Universal Lower Bounds and Optimal Rates: Achieving Minimax Clustering Error in Sub-Exponential Mixture Models

Maximilien Dreveton MAXIMILIEN.DREVETON@EPFL.CH
École Polytechnique Fédérale de Lausanne (EPFL)

Alperen Gözeten ALPERENGOZETEN@GMAIL.COM
Bilkent University

Matthias Grossglauser MATTHIAS.GROSSGLAUSER@EPFL.CH
École Polytechnique Fédérale de Lausanne (EPFL)

Patrick Thiran PATRICK.THIRAN@EPFL.CH
École Polytechnique Fédérale de Lausanne (EPFL)

Editors: Shipra Agrawal and Aaron Roth

Abstract

Clustering is a pivotal challenge in unsupervised machine learning and is often investigated through the lens of mixture models. The optimal error rate for recovering cluster labels in Gaussian and sub-Gaussian mixture models involves *ad hoc* signal-to-noise ratios. Simple iterative algorithms, such as Lloyd's algorithm, attain this optimal error rate. In this paper, we first establish a universal lower bound for the error rate in clustering *any* mixture model, expressed through Chernoff information, a more versatile measure of model information than signal-to-noise ratios. We then demonstrate that iterative algorithms attain this lower bound in mixture models with sub-exponential tails, notably emphasizing location-scale mixtures featuring Laplace-distributed errors. Additionally, for datasets better modelled by Poisson or Negative Binomial mixtures, we study mixture models whose distributions belong to an exponential family. In such mixtures, we establish that Bregman hard clustering, a variant of Lloyd's algorithm employing a Bregman divergence, is rate optimal.

Keywords: clustering, mixture models, k-means, iterative algorithms

1. Introduction

Clustering is the task of partitioning a set of data points into groups, called clusters, such that data points within the same cluster are more similar to each other than they are to data points in different clusters. Clustering is an important problem in statistics and machine learning (Hastie et al., 2009; Wu et al., 2008), with many applications.

Mixture models provide an elegant framework for the design and theoretical analysis of clustering algorithms (McLachlan et al., 2019; Bouveyron and Brunet-Saumard, 2014). Denoting by $z^* \in [k]^n$ the vector of cluster assignments, a mixture model assumes that the n observed data points $X_1, \cdots, X_n \in \mathcal{X}^n$, where $\mathcal{X} \subset \mathbb{R}^d$, are independently generated such that

$$\forall i \in [n] \colon X_i \,|\, z_i^* \sim f_{z_i^*}, \tag{1.1}$$

where $f_1, \cdots, f_k$ are the k probability distributions over $\mathcal{X}$. An *estimator* $\hat{z} = \hat{z}_n$ of z^* is a measurable function $\hat{z} \colon (X_1, \cdots, X_n) \mapsto \hat{z}(X_1, \cdots, X_n) \in [k]^n$. The loss of an estimator $\hat{z}$ of z^* is quantified by the number of disagreements between $\hat{z}$ and z^*, up to a global permutation of $\hat{z}$, *i.e.,*

$$\operatorname{loss}(z^*, \hat{z}) \;=\; \min_{\tau \in \operatorname{Sym}(k)} \operatorname{Ham}(z^*, \tau \circ \hat{z}), \tag{1.2}$$

where $\mathrm{Sym}(k)$ denotes the group of permutations on $[k]$ and Ham the Hamming distance. The *expected relative error* made by an estimator $\hat{z} \colon (X_1, \cdots, X_n) \to [k]^n$ is then defined as

$$\mathbb{E}\left[n^{-1} \mathrm{loss}\left(z^*, \hat{z}\right) \right],$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to the model (1.1).

Gaussian mixture models are an important class of mixture models, in which for all $a \in [k]$, the distribution f_a is Gaussian with mean $\mu_a \in \mathbb{R}^p$ and covariance matrix Σ_a. If the Gaussian mixture is *isotropic*, that is, $\Sigma_a = \sigma^2 I_p$ for all $a \in [k]$ (with $\sigma^2 > 0$), finding the maximum likelihood estimator (MLE) of (z^*, μ) is equivalent to solving the *k-means* problem

$$\hat{z}, \hat{\mu} \ = \ \underset{\substack{z \in [k]^n \\ \tilde{\mu}_1, \cdots, \tilde{\mu}_p \in \mathbb{R}^p}}{\arg\min} \sum_{i=1}^n \sum_{a=1}^k \mathbb{1}\{z_i = a\} \| X_i - \tilde{\mu}_{z_a} \|_2^2, \tag{1.3}$$

