j(\hat{\mathbf{c}}_{n+s,h})}\right)\right].$$

Then, for all j, $\|m_j\| \leq P_n\|u\|/\beta_1 \leq C_P$. Using Proposition 2, we may write

$$\begin{aligned}\hat{R}_{n+s,h}(\hat{\mathbf{c}}_{n+s,h}) &\geq n\hat{R}_{n,h_b^-}(\hat{\mathbf{c}}_{n+s,h})/(n+s)\\ &\geq \frac{n}{n+s}\left[\sum_{j=1}^{k} P_n\left(\hat{\tau}_{h_b^-}(\hat{\mathbf{c}}_{n+s,h})\mathbb{1}_{W_j(\hat{\mathbf{c}}_{n+s,h})}\right) d_\phi(m_j,\hat{c}_{n+s,h,j})\right.\\ &\qquad\left. + \hat{R}_{n,h_b^-}(\mathbf{m})\right].\end{aligned}$$

The last term satisfies

$$\begin{aligned}\hat{R}_{n,h_b^-}(\mathbf{m}) &\geq P_n d_\phi(u,\mathbf{m})\hat{\tau}_{h_b^-}(\mathbf{m}) \geq P_n d_\phi(u,\mathbf{m})\tau_{h_b^- - \beta_n}(\mathbf{m})\\ &\geq P d_\phi(u,\mathbf{m})\tau_{h_b^- - \beta_n}(\mathbf{m}) - \alpha_n \geq R^*_{h_b^- - \beta_n} - \alpha_n.\end{aligned}$$

Thus, for all $j \in [\![1,k]\!]$,

$$\beta_1 d_\phi(m_j,\hat{c}_{n+s,h,j}) \leq \left[2\alpha_n + R^*_{h_b^+ + \beta_n} - R^*_{h_b^- - \beta_n}\right]. \tag{5}$$

□

6.6 Proof of Corollary 16

With the same setting as Theorem 15, according to (5), we have, almost surely, for n large enough $d_\phi(K,\hat{\mathbf{c}}_{n+s,j}) \leq 2[2\alpha_n + R^*_{h_b^+ + \beta_n} - R^*_{h_b^- - \beta_n}]/[b(1-\kappa_1)(h \wedge (1-h)]$, whenever $s/(n+s) \leq b$ and $B_h\kappa_1 > b$. Now let c be defined as

$$c = \sup\{r > 0 \mid \{x \mid d_\phi(K,x) \leq r\} \subset B(K,1)\}.$$

If $c > 0$, then requiring b small enough and $s/(n+s) \leq b$ ensures that $d_\phi(K,\hat{\mathbf{c}}_{n+s,j}) \leq c/2$, almost surely, for n large enough, hence the result.

Thus, it remains to prove that $c > 0$. Assume that for every $r > 0$, $\{x \mid d_\phi(K,x) \leq r\} \nsubseteq B(K,1)$. Then there exists $x_0 \in K$ and a sequence v_n satisfying $d_\phi(x_0,v_n) \to 0$, along with $\|x_0 - v_n\| > 1$. Noting that, for $t \geq 0$, $d_\phi(x_0, v_n + t(v_n - x_0)) \geq d_\phi(x_0,v_n)$ yields $d_\phi(x_0,v_n) \geq d_\phi(x_0,v_n')$, where $v_n' = x_0 + (v_n - x_0)/\|v_n - x_0\|$. Therefore, $d_\phi(v_n,x_0) \geq \inf_{\|u-x_0\|=1} d_\phi(x_0,u) > 0$, hence the contradiction.

At last, if $c \mapsto d_\phi(x,c)$ is a proper map, then $d_\phi(K,\hat{\mathbf{c}}_{n+s,j}) \leq 2[2\alpha_n + R^*_{h_b^+ + \beta_n} - R^*_{h_b^- - \beta_n}]/[b(1-\kappa_1)(h \wedge (1-h)]$, whenever $s/(n+s) \leq b$ and $B_h\kappa_1 > b$ entails that, almost surely, for n large enough, $\|\hat{\mathbf{c}}_{n+s}\|_2 < +\infty$, thus $\widehat{BP}_{n,h} > b$. Hence $\widehat{BP}_{n,h} \geq B_h$, almost surely for n large enough. □

To ease understanding, the statements of the results are recalled before the proofs. In the sequel, references with numbers refer to statements in the main paper, whereas letters are devoted to internal references.

References

[1] David Arthur and Sergei Vassilvitskii. "`k-means++`: the advantages of careful seeding". In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. ACM, New York, 2007, pp. 1027–1035.

[2] A. Banerjee, X. Guo, and H. Wang. "On the optimality of conditional expectation as a Bregman predictor". In: *IEEE Transactions on Information Theory* 51 (2005).

[3] A. Banerjee et al. "Clustering with Bregman divergences". In: *Journal of Machine Learning Research* 6 (2005), pp. 1705–1749.

[4] Gérard Biau, Luc Devroye, and Gábor Lugosi. "On the performance of clustering in Hilbert spaces". In: *IEEE Trans. Inform. Theory* 54.2 (2008), pp. 781–790. ISSN: 0018-9448. DOI: 10.1109/TIT.2007.913516. URL: https://doi.org/10.1109/TIT.2007.913516.

[5] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. "Theory of classification: a survey of some recent advances". In: *ESAIM Probab. Stat.* 9 (2005), pp. 323–375. ISSN: 1292-8100. DOI: 10.1051/ps:2005018. URL: http://dx.doi.org/10.1051/ps:2005018.

[6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013, pp. x+481. 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.

[7] 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". In: *USSR Computational Mathematics and Mathematical Physics* 7 (1967), pp. 200–217.

[8] Christian Brownlees, Emilien Joly, and Gábor Lugosi. "Empirical risk minimization for heavy-tailed losses". In: *Ann. Statist.* 43.6 (2015), pp. 2507–2536. ISSN: 0090-5364. DOI: 10.1214/15-AOS1350. URL: https://doi.org/10.1214/15-AOS1350.

[9] H. Cardot, P. Cenac, and P-A. Zitt. "Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm." In: *Bernoulli* 19 (2013), pp. 18–43.

[10] Olivier Catoni and Ilaria Giulini. "Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector". In: *ArXiv e-prints* (2018). arXiv: 1802.04308 [math.ST].

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

[12] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. "Geometric Inference for Measures based on Distance Functions". In: *Foundations of Computational Mathematics* 11.6 (2011), pp. 733–751. DOI: 10.1007/s10208-011-9098-0. URL: https://hal.inria.fr/inria-00383685.

[13] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. "Geometric Inference for Probability Measures". In: *Foundations of Computational Mathematics archive* 11 (2011), pp. 733–751.

[14] Frédéric Chazal et al. "Persistence-based clustering in Riemannian manifolds". In: *J. ACM* 60.6 (2013), Art. 41, 38. ISSN: 0004-5411. DOI: 10.1145/2535927. URL: https://doi.org/10.1145/2535927.

[15] R. Coe and R. D. Stern. "Fitting Models to Daily Rainfall Data". In: *Journal of Applied Meteorology* 21.7 (1982), pp. 1024–1031. DOI: 10.1175/1520-0450(1982)021<1024:FMTDRD>2.0.CO;2. eprint: https://doi.org/10.1175/1520-0450(1982)021<1024:FMTDRD>2.0.CO;2. URL: https://doi.org/10.1175/1520-0450(1982)021<1024:FMTDRD>2.0.CO;2.

[16] J. A. Cuesta-Albertos, A. Gordaliza, and C. Matrán. "Trimmed k-means: an attempt to robustify quantizers". In: *Ann. Statist.* 25.2 (Apr. 1997), pp. 553–576. DOI: 10.1214/aos/1031833664. URL: http://dx.doi.org/10.1214/aos/1031833664.

[17] David Donoho and Peter J. Huber. "The notion of breakdown point". In: *A Festschrift for Erich L. Lehmann*. Wadsworth Statist./Probab. Ser. Wadsworth, Belmont, CA, 1983, pp. 157–184.

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

[19] Luis A. Garcí a Escudero et al. "A general trimming approach to robust cluster analysis". In: *Ann. Statist.* 36.3 (2008), pp. 1324–1345. ISSN: 0090-5364. DOI: 10.1214/07-AOS515. URL: https://doi.org/10.1214/07-AOS515.

[20] Stefan Evert. "A simple LNRE model for random character sequences". In: *In Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (Louvain-la-Neuve*. 2004, pp. 411–422.

[21] Aurélie Fischer. "Quantization and clustering with Bregman divergences". In: *J. Multivariate Anal.* 101.9 (2010), pp. 2207–2221. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2010.05.008. URL: https://doi.org/10.1016/j.jmva.2010.05.008.

[22] Heinrich Fritz, Luis A. Garcia-Escudero, and Agustin Mayo-Iscar. "tclust: An R Package for a Trimming Approach to Cluster Analysis". In: *Journal of Statistical Software* 47.12 (2012), pp. 1–26. URL: http://www.jstatsoft.org/v47/i12/.

[23] Alfonso Gordaliza. "Best approximations to random variables based on trimming procedures". In: *J. Approx. Theory* 64.2 (1991), pp. 162–180. ISSN: 0021-9045. DOI: 10.1016/0021-9045(91)90072-I. URL: https://doi.org/10.1016/0021-9045(91)90072-I.

[24] R. M. Gray et al. "Distortion measures for speech processing". In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 28 (1980), pp. 367–376.

[25] Michael Hahsler, Matthew Piekenbrock, and Derek Doran. "dbscan: Fast Density-Based Clustering with R". In: *Journal of Statistical Software* 91.1 (2019), pp. 1–30. DOI: 10.18637/jss.v091.i01.

[26] Amit Kumar and Ravindran Kannan. "Clustering with spectral norm and the k-means algorithm". In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science—FOCS 2010*. IEEE Computer Soc., Los Alamitos, CA, 2010, pp. 299–308.

[27] G. Lecué and M. Lerasle. "Robust machine learning by median-of-means : theory and practice". In: *ArXiv e-prints* (Nov. 2017). arXiv: 1711.10306 [math.ST].

[28] Clément Levrard. "Nonasymptotic bounds for vector quantization in Hilbert spaces". In: *Ann. Statist.* 43.2 (Apr. 2015), pp. 592–619. DOI: 10.1214/14-AOS1293. URL: https://doi.org/10.1214/14-AOS1293.

[29] Clément Levrard. "Quantization/Clustering: when and why does k-means work?" In: *JSFdS* 159.1 (2018), pp. 1–26.