where $\| \cdot \|_2$ denotes the Euclidean ℓ^2-norm. *Lloyd's algorithm* provides a simple way to find an approximate solution of this NP-hard minimisation problem iteratively (Lloyd, 1982). Under technical conditions on the initialisation and on the model parameters, Lu and Zhou (2016) show that the number of misclustered points by Lloyd's algorithm after $\Theta(\log n)$ iterations verify

$$\mathbb{E}\left[n^{-1} \mathrm{loss}\left(z^*, \hat{z}^{\mathrm{Lloyd}} \right) \right] \ \le \ e^{-(1+o(1)) \frac{1}{2} \mathrm{SNR}^2}, \tag{1.4}$$

where the signal-to-noise ratio (SNR) of this isotropic Gaussian mixture is defined as

$$\mathrm{SNR} \ = \ \frac{\min_{1 \le a \ne b \le k} \| \mu_a - \mu_b \|_2}{2\sigma}. \tag{1.5}$$

More generally, in a mixture of anisotropic Gaussians, where the probability distributions $f_1, \cdots, f_k$ are Gaussians with means $\mu_1, \cdots, \mu_k$ and share the same covariance matrix Σ, the MLE becomes

$$\hat{z}, \hat{\mu} \ = \ \underset{\substack{z \in [k]^n \\ \tilde{\mu}_1, \cdots, \tilde{\mu}_p \in \mathbb{R}^p}}{\arg\min} \sum_{i=1}^n \sum_{a=1}^k \mathbb{1}\{z_i = a\} \| X_i - \tilde{\mu}_{z_a} \|_\Sigma^2,$$

where $\|x - y\|_\Sigma^2 = (x-y)^T \Sigma^{-1} (x-y)$ denotes the square of the Mahalanobis distance. Because the formulation of the new MLE is similar to (1.3), it is natural to modify the Lloyd algorithm by (i) estimating the means and covariances in the estimation step and (ii) replacing the squared Euclidean distance by the Mahalanobis distance in the clustering step. The error of this new iterative algorithm is also upper-bounded as in (1.4), where the signal-to-noise ratio of this model is $\mathrm{SNR}_{\mathrm{anisotropic}} = 2^{-1} \min_{1 \le a \ne b \le k} \| \mu_a - \mu_b \|_\Sigma$ (Chen and Zhang, 2021). Moreover, for this anisotropic Gaussian mixture model, the upper bound (1.4) on the error attained by this iterative algorithm is tight and cannot be improved. More precisely, Chen and Zhang (2021) establish that

$$\inf_{\hat{z}} \sup_{z^* \in [k]^n} \mathbb{E}\left[n^{-1} \mathrm{loss}\left(z^*, \hat{z}\right) \right] \ \ge \ e^{-(1+o(1)) \frac{1}{2} \mathrm{SNR}^2_{\mathrm{anisotropic}}}, \tag{1.6}$$

where the $\inf$ is taken over all estimators $\hat{z}$. Combining (1.4) and (1.6) shows that a Lloyd-based scheme for solving the MLE in a mixture of anisotropic Gaussians achieves the minimax rate.

Error rates of iterative algorithms are now well understood for (sub)-Gaussian mixture models, but *what happens for the mixture of distributions with heavier tails than Gaussians?* As a first setting, we will study *location-scale mixture models*, for which the coordinates of each data point are generated as

$$\forall i \in [n], \, \forall \ell \in [d]: \quad X_{i\ell} \;=\; \mu_{z_i^* \ell} + \sigma_{z_i^* \ell} \epsilon_{i\ell}, \tag{1.7}$$

where $(\mu_a, \sigma_a) \in \mathbb{R}^d \times (0, \infty)^d$ denote the location and the scale of block $a \in [k]$, and the random variables $\epsilon_{i\ell}$ are independently sampled from a distribution with mean 0 and variance 1. *When the $\epsilon_{i\ell}$'s are not Gaussians, does the minimax error rate also relate to a signal-to-noise ratio?* Furthermore, parametric mixture models do not always involve location and scale parameters. For example, single-cell RNA sequencing datasets are represented by a matrix $X \in \mathbb{Z}^{n \times d}$ where n is the number of cells, d is the number of genes, and $X_{i\ell}$ records the number of unique molecular identifiers from the i-th cell that map to the ℓ-th gene. Cells can be of different types, and entries X_{ij} are often assumed to come from a negative binomial, whose parameters depend on the types of the cell i and of the gene ℓ (Grün et al., 2014; Haque et al., 2017). This motivates the study of negative binomial mixture models and, more generally, of mixture models of the form (1.1) where the pdfs $f_1, \cdots, f_k$ belong to an exponential family, but not necessarily Gaussian.

Our first contribution is the characterisation of a fundamental limit for the misclustering error. Denote by $\mathrm{Chernoff}(f, g)$ the *Chernoff information* between two probability distributions f and g. We establish that the classification error made by *any* clustering algorithm when applied to a mixture model defined in (1.1) is lower bounded as

$$\inf_{\hat{z}} \sup_{z^* \in [k]^n} \mathbb{E}\left[n^{-1} \mathrm{loss}\left(z^*, \hat{z}\right)\right] \;\geq\; e^{-(1+o(1)) \min\limits_{1 \leq a \neq b \leq k} \mathrm{Chernoff}(f_a, f_b)}. \tag{1.8}$$

This lower bound involves the Chernoff information instead of signal-to-noise ratios, but we show that these two quantities are related in many models of interest. In particular, for anisotropic Gaussian mixture models, the lower bounds (1.8) and (1.6) are the same. However, expressing the lower bound in terms of the Chernoff information instead of SNR is more versatile as it does not require making any assumption on the pdfs $f_1, \cdots, f_k$. The rationale for finding the Chernoff information in the lower bound (1.8) lies in the reformulation of the problem of assigning a data point X_i to a cluster $\hat{z}_i$ as an equivalent hypothesis testing problem, which tests the k different hypothesis $H_k \colon \hat{z}_i = \ell$ for $\ell \in [k]$. The difficulty of this problem is defined as the error made by the best estimator, which is asymptotically $\exp(-(1+o(1)) \min_{1 \leq a \neq b \leq k} \mathrm{Chernoff}(f_a, f_b))$.

Our second contribution is to show that iterative clustering algorithms can attain this error rate in sub-exponential mixture models. More precisely, we establish that an iterative algorithm achieves the lower bound (1.8) in the location-scale mixture model (1.7) when the $\epsilon_{i\ell}$'s are Laplace-distributed. An interesting example is when each dimension has the same variance, *i.e.,* $\sigma_{1\ell} = \cdots = \sigma_{k\ell} =: \sigma_\ell$. In such a model, the minimax error rate can be written as $\exp(-(1+o(1))\mathrm{SNR}_{\mathrm{Laplace}})$, where

$$\mathrm{SNR}_{\mathrm{Laplace}} \;=\; \min_{1 \leq a \neq b \leq k} \|\Sigma^{-1}(\mu_a - \mu_b)\|_1,$$

where Σ is the diagonal matrix whose diagonal elements are $\sigma_1, \cdots, \sigma_d$. Furthermore, for the mixture model defined in (1.1) whose pdfs belong to an exponential family, we show that the lower bound (1.8) is attained by a variant of Lloyd's algorithm that replaces the minimisation of the squared Euclidean distance by the minimization of a Bregman divergence. This algorithm is commonly called *Bregman hard clustering* in the literature, and the choice of the Bregman divergence depends on the exponential family considered (Banerjee et al., 2005).

Paper structure The paper is structured as follows. In Section 2, we establish a lower bound on the error rate made by any algorithm in clustering mixture models. We show in Section 3 that iterative algorithms attain this lower bound in various mixture models, such as the Laplace mixture model (Section 3.2) and the exponential family mixture models (Section 3.3). We discuss these results in Section 4 and compare them with the existing literature.

Notations The notation 1_n denotes the vector of size $n \times 1$ whose entries are all equal to one. For a vector x, we denote by $\|x\|_p$ its ℓ^p norm (with $1 \le p \le \infty$). The standard scalar product between two vectors x, y is denoted $< x, y >$. The indicator of an event A is denoted $\mathbb{1}\{A\}$. We abbreviate "random variable" by rv and "probability density function" by pdf. Laplace and Gaussian random variables with mean μ and scale σ are denoted by $\mathrm{Lap}(\mu, \sigma)$ and $\mathrm{Nor}(\mu, \sigma^2)$. Given a pdf f, we write $X \sim f$ if X is a rv whose pdf is f. A real-valued rv X is *sub-exponential* if there exists $C > 0$ such that for all $x \ge 0$ we have $\mathbb{P}\left(|X| \ge x\right) \le 2e^{-Ct}$. Finally, we use the standard Landau notations o and O, and write $a = \omega(b)$ when $b = o(a)$ and $a = \Omega(b)$ when $b = O(a)$. We also write $a = \Theta(b)$ when $a = O(b)$ and $b = O(a)$.