[30] T. Linder. "Learning-theoretic methods in vector quantization". In: *Principles of nonparametric learning (Udine, 2001)*. Vol. 434. CISM Courses and Lect. Springer, Vienna, 2002, pp. 163–210.

[31] S. P. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28 (1982), pp. 129–137.

[32] Ricardo A. Maronna, R. Douglas Martin, and Victor J. Yohai. *Robust statistics*. Wiley Series in Probability and Statistics. Theory and methods. John Wiley & Sons, Ltd., Chichester, 2006, pp. xx+403. ISBN: 978-0-470-01092-1; 0-470-01092-4. DOI: 10.1002/0470010940. URL: https://doi.org/10.1002/0470010940.

[33] S. Mendelson and R. Vershynin. "Entropy and the combinatorial dimension". In: *Invent. Math.* 152.1 (2003), pp. 37–55. ISSN: 0020-9910. DOI: 10.1007/s00222-002-0266-3. URL: http://dx.doi.org/10.1007/s00222-002-0266-3.

[34] F. Nielsen, J.D. Boissonnat, and R. Nock. *Bregman Voronoi diagrams: properties, algorithms and applications*. Tech. rep. 6154. INRIA, 2007.

[35] R. T. Rockafellar. *Convex Analysis*. Princeton, New Jersey: Princeton University Press, 1970.

[36] A. Strehl and J. Ghosh. "Cluster ensembles - A knowledge reuse framework for combining multiple partitions". In: *Journal of Machine Learning Research* 3 (2002), pp. 583–617.

[37] Cheng Tang and Claire Monteleoni. "On Lloyd's Algorithm: New Theoretical Insights for Clustering in Practice". In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Vol. 51. PMLR, 2016, pp. 1280–1289.

[38] Lauren Tilton Taylor Arnold. *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text*. 1st ed. 2015. Quantitative Methods in the Humanities and Social Sciences. Springer International Publishing, 2015. ISBN: 3319207016,978-3-319-20701-8,978-3-319-20702-5.

[39] Aad van der Vaart and Jon A. Wellner. "A note on bounds for VC dimensions". In: *High dimensional probability V: the Luminy volume*. Vol. 5. Inst. Math. Stat. (IMS) Collect. Inst. Math. Statist., Beachwood, OH, 2009, pp. 103–107. DOI: 10.1214/09-IMSCOLL508. URL: https://doi.org/10.1214/09-IMSCOLL508.

[40] D. Vandev. "A note on the breakdown point of the least median of squares and least trimmed squares estimators". In: *Statist. Probab. Lett.* 16.2 (1993), pp. 117–119. ISSN: 0167-7152. DOI: 10.1016/0167-7152(93)90155-C. URL: https://doi.org/10.1016/0167-7152(93)90155-C.

7 Technical proofs for Section 2

7.1 Proof of Lemma 7.1

Lemma (7.1). *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 $\tilde{P} \in \mathcal{P}_h(\mathbf{c})$.

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,

$$\begin{aligned} F_{\mathbf{c}}^{-1}(u) &= \inf\{s \geq 0 \mid \text{ with probability } \geq u, d_\phi(X, \mathbf{c}) \leq s\} \\ &= \inf\left\{r^2 \geq 0 \mid P(\bar{B}_\phi(\mathbf{c}, r)) \geq u\right\}. \end{aligned}$$

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

$$R(\tilde{P}_{\mathbf{c}}, \mathbf{c}) = \mathbb{E}_{\tilde{X}^*} d_\phi(\tilde{X}^*, \mathbf{c}) = \int_0^1 F_{\mathbf{c}}^{-1}(hu) du.$$

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 $P(B_\phi(\mathbf{c}, u)) \geq h\tilde{P}(B_\phi(\mathbf{c}, u))$, it holds that $\tilde{F}_{\mathbf{c}}^{-1}(u) \geq F_{\mathbf{c}}^{-1}(hu)$. Thus, we may write

$$R(\tilde{P}, \mathbf{c}) = \int_0^1 \tilde{F}_{\mathbf{c}}^{-1}(u) du \geq R(\tilde{P}_{\mathbf{c}}, \mathbf{c}).$$

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

8 Technical proofs for Section 3

8.1 Proofs for Example 17

Example (17). *Let* $\phi_1 = \|.\|^2$, $\phi_2 = \exp(-.)$, $\Omega = \mathbb{R}$, $P = (1-p)\delta_{-1} + p\delta_1$, *with* $p \leq 1/2$. *Then, for* $\phi = \phi_j$, $j \in \{1,2\}$, $k = 2$ *and* $h > (1-p)$, *we have* $B_h = \frac{h+p-1}{p} \wedge (1-h)$. *Let* $Q_{\gamma,N} = (1-\gamma)P + \gamma\delta_N$. *The following holds.*

- *If* $(1+p)h > 1$, $B_h = 1-h$, *and for every* $\gamma > 1-h$, *any sequence of optimal* 2*-points* h*-trimmed codebook* $\mathbf{c}_2^*(Q_{\gamma,N})$ *for* $Q_{\gamma,N}$ *satisfies*

$$\lim_{N\to+\infty} \|\mathbf{c}_2^*(Q_{\gamma,N})\| = +\infty.$$

- *If* $(1+p)h \leq 1$, *then* $B_h = \frac{h+p-1}{p}$, *and, for* $\gamma = B_h$, $(-1, N)$ *is an optimal* 2*-points* h*-trimmed codebook for* $Q_{\gamma,N}$.

We have, for P and any $s \in [0,1]$, $R_{2,s}^* = 0$, and $R_{1,s}^* > 0$ if and only if $s > (1-p)$. Thus, for any $h > (1-p)$ and $b \leq (1-h)$, $R_{1,h_b^-}^* > R_{2,h_b^+}^*$ if and only if $h_b^- > (1-p)$ that is equivalent to $b < (h+p-1)/p$. We deduce that $B_h = \frac{h+p-1}{p} \wedge (1-h)$.

Assume that $(h+p-1)/p \leq (1-h)$. Then $B_h = (h+p-1)/p$. For $\gamma = B_h$ and $\mathbf{c} = (-1, N)$, we have that $Q_{\gamma,N}(\{-1, N\}) = (1-\gamma)(1-p) + \gamma = h$, hence $R_{h,\gamma,N}(\mathbf{c}) = 0$, where $R_{h,\gamma,N}$ denotes the h-trimmed distortion with respect to $Q_{\gamma,N}$.

Now assume that $(h+p-1)/p > (1-h)$. Let $\gamma > (1-h)$. Then, for any $\mathbf{c} \in \mathbb{R}^{(2)}$, if $\tau_h(\mathbf{c})$ is such that $Q_{\gamma,N}\tau_h(\mathbf{c})$ is a submeasure of $Q_{\gamma,N}$ with total mass h, then $Q_{\gamma,N}\tau_h(\mathbf{c})(u)\mathbb{1}_{\{N\}}(u) \geq \gamma - (1-h)$. Now, if $\mathbf{c}_2^*(Q_{\gamma,N})$ is an optimal two-points quantizer for $Q_{\gamma,N}$, Proposition 9 ensures that for every $j \in \{1,2\}$,

$$c_{2,j}^*(Q_{\gamma,N}) = \frac{Pu\tau_h(\mathbf{c}_2^*(Q_{\gamma,N}))(u)\mathbb{1}_{W_j(\mathbf{c}_2^*(Q_{\gamma,N}))}(u)}{P\tau_h(\mathbf{c}_2^*(Q_{\gamma,N}))(u)\mathbb{1}_{W_j(\mathbf{c}_2^*(Q_{\gamma,N}))}(u)}.$$

Thus, we may assume that $Q_{\gamma,N}\tau_h(\mathbf{c}_2^*(Q_{\gamma,N}))(u)\mathbb{1}_{W_2(\mathbf{c}_2^*(Q_{\gamma,N}))}(u)\mathbb{1}_{\{N\}}(u) \geq \gamma - (1-h)$, without loss of generality. Hence, for N large enough,

$$c_{2,2}^*(Q_{\gamma,N}) \geq -(1-p) + N(\gamma - (1-h)) \underset{N\to\infty}{\longrightarrow} +\infty.$$

□

9 Technical proofs for Section 5

To ease the understanding the statement of the results are recalled before their proofs.

9.1 Proof of Lemma 20

Lemma (20)**.** *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*

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

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* $s \leq h$, *then* $r_s(\mathbf{c}) \leq r^+$.

Proof of Lemma 20. Let K_+ be such that $P(B(0,K_+)) > h$. Thus, if $c \in \bar{B}(0,K)$, $P(B(c,K+K_+)) > h$. 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, for all $x, y \in B(c,K+K_+) \cap F_0$, $d_\phi(x,y) \leq C_+\|x-y\|$. Therefore, for every $c \in \bar{B}(0,K)$, $P\left(B_\phi\left(c, \sqrt{C_+(2K+K_+)}\right)\right) > h$. Hence $r_s(c) \leq r_h(c) \leq \sqrt{C_+(2K+K_+)} = r^+$.

At last, if $\mathbf{c}$ is such that $c_{j_0} \in \bar{B}(0,K)\cap F_0$, then $\bar{B}_\phi(c_{j_0}, r_h(c_{j_0})) \subset \bar{B}_\phi(\mathbf{c}, r^+)$. Therefore $P(\bar{B}_\phi(\mathbf{c}, r^+)) \geq s$ for every $s \leq h$, hence $r_s(\mathbf{c}) \leq r^+$. □

9.2 Proof of Lemma 21

Lemma (21). *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 $\bar{B}(0,K) \cap F_0$, and $x \in \Omega$,*

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

where $D(\mathbf{c},\mathbf{c}') = \min_{\sigma\in\Sigma_k} \max_{j\in[\![1,k]\!]} |c_j - c'_{\sigma(j)}|$ (cf. Theorem 10).

Proof of Lemma 21. The set $F_0 \cap \bar{B}(0,K)$ is a convex compact subset of $\mathring{\Omega}$. Let $x \in \mathbb{R}^d$ and $\mathbf{c},\mathbf{c}' \in \big(F_0 \cap \bar{B}(0,K)\big)^{(k)}$. Since ϕ and $x \mapsto \nabla\phi(x)$ are $\mathcal{C}^1$, the mean value theorem yields that for every $j \in [\![1,k]\!]$,

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

for some constant C_K. Thus,

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

□

9.3 Proof of Lemma 22

Lemma (22). *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 $s_0 < h$ such that*

$$\forall s_0 < s < h, \quad \sup_{\mathbf{c}\in\left(F_0\cap\bar{B}(0,K)\right)^{(k)}} R_h(\mathbf{c}) - R_s(\mathbf{c}) \leq \epsilon.$$

Proof of Lemma 22. According to Lemma 7.1 and Lemma 21, for every $h \in (0,1)$, $\mathbf{c},\mathbf{c}' \in \big(F_0 \cap \bar{B}(0,K)\big)^{(k)}$ and $\tilde{P}_{\mathbf{c}'} \in \mathcal{P}_h(\mathbf{c}')$,

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

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, let $s < h$, and let α_s and α_h be such that $\frac{1}{h}(P\mathbb{1}_{B_\phi(\mathbf{c},r_h(\mathbf{c}))} + \delta_h P\mathbb{1}_{\partial B_\phi(\mathbf{c},r_h(\mathbf{c}))}) \in \mathcal{P}_h(\mathbf{c})$ (resp. $\frac{1}{s}(P\mathbb{1}_{B_\phi(\mathbf{c},r_s(\mathbf{c}))} + \delta_s P\mathbb{1}_{\partial B_\phi(\mathbf{c},r_s(\mathbf{c}))}) \in \mathcal{P}_s(\mathbf{c})$). Then

$$\begin{aligned}R_h(\mathbf{c}) - R_s(\mathbf{c}) =\ & Pd_\phi(u,\mathbf{c})\left(\mathbb{1}_{B_\phi(\mathbf{c},r_h(\mathbf{c}))}(u) + \delta_h\mathbb{1}_{\partial B_\phi(\mathbf{c},r_h(\mathbf{c}))}(u)\right) \\ & - Pd_\phi(u,\mathbf{c})\left(\mathbb{1}_{B_\phi(\mathbf{c},r_s(\mathbf{c}))}(u) + \delta_s\mathbb{1}_{\partial B_\phi(\mathbf{c},r_s(\mathbf{c}))}(u)\right) \\ \leq\ & r_h^2(\mathbf{c})\,(h-s)\,.\end{aligned}$$

Moreover, according to Lemma 20, $\sup_{\mathbf{c}\in\left(F_0\cap\bar{B}(0,K)\right)^{(k)}} r_h(\mathbf{c}) \leq r^+$ for some $r^+ < \infty$, hence the result. □

9.4 Proof of Lemma 23

Throughout this section, for any $\mathbf{c} \in \Omega^{(k)}$ and $s \in]0,1]$, we denote by $\tau_s(\mathbf{c})$ a map in $[0,1]$ such that $\frac{1}{s}P\tau_s(\mathbf{c}) \in \mathcal{P}_s(\mathbf{c})$, and by T_s the operator

$$T_s : \begin{cases}\Omega^{(k)} & \to & F_0^{(k)} \\ \mathbf{c} & \mapsto & \left(\frac{Pu\mathbb{1}_{W_j(\mathbf{c})}(u)\tau_s(\mathbf{c})(u)}{P\tau_s(\mathbf{c})(u)\mathbb{1}_{W_j(\mathbf{c})}(u)}\right)_{j\in[\![1,k]\!]}\end{cases},$$

with the convention $T_s(\mathbf{c})_j = c_j$ whenever $P\tau_s(\mathbf{c})\mathbb{1}_{W_j(\mathbf{c})} = 0$. The statement of Lemma 23 is recalled below.

Lemma (23). *Assume that the requirements of Theorem 8 are satisfied. For every $k \geq 2$, if $R^*_{k-1,h} - R^*_{k,h} > 0$, then*

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

Moreover there exists $0 < h^- < h < h^+ < 1$ such that, for every $j \in [\![2,k]\!]$, $R^*_{j-1,h^-} - R^*_{j,h^+} \geq \frac{\alpha}{2}$.

For any $h \wedge (1-h) \geq b > 0$, denote by $h_b^- = (h-b)/(1-b)$, $h_b^+ = h/(1-b)$. Let b be such that $\min_{j \in [\![2,k]\!]} R^*_{j-1,h_b^-} - R^*_{j,h_b^+} > 0$, and, for κ_1 in $(0,1)$, denote by $b_1 = \kappa_1 b$. Then, for any $s \in [h_{b_1}^-, h_{b_1}^+]$ and $j \in [\![1,k]\!]$, there exists a minimizer $\mathbf{c}^*_{j,s}$ of $R_{j,s}$ satisfying

$$\forall p \in [\![1,j]\!], \quad \|c^*_{j,s,p}\| \leq \frac{P\|u\|}{b(1-h)(1-\kappa_1)}.$$

The proof of Lemma 23 proceeds from two intermediate results, Lemma 30 and Lemma 31 below. First, Lemma 30 ensures that there exists bounded optimal codebooks whenever $R^*_{k-1,h} - R^*_{k,h} > 0$.

Lemma 30. *For every $k \geq 2$, if $R^*_{k-1,h} - R^*_{k,h} > 0$, then*

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

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 $\forall p \in [\![1,j]\!]$, $\|c^*_{j,s,p}\| \leq C_{h^-,h^+}$ and $c^*_{j,s,p} \in F_0$.*

A proof of Lemma 30 is given in Section 9.5. The other intermediate result, Lemma 31, ensures that optimal codebooks cells have enough mass.

Lemma 31. *Assume that $R^*_{k-1,h} - R^*_{k,h} > 0$, and let b be such that $\min_{j \in [\![2,k]\!]} R^*_{j-1,h_b^-} - R^*_{j,h_b^+} > 0$. Let $\kappa_1 < 1$ and $b_1 = \kappa_1 b$. Then, for any $s \in [h_{b_1}^-, h_{b_1}^+]$ and $j \in [\![1,k]\!]$, if $\mathbf{c}^*_{j,s}$ is a minimizer of $R_{j,s}$, we have*

$$\forall p \in [\![1,j]\!] \quad P\tau_s(\mathbf{c}^*_{j,s}) \mathbb{1}_{W_p(\mathbf{c}^*_{j,s})} \geq b(1-h)(1-\kappa_1).$$

The proof of Lemma 31 is given in Section 9.6. Equipped with these two lemmas, we are in position to prove Lemma 23.

Proof of Lemma 23. Assume that $R^*_{k-1,h} - R^*_{k,h} > 0$, then Lemma 30 entails that there exists b such that $\min_{j \in [\![2,k]\!]} R^*_{j-1,h_b^-} - R^*_{j,h_b^+} > 0$. Moreover, for any $s \in [h_{b_1}^-, h_{b_1}^+]$, and $j \in [\![1,k]\!]$, Lemma 30 provides a minimizer $\mathbf{c}^*_{j,s}$ of $R_{j,s}$. According to Proposition 9, $T_s(\mathbf{c}^*_{j,s})$ is a $R_{j,s}$-minimizer. According to Lemma 31, it satisfies, for all $p \in [\![1,j]\!]$,

$$\|(T_s(\mathbf{c}^*_{j,s}))_p\| \leq \frac{P\|u\|}{b(1-h)(1-\kappa_1)}.$$

□

9.5 Proof of Lemma 30

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 the restriction of P to the set A, $P_{|A}$, 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 $\left(\bar{B}(0, C_{h^-,h^+}) \cap F_0\right)^{(k)}$ is a compact set, the limit in $\left(\bar{B}(0, C_{h^-,h^+}) \cap F_0\right)^{(k)}$ 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 codebook c, $\|T_s(c)\| \leq P\|u\|/s$, optimal 1-point codebooks can be found in $\bar{B}(0, C_1) \cap F_0$, with $C_1 = \frac{P\|u\|}{s}$. 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 22, 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,

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

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 22, 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

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

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

Now, if $\mathbf{c} = (c_1, c_2)$ is an $\alpha/4$-minimizer of $R^*_{2,s}$, for $h^+ \geq s \geq h - (h-h_2)/2 =: h^-$, it holds $P\tau_s(\mathbf{c})\mathbb{1}_{W_j(\mathbf{c})} \geq h - h^-$, for $j \in \{1,2\}$. Indeed, suppose that $P\tau_s(\mathbf{c})\mathbb{1}_{W_1(\mathbf{c})} < h - h^-$. Then

$$R^*_{2,h^+} \geq R_s(\mathbf{c}) - \frac{\alpha}{4} \geq P d_\phi(u, c_2)\tau_s(\mathbf{c})(u)\mathbb{1}_{W_2(\mathbf{c})}(u) - \frac{\alpha}{4} \geq R^*_{1,h_2} - \frac{\alpha}{4},$$

since $s - h + h^- \geq h_2$, hence the contradiction. Choosing $h^+ \geq s \geq h^-$ entails that $\|T_s(\mathbf{c}^*_{j,s})_p\| \leq \frac{P\|u\|}{h-h^-}$ for every $p \in \{1,2\}$, 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 $\alpha = \min_{j \in [\![2,k]\!]} R^*_{j-1,h} - R^*_{j,h} > 0$. Let h^{--} and h^{++} be 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 $\frac{\alpha}{8}$-minimizer of $R^*_{k,h}$. According to Lemma 22, 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 22 provides $h > h_1 > h^{--}$ such that

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

Then, according to step $k-1$, since $\|(\mathbf{c}^*_{k-1,h_1})_j\| \leq C_{h^{--},h^{++}}$ for $j \in [\![1,k]\!]$, 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 an $\frac{\alpha}{4}$-minimizer of $R^*_{k,s}$, for $h^+ \geq s \geq h^- = \frac{h+h_1}{2}$, and assume that $P\tau_s(\mathbf{c})\mathbb{1}_{W_1(\mathbf{c})} < h - h^-$. Then

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

Then, $R^*_{k,h^+} \geq R^*_{k-1,h_1} - \frac{\alpha}{4}$, hence the contradiction. Thus, for such a choice of h^- and $h^- \leq s \leq h^+$, $P\tau_s(\mathbf{c})\mathbb{1}_{W_p(\mathbf{c})} \geq h - h^-$, which entails $\|(T_s(\mathbf{c}))_p\| \leq P\|u\|/(h - h^-)$, for every $p \in [\![1,k]\!]$. □

9.6 Proof of Lemma 31

Let $s \in [h_{b_1}^-, h_{b_1}^+]$, $j \in [\![1,k]\!]$ and $\mathbf{c}_{j,s}^*$ be a $R_{j,s}$ minimizer. If $j=1$, then $P\tau_s(\mathbf{c}_{j,s}^*) = s \geq h_{b_1}^- \geq b(1-\kappa_1)(1-h)$. Now assume that $j \geq 2$, and, without loss of generality, that $P\tau_s(\mathbf{c}_{j,s}^*)\mathbb{1}_{W_1(\mathbf{c}_{j,s}^*)} < b(1-\kappa_1)(1-h) < h_{b_1}^- - h_b^-$. We may write

$$\begin{aligned} R_{j,h_b^+}^* \geq R_{j,s}^* &\geq \sum_{p=2}^{j} Pd_\phi(u, c_{j,s,p}^*)\tau_s(\mathbf{c}_{j,s}^*)(u)\mathbb{1}_{W_p(\mathbf{c}_{j,s}^*)}(u) \\ &\geq R_{j-1,s-(h_{b_1}^- - h_b^-)}^* \geq R_{j-1,h_b^-}^*, \end{aligned}$$

hence the contradiction. □

10 Technical proofs for Section 6

10.1 Proof of Proposition 24

Proposition (24). *With probability larger than $1-e^{-x}$, we have, for $k \geq 2$,*

$$\begin{aligned} \sup_{\mathbf{c}\in(\mathbb{R}^d)^{(k)}, r\geq 0} \left|(P-P_n)\mathbb{1}_{B_\phi(\mathbf{c},r)}\right| &\leq C\sqrt{\frac{k(d+1)\log(k)}{n}} + \sqrt{\frac{2x}{n}}, \\ \sup_{\mathbf{c}\in(\mathbb{R}^d)^{(k)}, r\geq 0} \left|(P-P_n)\mathbb{1}_{\partial B_\phi(\mathbf{c},r)}\right| &\leq C\sqrt{\frac{k(d+1)\log(k)}{n}} + \sqrt{\frac{2x}{n}}, \end{aligned} \tag{6}$$

where $\partial B_\phi(\mathbf{c},r)$ denotes $\left\{x \mid d_\phi(x,\mathbf{c}) = r^2\right\}$ and C denotes a universal constant. Moreover, if r^+ and K are fixed, and $P\|u\|^2 \leq M_2 < \infty$, we have, with probability larger than $1-e^{-x}$,

$$\begin{aligned} \sup_{\mathbf{c}\in\left(\bar{B}(0,K)\cap F_0\right)^{(k)}, r\leq r^+} \left|(P-P_n)d_\phi(.,\mathbf{c})\mathbb{1}_{B_\phi(\mathbf{c},r)}\right| &\leq (r^+)^2\left[C_{K,r^+,M_2}\frac{\sqrt{kd\log(k)}}{\sqrt{n}} + \sqrt{\frac{2x}{n}}\right], \\ \sup_{\mathbf{c}\in\left(\bar{B}(0,K)\cap F_0\right)^{(k)}, r\leq r^+} \left|(P-P_n)d_\phi(.,\mathbf{c})\mathbb{1}_{\partial B_\phi(\mathbf{c},r)}\right| &\leq (r^+)^2\left[C_{K,r^+,M_2}\frac{\sqrt{kd\log(k)}}{\sqrt{n}} + \sqrt{\frac{2x}{n}}\right], \end{aligned} \tag{7}$$

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

The proof of Proposition 24 will make use of the following results. The first one deals with VC dimension of Bregman balls.

Lemma 32. *Let $\mathcal{C}$ (resp. $\bar{\mathcal{C}}$) denote the class of open (resp. closed) Bregman balls $B_\phi(x,r) = \{y \in \mathbb{R}^d \mid \sqrt{d_\phi(y,x)} < r\}$ (resp. $\bar{B}_\phi(x,r) = \{y \in \mathbb{R}^d \mid \sqrt{d_\phi(y,x)} < r\}$), $x \in \mathbb{R}^d$, $r \geq 0$. Then*

$$\begin{aligned} d_{VC}(\mathcal{C}) &\leq d+1, \\ d_{VC}(\bar{\mathcal{C}}) &\leq d+1, \end{aligned}$$

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

Proof of Lemma 32. 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

$$\begin{aligned} A_1 &= S \cap B_\phi(c_1,r_1) \cap B_\phi(c_2,r_2)^c \\ A_2 &= S \cap B_\phi(c_2,r_2) \cap B_\phi(c_1,r_1)^c, \end{aligned}$$

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

$$\ell_{1,2}(x) := d_\phi(x, c_1) - d_\phi(x, c_2) < 0.$$

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

Next, to bound expectation of suprema of empirical processes, we will need the following result. For any set of real-valued functions $\mathcal{F}$, let $\mathcal{N}(\mathcal{F}, \varepsilon, L_2(P_n))$ denote its ε covering number with respect to the $L_2(P_n)$ norm. In addition, the pseudo-dimension of $\mathcal{F}$, $d_{VC}(\mathcal{F})$, is defined as the Vapnik dimension of the sets $\{(x,t) \mid f(x) \leq t\}$.

Theorem 33. *[33, Theorem 1] If $\mathcal{F}$ is a set of functions taking values in $[-1,1]$. Then, for all $\varepsilon \leq 1$*

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

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

We are now in a position to prove Proposition 24.

Proof of Proposition 24. Let $\mathbf{c} \in \mathbb{R}^{(k)}$. Since $B_\phi(\mathbf{c}, r) = \bigcup_{j=1}^k B_\phi(c_j, r)$, according to [39, Theorem 1.1] and Lemma 32, we may write

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

where c_1 and c_2 are universal constant. Thus, applying [5, Theorem 3.4] gives the first inequality of (6). The second inequality of (6) follows from

$$\sup_{\mathbf{c}\in(\mathbb{R}^d)^{(k)}, r\geq 0} \left|(P - P_n)\mathbb{1}_{\partial B_\phi(\mathbf{c},r)}\right| \leq \sup_{\mathbf{c}\in(\mathbb{R}^d)^{(k)}, r\geq 0} \left|(P - P_n)\mathbb{1}_{B_\phi(\mathbf{c},r)}\right| + \sup_{\mathbf{c}\in(\mathbb{R}^d)^{(k)}, r\geq 0} \left|(P - P_n)\mathbb{1}_{\bar{B}_\phi(\mathbf{c},r)}\right|.$$

The inequalities of (7) are more involved. Denote by Z the left-hand side of the first inequality. A bounded difference inequality (see, e.g., [6, Theorem 6.2]) yields

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

It remains to bound

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

according to the symmetrization principle (see, e.g., [6, Lemma 11.4]), where for short $\mathbb{E}_Y$ denotes expectation with respect to the random variable Y. Let Γ_0 denote the set of functions $\left\{\frac{d_\phi(.,\mathbf{c})}{(r^+)^2}\mathbb{1}_{B_\phi(\mathbf{c},r)} \mid \mathbf{c} \in B(0,K)^{(k)}, r \leq r^+\right\}$. We have to assess the covering number $\mathcal{N}(\Gamma_0, \varepsilon, L_2(P_n))$. 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 $\Gamma_1 = \left\{ \frac{d_\phi(.,\mathbf{c})}{(r^+)^2} \wedge 1 \right\}$ and $\Gamma_2 = \left\{ \mathbb{1}_{B_\phi(\mathbf{c},r)} \right\}$. On one hand, we have

$$\begin{aligned}
\mathcal{N}(\Gamma_2, u, L_2(P_n)) &= \mathcal{N}(1-\Gamma_2, u, L_2(P_n)) \\
&= \mathcal{N}\left(\left\{ \prod_{j=1}^{k} \mathbb{1}_{B_\phi(c_j,r)^c} \mid \mathbf{c}, r \right\}, u, L_2(P_n) \right) \\
&\leq \mathcal{N}\left(\left\{ \mathbb{1}_{B_\phi(c,r)} \mid c \in \mathbb{R}^d, r \geq 0 \right\}, u/k, L_2(P_n) \right)^k \\
&\leq \left(\frac{2k}{u} \right)^{\kappa(d+1)k},
\end{aligned}$$

according to Theorem 33.

Now turn to Γ_1. According to Lemma 21, we may write

$$\begin{aligned}
\mathcal{N}(\Gamma_1, u, L_2(P_n)) &\leq \mathcal{N}\left(\left\{ \frac{d_\phi(.,\mathbf{c})}{(r^+)^2} \mid \mathbf{c} \in (F_0 \cap B(0,K))^{(k)} \right\}, u, L_2(P_n) \right) \\
&\leq \mathcal{N}\left(B(0,K)^k, \frac{(r^+)^2 u}{C_K(1+\|x\|_{L_2(P_n)})}, d_H \right).
\end{aligned}$$

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

$$\mathcal{N}(\Gamma_1, u, L_2(P_n)) \leq \left(\frac{3KC_K(1+\|x\|_{L_2(P_n)}))}{(r^+)^2 u} \right)^{kd},$$

hence

$$\mathcal{N}(\Gamma_0, \varepsilon, L_2(P_n)) \leq \left(\frac{6KC_K(1+\|x\|_{L_2(P_n)})}{(r^+)^2 \varepsilon} \right)^{kd} \times \left(\frac{4k}{\varepsilon} \right)^{\kappa(d+1)k}.$$

Using Dudley's entropy integral (see, e.g., [6, Corollary 13.2]) yields, for $k \geq 2$,

$$\begin{aligned}
&\mathbb{E}_\sigma \frac{1}{n} \sup_{\mathbf{c} \in (B(0,K) \cap F_0)^{(k)}, r \leq r^+} \sum_{i=1}^{n} \sigma_i \frac{d_\phi(X_i, \mathbf{c})}{(r^+)^2} \mathbb{1}_{B_\phi(\mathbf{c},r)}(X_i) \\
&\qquad \leq C \frac{(r^+)^2}{\sqrt{n}} \sqrt{kd \log\left(\frac{C_K(1+\|x\|_{L_2(P_n)})}{(r^+)^2} \right) + \kappa(d+1)k\log(4k)}.
\end{aligned}$$

Thus, applying Jensen's inequality leads to

$$\mathbb{E}_X \mathbb{E}_\sigma \frac{1}{n} \sup_{\mathbf{c} \in (B(0,K) \cap F_0)^{(k)}, r \leq r^+} \sum_{i=1}^{n} \sigma_i \frac{d_\phi(X_i, \mathbf{c})}{(r^+)^2} \mathbb{1}_{B_\phi(\mathbf{c},r)}(X_i) \leq (r^+)^2 C_{K,r^+,M_2} \sqrt{\frac{kd\log(k)}{n}}.$$

Replacing $\mathbb{1}_{B_\phi(\mathbf{c},r)}$ with $\mathbb{1}_{\partial B_\phi(\mathbf{c},r)}$ in the definition of Γ_0 and Γ_2 gives the same inequality, and (7). □

10.2 Proof of Proposition 25

Proposition (25). *Assume that $P\|u\|^p < +\infty$ for some $p \geq 2$, and let $b > 0$ be such that $\min_{j \in [\![2,k]\!]} R^*_{j-1,h_b^-} - R^*_{j,h_b^+} > 0$, where $h_b^- = (h-b)/(1-b)$, $h_b^+ = h/(1-b)$, as in Lemma 23. Let $\kappa_2 < 1$, and denote by $b_2 = \kappa_2 b$. Then there exists C_{P,h,k,κ_2} such that, for n large enough, with probability larger than $1 - n^{-\frac{p}{2}}$, we have, for all $j \in [\![2,k]\!]$, and $i \in [\![1,j]\!]$,*

$$\sup_{h_{b_2^-} \leq s \leq h} \|\hat{c}_{j,s,i}\| \leq C_{P,h,k,\kappa_2,b},$$

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

Proof of Proposition 25. For a codebook $\mathbf{c}$, let $\hat{\tau}_s(\mathbf{c})$ denote the trimming function $\mathbb{1}_{B_\phi(\mathbf{c},r_n(\mathbf{c}))}+\alpha_{n,s}(\mathbf{c})\mathbb{1}_{\partial B_\phi(\mathbf{c},r_n(\mathbf{c}))}$, so that $\frac{1}{s}P_n\hat{\tau}_s \in \hat{\mathcal{P}}_s(\mathbf{c})$, and let $\tau_s(\mathbf{c})$ denote the trimming function for the distribution P. Similarly to the proof of Theorem 8, we denote by $\hat{T}_s$ the operator that maps $\hat{\mathbf{c}}$ to the empirical means of its Bregman-Voronoi cells, that is

$$(\hat{T}_s(\mathbf{c}))_j = \frac{P_n u\hat{\tau}_s(\mathbf{c})\mathbb{1}_{W_j(\mathbf{c})}}{P_n\hat{\tau}_s(\mathbf{c})\mathbb{1}_{W_j(\mathbf{c})}}.$$

We let $b_2 = \kappa_2 b$, $\kappa_2 < 1$, and choose $b_1 = (\kappa_2 + \frac{1-\kappa_2}{2})b$ so that $b_2 < b_1 < b$. At last we denote by $\eta = \left[(h^-_{b_2} - h^-_{b_1}) \wedge (h^+_{b_1} - h^+_{b_2})\right]/2$, and will prove recursively that, for $j \in [\![1,k]\!]$ and $i \in [\![1,j]\!]$,

$$\sup_{h^-_{b_1}+\eta+\frac{j\eta}{k}\le s\le h} \|\hat{c}_{j,s,i}\| \le C_{P,h,k,\kappa_2,b},$$

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

Since $P\|u\|^p < +\infty$, Lemma 26 yields that $P_n\|u\| \le C_1$, for C_1 large enough, with probability larger than $1 - \frac{1}{8n^{\frac{p}{2}}}$. We set

$$C_{P,h,k,\kappa_2,b} = \frac{C_1}{h^-_{b_1} + \eta(1+1/k)} \vee C_{b_1,P,h} \vee \frac{2kC_1}{\eta},$$

where $C_{b_1,P,h}$ is given by Lemma 23. According to Proposition 24, for n large enough, we have that,

$$\begin{aligned} \sup_{\mathbf{c}\in(\mathbb{R}^d)^{(k)},r\ge 0} |(P-P_n)B_\phi(\mathbf{c},r)| &\le \frac{\eta}{4k} \\ \sup_{\mathbf{c}\in(\mathbb{R}^d)^{(k)},r\ge 0} |(P-P_n)\partial B_\phi(\mathbf{c},r)| &\le \frac{\eta}{4k}, \end{aligned} \tag{8}$$

with probability larger than $1 - \frac{1}{8n^{\frac{p}{2}}}$. On this probability event, from Lemma 20 and the fact that $s \mapsto r_s(\mathbf{c})$ is non-decreasing, we deduce that for some $r^+ > 0$,

$$\sup_{c\in B(0,C_{P,h,k,\kappa_2,b})\cap F_0, s\le h^+_{b_1}} r_{n,s}(c) \vee r_s(c) \le r^+.$$

We recall that since $P\|u\|^p < +\infty$, Lemma 26 yields that $P_n\|u\| \le C_1$, 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 24, we also have, with probability larger than $1 - \frac{1}{8n^{\frac{p}{2}}}$,

$$\begin{aligned} \sup_{\mathbf{c}\in(\bar{B}(0,C_{P,h,k,\kappa_2,b})\cap F_0)^{(k)},r\le r^+} \left|(P-P_n)d_\phi(u,\mathbf{c})\mathbb{1}_{B_\phi(\mathbf{c},r)}(u)\right| &\le \alpha_n \\ \sup_{\mathbf{c}\in(\bar{B}(0,C_{P,h,k,\kappa_2,b})\cap F_0)^{(k)},r\le r^+} \left|(P-P_n)d_\phi(u,\mathbf{c})\mathbb{1}_{\partial B_\phi(\mathbf{c},r)}(u)\right| &\le \alpha_n, \end{aligned} \tag{9}$$

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 has probability larger than $1 - \frac{1}{n^{\frac{p}{2}}}$. We proceed recursively on j.

For $j = 1$ and $h \ge s \ge h^-_{b_1} + \eta(1+1/k)$, according to Proposition 9, $\hat{T}_s(\hat{\mathbf{c}}_{1,s}) = \hat{\mathbf{c}}_{1,s}$, hence

$$\|\hat{\mathbf{c}}_{1,s}\| \le \frac{P_n\|u\|}{h^-_{b_1} + \eta(1+1/k)} \le \frac{C_1}{h^-_{b_1} + \eta(1+1/k)} \le C_{P,h,k,\kappa_2,b}.$$

Now assume that the statement of Proposition 25 holds up to order $j-1$. Let $\hat{\mathbf{c}}_{j,s}$ be a j-points empirically optimal codebook with trimming level $h \ge s \ge h^-_{b_1} + \eta(1+j/k)$. Assume that there exists one cell (say W_1) such that $P_n(\mathbb{1}_{W_1(\hat{\mathbf{c}}_{j,s})}\hat{\tau}_s(\hat{\mathbf{c}}_{j,s})) \le \frac{\eta}{k}$. On the one hand, we may write

$$\begin{aligned} \hat{R}_s(\hat{\mathbf{c}}_{j,s}) &\le \hat{R}_s(\mathbf{c}^*_{j,h^+_{b_1}}) \le P_n d_\phi(u,\mathbf{c}^*_{j,h^+})\hat{\tau}_s(\mathbf{c}^*_{j,h^+_{b_1}})(u) \\ &\le P_n d_\phi(u,\mathbf{c}^*_{j,h^+})\tau_{s+2\eta/k}(\mathbf{c}^*_{j,h^+_{b_1}})(u) \le R^*_{j,h^+_{b_1}} + \alpha_n, \end{aligned}$$

where $\mathbf{c}^*_{j,h^+_{b_1}}$ is a $R_{j,h^+_{b_1}}$ minimizer provided by Theorem 8.

On the other hand, we have

$$\begin{aligned}\hat{R}_s(\hat{\mathbf{c}}_{j,s}) &\geq \sum_{p=2}^{j} P_n d_\phi(u, \hat{c}_{j,s,p}) \mathbb{1}_{W_p(\hat{\mathbf{c}}_{j,s})} \hat{\tau}_{j,h^-_{b_1}+\eta(1+(j-1)/k)}(u) \\ &\geq \hat{R}_{h^-_{b_1}+\eta(1+(j-1)/k)}(\hat{\mathbf{c}}_{j-1,h^-_{b_1}+\eta(1+(j-1)/k)}).\end{aligned}$$

Thus,

$$\hat{R}_s(\hat{\mathbf{c}}_{j,s}) \geq P d_\phi(u, \hat{\mathbf{c}}_{j-1,h^-_{b_1}+\eta(1+(j-1)/k)}) \tau_{h^-_{b_1}+\eta(1+(j-1)/k)-\eta/2k}(\hat{\mathbf{c}}_{j-1,h^-_{b_1}+\eta(1+(j-1)/k)}) - \alpha_n,$$

according to the recursion assumption and (9). It comes

$$\hat{R}_s(\hat{\mathbf{c}}_{j,s}) \geq R^*_{j-1,h^-_{b_1}} - \alpha_n,$$

hence $R^*_{j-1,h^-_{b_1}} \leq R^*_{j,h^+_{b_1}} + 2\alpha_n$, that is impossible for n large enough. Therefore, for n large enough and every $p \in [\![1, j]\!]$,

$$P_n(\mathbb{1}_{W_p(\hat{\mathbf{c}}_{j,s})} \hat{\tau}_s(\hat{\mathbf{c}}_{j,s})) \geq \frac{\eta}{k}.$$

According to Proposition 9, equality $\hat{T}_s(\hat{\mathbf{c}}_{j,s}) = \hat{\mathbf{c}}_{j,s}$ holds and entails $\|\hat{c}_{j,s,p}\| \leq \frac{2kP_n\|u\|}{\eta} \leq C_{P,k,b,\kappa_2}$. □

10.3 Proof of Lemma 26

Lemma (26). *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 of Lemma 26. 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

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

From the Marcinkiewicz-Zygmund inequality applied to the real-valued centered random variables $Y_i = \|X_i\| - P\|u\|$ and the Minkowski inequality, it follows that

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{i=1}^n \left(\|X_i\| - P\|u\|\right)\right|^p\right] &= \mathbb{E}\left[\left|\sum_{i=1}^n Y_i\right|^p\right] \\ &\leq B_p \mathbb{E}\left[\left(\sum_{i=1}^n Y_i^2\right)^{\frac{p}{2}}\right] \\ &\leq B_p \left(\sum_{i=1}^n \left(\mathbb{E}|Y_i|^p\right)^{\frac{2}{p}}\right)^{\frac{p}{2}} \\ &= B_p n^{\frac{p}{2}} \mathbb{E}\left[|Y|^p\right] \\ &= B_p n^{\frac{p}{2}} P\left(|\|u\| - P\|u\||^p\right),\end{aligned}$$

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. □

10.4 Proof of Lemma 27

Lemma (27). *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 $n \in \mathbb{N}$,*

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

Proof of Lemma 27. Set $c \in \bar{B}(0,K) \cap F_0$. 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 $F_0 \cap \bar{B}(0,2K+K_+)$, the mean value theorem yields

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

for $C_+ = \sup_{c\in F_0\cap\bar{B}(0,2K+K_+)} \|\nabla_c\phi\| < +\infty$. Thus, it follows that

$$B(c,K+K_+) \subset B_\phi(c,r_+). \tag{10}$$

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 $r_{n,\phi,h}(\mathbf{c}) \leq r_+$. □

10.5 Proof of Lemma 28

Lemma (28). *Under the assumptions of Corollary 12, if $P\|u\|^q\psi^q(k\|u\|/h) < \infty$, then there exists a constant C_q such that $\mathbb{E}R_h^q(\hat{\mathbf{c}}_n) \leq C_P^q$.*

Let $\hat{\tau}_h(\hat{\mathbf{c}}_n)$ be such that $\frac{1}{h}P_n\hat{\tau}_h(\hat{\mathbf{c}}_n) \in \mathcal{P}_{n,h}(\hat{\mathbf{c}}_n)$, and $\hat{j}$ be such that $P_n(\hat{\tau}_h(\hat{\mathbf{c}}_n)\mathbb{1}_{W_{\hat{j}}(\hat{\mathbf{c}}_n)}) \geq \frac{nh}{k}$. According to the mean-value theorem and since $q \geq 1$ we may write

$$\begin{aligned} \mathbb{E}\left(R_h(\hat{\mathbf{c}}_n)\right)^q &\leq \mathbb{E}Pd_\phi^q(u,\hat{c}_{n,\hat{j}}) \\ &\leq \mathbb{E}P3^{q-1}\left[\phi^q(u) + \phi(\hat{c}_{n,\hat{j}})^q + \psi^q(\|\hat{c}_{n,\hat{j}}\|)\|u-\hat{c}_{n,\hat{j}}\|^q\right] \\ &\leq \mathbb{E}P3^{q-1}\left[\phi^q(u) + \phi(\hat{c}_{n,\hat{j}})^q + \psi^q(\|\hat{c}_{n,\hat{j}}\|)2^{q-1}\left(\|u\|^q + \|\hat{c}_{n,\hat{j}}\|^q\right)\right] \\ &\leq 3^{q-1}P\|u\|^q\psi^q(u) + 3^{q-1}(1+2^{q-1})\mathbb{E}\|\hat{c}_{n,\hat{j}}\|^q\psi^q(\|\hat{c}_{n,\hat{j}}\|) \\ &\qquad + 6^{q-1}P\|u\|^q\mathbb{E}\psi^q(\|\hat{c}_{n,\hat{j}}\|). \end{aligned}$$

Since $\psi(t) \leq \psi\left(\frac{kt}{h}\right)$, the first term is bounded. Also, note that since $p \geq 2$, $q \leq 2 \leq p$ so that $P\|u\|^q < \infty$. Next, since $\hat{\mathbf{c}}_n$ satisfies the centroid condition, we have

$$\|\hat{\mathbf{c}}_{n,\hat{j}}\| \leq \frac{P_nu\hat{\tau}_h(\hat{\mathbf{c}}_n)(u)\mathbb{1}_{W_{\hat{j}}(\hat{\mathbf{c}}_n)}(u)}{P_n\hat{\tau}_h(\hat{\mathbf{c}}_n)(u)\mathbb{1}_{W_{\hat{j}}(\hat{\mathbf{c}}_n)}(u)} \leq \frac{k}{nh}\sum_{i=1}^n \|X_i\|.$$

Since $u \mapsto \|u\|^q\psi^q(u)$ is convex we may write

$$\begin{aligned} \mathbb{E}\|\hat{c}_{n,\hat{j}}\|^q\psi^q(\|\hat{c}_{n,\hat{j}}\|) &\leq \mathbb{E}\left[\left(\frac{k}{nh}\sum_{i=1}^n\|X_i\|\right)^q\psi^q\left(\frac{k}{nh}\sum_{i=1}^n\|X_i\|\right)\right] \\ &\leq \left(\frac{k}{h}\right)^q P\left(\|u\|^q\psi^q(k\|u\|/h)\right) < \infty. \end{aligned}$$

At last, note that

$$\begin{aligned} P\psi^q(k\|u\|/h) &\leq P\left(\left(\|u\|^q \vee 1\right)\psi^q(k\|u\|/h)\right) \\ &\leq P\left(\|u\|^q\psi^q(k\|u\|/h)\mathbb{1}_{\|u\|>1}\right) + P\left(\psi^q(k\|u\|/h)\mathbb{1}_{\|u\|\leq 1}\right) \\ &\leq P\|u\|^q\psi^q(k\|u\|/h) + \psi^q(k/h) < \infty, \end{aligned}$$

so that, using convexity of ψ,

$$\begin{aligned} \mathbb{E}\psi^q(\|\hat{c}_{n,\hat{j}}\|) &\leq \mathbb{E}\left[\psi^q\left(\frac{k}{nh}\sum_{i=1}^n \|X_i\|\right)\right] \\ &\leq P\psi^q(k\|u\|/h) < \infty. \end{aligned}$$

Combinining all pieces entails that $\mathbb{E}(R_h(\hat{\mathbf{c}}_n)) < \infty$. □

10.6 Proof of Lemma 29

Lemma (29). *Assume that $B_h > 0$ (see Definition 13), let $b < B_h$ and $b < b_1 < B_h$ such that $b = \kappa_1 b_1$, with $\kappa_1 < 1$. Denote by $\beta_1 = (1-\kappa_1)b_1\left[h \wedge (1-h)\right]/2$. Assume that $s/(n+s) \leq b$. Then, for n large enough, with probability larger than $1 - n^{-\frac{p}{2}}$, we have, for all $j \in [\![1, k]\!]$,*

$$P_n\left(\hat{\tau}_{h_b^-}(\hat{\mathbf{c}}_{n+s,h})\mathbb{1}_{W_j(\hat{\mathbf{c}}_{n+s,h})}\right) \geq \beta_1.$$

Proof of Lemma 29. As in the proof of Proposition 25, we assume that

$$\sup_{\mathbf{c}\in(\mathbb{R}^d)^{(k)}, r\geq 0} |(P-P_n)B_\phi(\mathbf{c},r)| \vee |(P-P_n)\partial B_\phi(\mathbf{c},r)| \leq \beta_n \leq \beta_1.$$

According to Proposition 24, for n large enough, this occurs with $\beta_n = O\left(\sqrt{\log(n)/n}\right)$, and with probability larger than $1 - \frac{1}{8n^{\frac{p}{2}}}$. On this probability event, we deduce as well that $\sup_{c\in B(0,C_P)\cap F_0, s\leq h_{b_1}^+} r_{n,s}(c) \vee r_s(c) \leq r^+$, for some $r^+ > 0$. We also assume that $P_n\|u\| \leq C_1$, for C_1 large enough, and

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

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 $1 - \frac{1}{n^{\frac{p}{2}}}$, according to Proposition 24 and Lemma 26. We let $\alpha_1 > 0$ be such that $\min_{j\in[\![2,k]\!]} R^*_{k-1,h_{b_1}^-} - R^*_{k,h_{b_1}^+} \geq \alpha_1$, according to Lemma 23, and let $b < B_h$ such that $s/(n+s) \leq b = \kappa_1 b_1$. Let $\hat{\mathbf{c}}_{n+s,h}$ denote an h-trimmed empirical risk minimizer based on $\{X_1, \ldots, X_n, x_{n+1}, \ldots, x_{n+s}\}$, and $\mathbf{c}^*_{h_{b_1}^+}$ a $h_{b_1}^+$-trimmed optimal codebook. Then

$$\hat{R}_{n+s,h}(\hat{\mathbf{c}}_{n+s,h}) \leq \hat{R}_{n+s,h}(\mathbf{c}^*_{h_{b_1}^+}) \leq \frac{1}{n+s}\left[\sum_{i=1}^n d_\phi(X_i, \mathbf{c}^*_{h_{b_1}^+})\hat{\tau}_{h_b^+}(\mathbf{c}^*_{h_{b_1}^+})(X_i)\right],$$

since $(n+s)h \le nh_b^+ < nh_{b_1}^+ \le n$. We may write

$$\begin{aligned}
\hat{R}_{n+s,h}(\hat{\mathbf{c}}_{n+s,h}) &\le \frac{n}{n+s}\left(P_n d_\phi(u, \mathbf{c}^*_{h_{b_1}^+})\tau_{h_b^+ + \beta_n}(\mathbf{c}^*_{h_{b_1}^+})(u)\right) \\
&\le \frac{n}{n+s}\left(P_n d_\phi(u, \mathbf{c}^*_{h_{b_1}^+})\tau_{h_{b_1}^+}(\mathbf{c}^*_{h_{b_1}^+})(u)\right) \\
&\le \frac{n}{n+s}\left(P d_\phi(u, \mathbf{c}^*_{h_{b_1}^+})\tau_{h_{b_1}^+}(\mathbf{c}^*_{h_{b_1}^+})(u) + \alpha_n\right) \\
&\le \frac{n}{n+s}\left(R^*_{h_{b_1}^+} + \alpha_n\right),
\end{aligned}$$

for n large enough. Now assume that $P_n\left(\hat{\tau}_{h_b^-}(\hat{\mathbf{c}}_{n+s,h})\mathbb{1}_{W_1(\hat{\mathbf{c}}_{n+s,h})}\right) < \beta_1$. Then,

$$\hat{R}_{n+s,h}(\hat{\mathbf{c}}_{n+s,h}) \ge \frac{n}{n+s}\hat{R}_{n,h_b^-}(\hat{\mathbf{c}}_{n+s,h}),$$

since $n-(n+s)(1-h) \ge n(1-h_b^-)$. Thus, removing one quantization point,

$$\begin{aligned}
\hat{R}_{n+s,h}(\hat{\mathbf{c}}_{n+s,h}) &\ge \frac{n}{n+s}P_n\left[d_\phi(u, \hat{\mathbf{c}}^{(k-1)}_{n,h_b^- - \beta_1})\hat{\tau}_{h_b^- - \beta_1}(\hat{\mathbf{c}}^{(k-1)}_{n,h_b^- - \beta_1})(u)\right] \\
&\ge \frac{n}{n+s}P_n\left[d_\phi(u, \hat{\mathbf{c}}^{(k-1)}_{n,h_b^- - \beta_1})\tau_{h_{b_1}^-}(\hat{\mathbf{c}}^{(k-1)}_{n,h_b^- - \beta_1})(u)\right],
\end{aligned}$$

where $\hat{\mathbf{c}}^{(k-1)}_{n,h_b^- - \beta_1}$ denotes a $h_b^- - \beta_1$-trimmed empirical risk minimizer with $k-1$ codepoints. Since $h_b^- - \beta_1 \ge h_b^- - 2\beta_1 \ge h_{b_1^-}$, Proposition 25 implies

$$\begin{aligned}
\hat{R}_{n+s,h}(\hat{\mathbf{c}}_{n+s,h}) &\ge \frac{n}{n+s}P_n\left[d_\phi(u, \hat{\mathbf{c}}^{(k-1)}_{n,h_b^- - \beta_1})\tau_{h_{b_1}^-}(\hat{\mathbf{c}}^{(k-1)}_{n,h_b^- - \beta_1})(u)\right] \\
&\ge \frac{n}{n+s}P\left[d_\phi(u, \hat{\mathbf{c}}^{(k-1)}_{n,h_b^- - \beta_1})\tau_{h_{b_1}^-}(\hat{\mathbf{c}}^{(k-1)}_{n,h_b^- - \beta_1})(u) - \alpha_n\right].
\end{aligned}$$

Thus, $\hat{R}_{n+s,h}(\hat{\mathbf{c}}_{n+s,h}) \ge n(R^*_{k-1,h_{b_1}^-} + \alpha_n)/(n+s)$ hence the contradiction for $2\alpha_n < \alpha_1$. □

11 Supplementary material for Section 4

11.1 Additional files for the comparison of Bregman clusterings for mixtures with noise

11.1.1 Details on the different clustering procedures

In Section 4.4, we compared our trimmed Bregman procedures with the following clustering schemes : trimmed k-median [9], `tclust` [22], single linkage, ToMATo [14] and `dbscan` [25]. Trimmed k-median denotes the k-median clustering trimmed afterwards. Actually, we keep the $q = 110$ points which l_1-norm to their center is the smallest. In order to compute the centers, we use the function `kGmedian` from the R package `Gmedian`, with parameters $gamma = 1$, $alpha = 0.75$ and $nstart = nstartkmeans = 20$. For tclust, we use the function `tclust` from the R package `tclust` with parameters $k = 3$ (number of clusters) and $alpha = 10/120$, the proportion of points to consider as outliers. We use the C++ ToMATo algorithm, available at https://geometrica.saclay.inria.fr/data/ToMATo/. We compute the inverse of the distance-to-measure function [12] with paramter $m0 = 10/120$ (that can be considered as a density) at every sample point, and keep the 110 highest valued points. We use the first parameter 5 (the radius for the Rips graph built from the resulting sample points) and the second parameter 0.01 (related to the number of clusters). For the single linkage method, we first keep the 110 points with the smallest distance to their 10th nearest

neighbor, then, cluster points according to the R functions `hclust` with the method "single" and `cutree` with parameter $h = 4$ (related to the number of clusters). For the dbscan method, we use the `dbscan` function from the R package `dbscan`. We set the parameters *eps* to the 110-th smallest distance to a third nearest neighbor among points in the sample, $minPts = 3$ and $borderPoints = FALSE$.

For these three last methods, we cannot calibrate the parameters so that the algorithms return 3 clusters, because of the systemmatic presence of many additional small clusters.

11.1.2 Clustering for 12000 sample points

This section exposes additional experimental results. We proceed exactly like in Section 4.4, but with samples made of 10000 signal points and 2000 noise points.

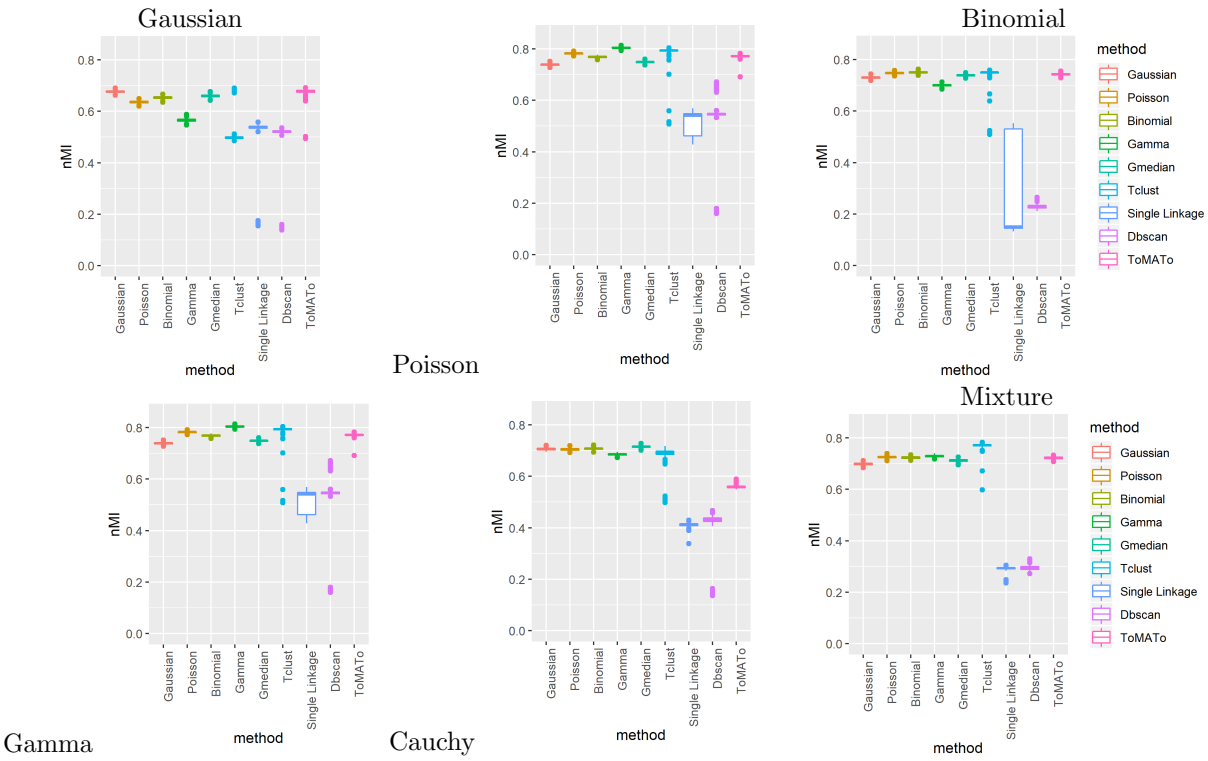


Figure 8: NMI for samples of 12000 points

For each clustering procedure, we decide to consider 11000 points as signal. The parameters for the different procedures are the same as described in Section 11.1.1, except for the ToMATo algorithm, we set the parameter $m0 = 200/12000$. As well, the number of nearest neighbors for the single linkage method is set to 200 and the parameter h is set to 0.8 for Gaussian distribution, 0.8 for Poisson, 1.1 for Binomial, 0.6 for Gamma, 0.4 for Cauchy and 0.4 for the mixture of 3 different distributions. For dbscan, *eps* is the 11000-th smallest distance of a point to its third nearest neighbor.

The NMI over 1000 replications of the experiments are represented via boxplots in Figure 8. Algorithm 1 with the proper Bregman divergence systematically (slightly) outperforms other clustering schemes.

11.2 Discussion about the choice of the Bregman divergence

We consider three mixtures of Gaussian distributions $\mathcal{L}(c, \sigma) = \frac{1}{3}\mathcal{N}(c_1, \sigma_1 I_2) + \frac{1}{3}\mathcal{N}(c_2, \sigma_2 I_2) + \frac{1}{3}\mathcal{N}(c_3, \sigma_3 I_2)$ with $c = (c_1, c_2, c_3)$ for $c_1 = (10, 10)$, $c_2 = (25, 25)$ and $c_3 = (40, 40)$, I_2 the identity matrix on $\mathbb{R}^2$ and

$\sigma = (\sigma_1, \sigma_2, \sigma_3)$. The first distribution $\mathcal{L}_1$ corresponds to clusters with the same variance, with $\sigma = (5, 5, 5)$, the second distribution $\mathcal{L}_2$ to clusters with increasing variance, with $\sigma = (1, 4, 7)$, and the third distribution $\mathcal{L}_3$ to clusters with increasing and decreasing variance, with $\sigma = (2, 7, 2)$. We cluster samples of 100 points from $\mathcal{L}_1$, $\mathcal{L}_2$ and then $\mathcal{L}_3$. We use Algorithm 1 with the Gaussian, Poisson, Binomial and Gamma Bregman divergences. Note that we set $N = 50$ for the Binomial divergence, so that we expect a clustering with clusters size symmetric with respect to 25. The performance of the clustering in terms of NMI is represented in Figure 9, after 1000 replications of the experiments.

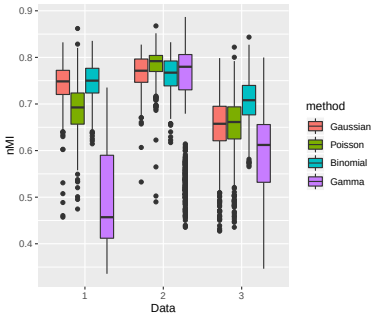
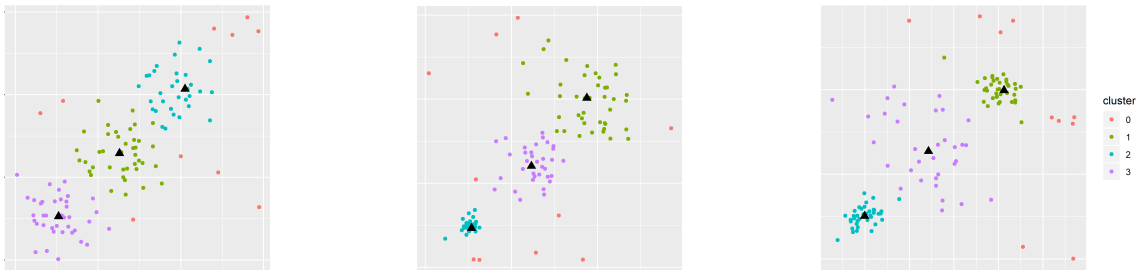


Figure 9: Comparison of Bregman divergences efficiency for different clusters variances.

The corresponding clustering with the best suited Bregman divergence is represented in Figure 10. In particular, we used the Gaussian divergence for $\mathcal{L}_1$, the Poisson divergence for $\mathcal{L}_2$ and the Binomial divergence for $\mathcal{L}_3$.



$\mathcal{L}_1$ – Gaussian divergence $\mathcal{L}_2$ – Poisson divergence $\mathcal{L}_3$ – Binomial divergence

Figure 10: Clustering with the best suited Bregman divergence

Since the sum of two Bregman divergences is a Bregman divergence, it is also possible to cluster data with Algorithm 1, with a different divergence on the different coordinates. For instance, we sample 100 points from $\frac{1}{3}\mathcal{N}(c_1, \Sigma_1) + \frac{1}{3}\mathcal{N}(c_2, \Sigma_2) + \frac{1}{3}\mathcal{N}(c_3, \Sigma_3)$, with for every $i \in [\![1, 3]\!]$, Σ_i diagonal with coefficients $(\sigma_i^{(2)}, \sigma_i^{(3)})$ with $\sigma^{(2)} = (1, 4, 7)$ and $\sigma^{(3)} = (2, 7, 2)$. In Figure 11, we represented the clustering obtained with Algorithm 1 with the Poisson divergence on the first coordinate and the Binomial divergence on the second coordinate. We observe that the shape of the clusters obtained correspond roughly to the shape of the sublevel sets of the sampling distribution.

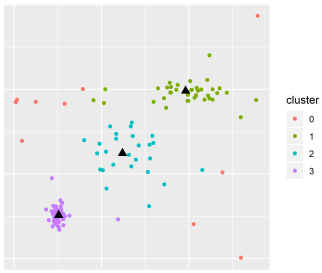


Figure 11: Clustering with hybrid Bregman divergence

11.3 Additional files for Stylometric author clustering

This section exposes the graphics and additional numerics that support several results from Section 4.6, for instance about the calibration of parameters.

Trimmed k-median:

In Figure 12 we plot the cost and the NMI as a function of q for different numbers of clusters k, in Figure 13 we focus on the case $k = 4$. Finally, in Figure 14 we plot the best clusterings in terms of NMI for $k = 4$ and $k = 6$.

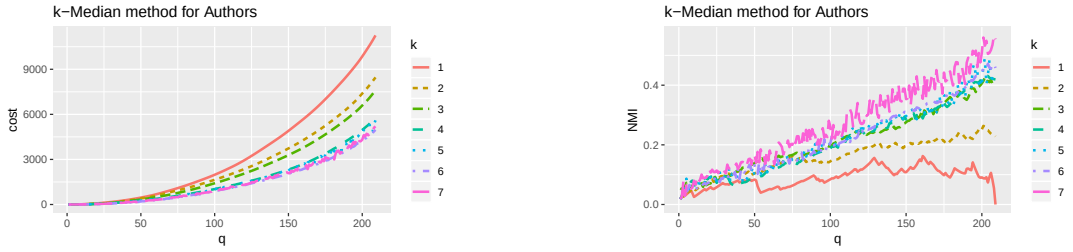


Figure 12: Cost and NMI for Author clustering with k-Median method

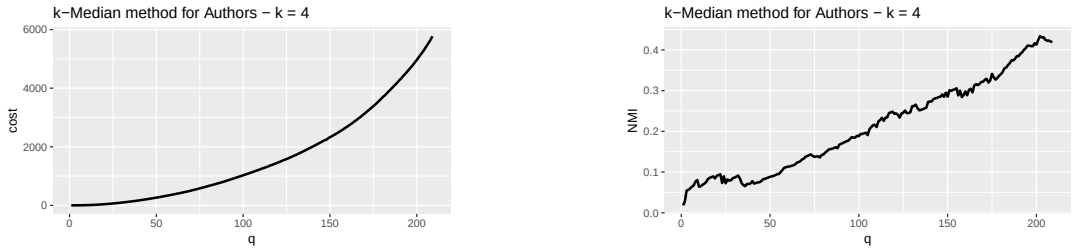


Figure 13: Cost and NMI for Author clustering with k-Median method – $k = 4$

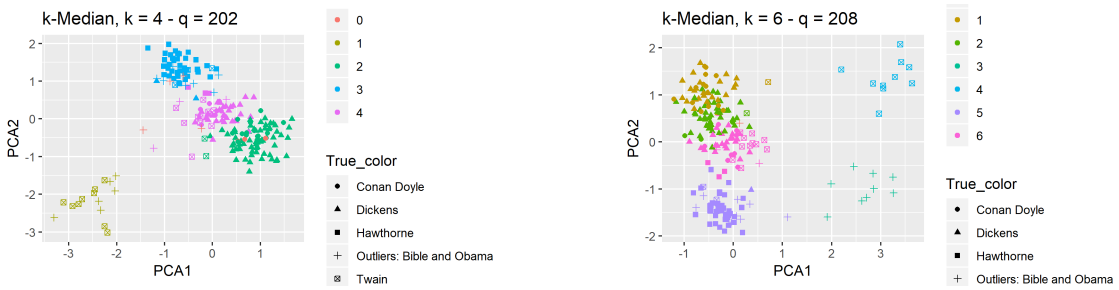


Figure 14: Examples of Author clusterings obtained with k-Median method

These graphics suggest that $k = 4$ and $k = 6$ are possible choices. The corresponding q that minimize NMI's are respectively $q = 202$ ($NMI = 0.4334372$), and $q = 208$ ($NMI = 0.4721967$).

tclust:

Figure 15 and 16 do not allow to select k. If $k = 4$ is chosen, Figure 16 suggests that $q \simeq 184$ is a relevant choice. Figure 17 provides the associated clustering, whose NMI is 0.4912537.

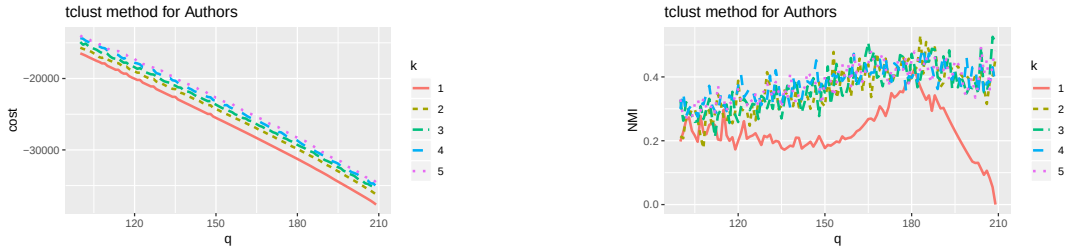


Figure 15: Cost and NMI for Author clustering with tclust algorithm

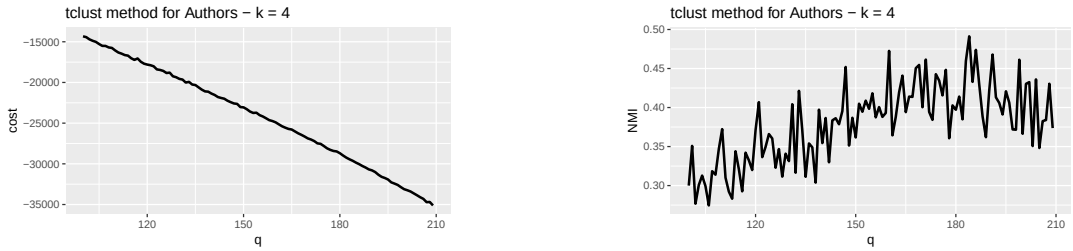


Figure 16: Cost and NMI for Author clustering with tclust algorithm – $k = 4$

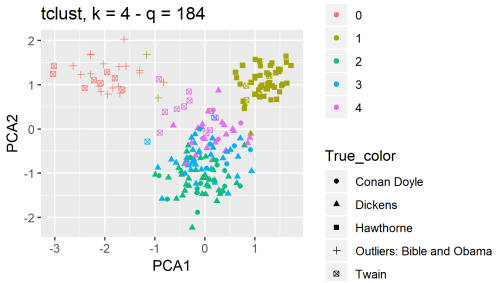


Figure 17: Examples of Author clusterings obtained with tclust algorithm

Trimmed k-means:

Figure 18 suggests the choice $k = 4$, and Figure 19 shows that $q = 190$ yields a slope jump and NMI peak. The associated clustering is depicted in Figure 20, its NMI is 0.5336308.

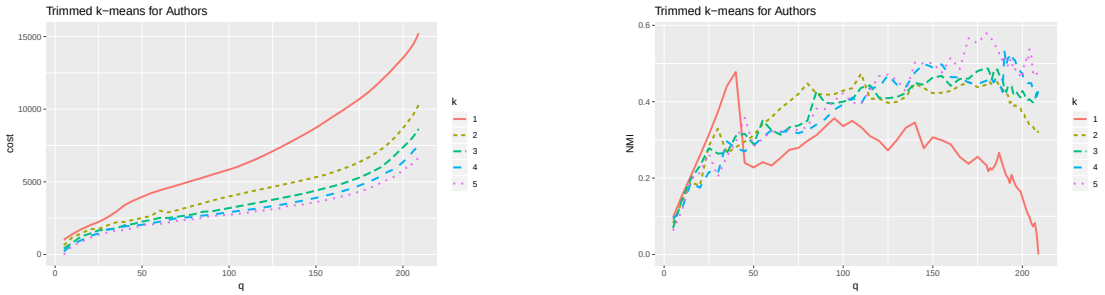


Figure 18: Cost and NMI for Author clustering with trimmed k-means

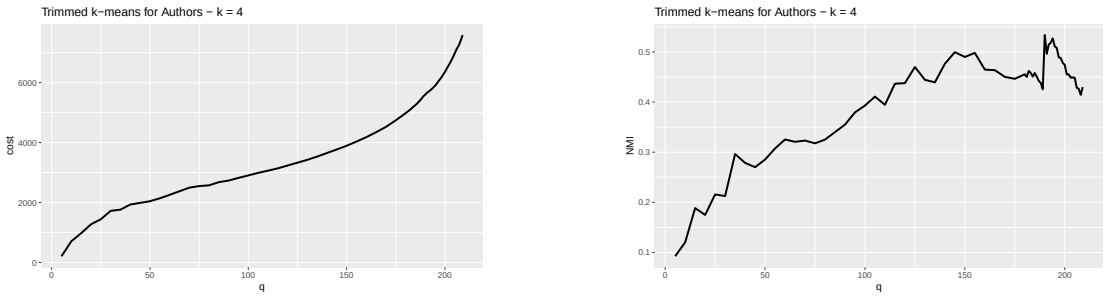


Figure 19: Cost and NMI for Author clustering with trimmed k-means – $k = 4$

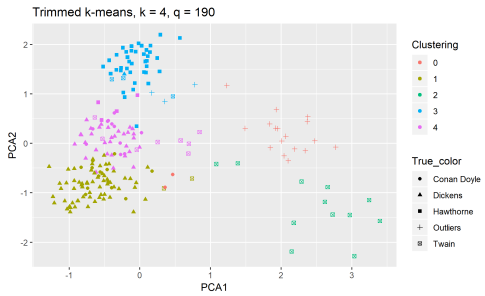


Figure 20: Examples of Author clusterings obtained with trimmed k-means