2. Minimax rate of the clustering error in mixture models

Let f and g be two pdfs with respect to a reference dominating measure ν. The *Chernoff information* between f and g is defined as

$$\mathrm{Chernoff}(f, g) \;=\; -\log \inf_{t \in (0,1)} \int f^t(x) g^{1-t}(x) d\nu(x).$$

For a family $\mathcal{F} = (f_1, \cdots, f_k)$ composed of k probability distributions, we define

$$\mathrm{Chernoff}(\mathcal{F}) \;=\; \min_{1 \le a \ne b \le k} \mathrm{Chernoff}\left(f_a, f_b\right). \tag{2.1}$$

The following theorem establishes an asymptotic lower bound on the clustering error in a mixture model composed of the distributions belonging to the family $\mathcal{F}$.

Theorem 1 *Consider the mixture model defined in (1.1) and let $\mathcal{F} = (f_1, \cdots, f_k)$ be the family of k probability distributions that comprise the mixture, where k and the distributions f_a scale with n. Suppose that* $\mathrm{Chernoff}(\mathcal{F}) = \omega(\log k)$. *Then,*

$$\inf_{\hat{z}} \sup_{z \in [k]^n} \mathbb{E}\left[n^{-1} \mathrm{loss}(z, \hat{z})\right] \;\ge\; e^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})},$$

where the inf *is taken over all estimators* $\hat{z} \colon (X_1, \cdots, X_n) \to [k]^n$.

The proof of Theorem 1 is given in Section A. The proof of Theorem 1 has two main steps. The first challenge is to address the minimum over all permutations in the definition of the error loss. Hence, rather than directly examining $\inf_{\hat{z}} \sup_{z \in [k]^n} \mathbb{E}\left[n^{-1} \mathrm{loss}(z, \hat{z})\right]$, we focus on a sub-problem

$\inf_{\hat{z}} \sup_{z\in\tilde{\mathcal{Z}}} \mathbb{E}\left[n^{-1}\mathrm{loss}(z,\hat{z})\right]$, where $\tilde{\mathcal{Z}} \subset [k]^n$ is chosen such that $\mathrm{loss}(z,\hat{z}) = \mathrm{Ham}(z,\hat{z})$ for all $z,\hat{z} \in \tilde{\mathcal{Z}}$. The idea is that this sub-problem is simple enough to analyze, but it still captures the hardness of the original clustering problem. Next, we bound the minimax risk of this sub-problem by the Bayes risk and we demonstrate that it is sufficient to lower-bound the testing error between each pair. The optimal error of this pairwise testing problem follows naturally from Lemma 2.

In Gaussian mixture models, the following example shows that the quantity $\mathrm{Chernoff}(\mathcal{F})$ is related to the more commonly used signal-to-noise ratios.

Example 1 *Suppose that $f_a = \mathrm{Nor}(\mu_a, \Sigma_a)$ where $\mu_1, \cdots, \mu_k \in \mathbb{R}^d$ and $\Sigma_1, \cdots, \Sigma_k$ are k-by-k positive definite matrices. Then,*

$$\mathrm{Chernoff}(\mathcal{F}) \;=\; \max_{a\neq b\in[k]} \sup_{t\in(0,1)} (1-t)\left\{\frac{t}{2}(\mu_a-\mu_b)^T\Sigma_{abt}^{-1}(\mu_a-\mu_b) - \frac{1}{2(t-1)}\log\frac{|t\Sigma_a+(1-t)\Sigma_b|}{|\Sigma_a|^{1-t}\cdot|\Sigma_b|^t}\right\}.$$

When $\Sigma_1 = \cdots = \Sigma_k$ the $\sup$ is achieved for $t = 2^{-1}$ and we obtain $\mathrm{Chernoff}(\mathcal{F}) = 2^{-1}\mathrm{SNR}^2_{\mathrm{anisotropic}}$ where $\mathrm{SNR}_{\mathrm{anisotropic}} = 2^{-1}\max_{a,b\in[k]}\|\Sigma^{-1/2}(\mu_a-\mu_b)\|_2$. This recovers the minimax lower bound for clustering Gaussian mixtures established in Chen and Zhang (2021).

Theorem 1 is closely related to hypothesis testing. Indeed, suppose that the probability densities $f_1, \cdots, f_k$ are known. By the Neyman-Pearson lemma, the optimal clustering $\hat{z}^{\mathrm{MLE}}$ verifies

$$\hat{z}_i^{\mathrm{MLE}} \;=\; \arg\max_{a\in[k]} f_a(X_i),$$

and $\hat{z}_i^{\mathrm{MLE}}$ is a function of X_i only, and not of the other data points $X_{-i} = (X_j)_{j\neq i}$.

Yet, hypothesis testing conventionally operates within the framework of *fixed* pdfs f and g, where observations consist of n data points $Y_1, \cdots, Y_n$, independently sampled from either f or g. It is standard to quantify the optimal error rate of this problem using the Chernoff information. We focus on an alternative scenario: when we have two sequences of distributions f_m and g_m, indexed by a parameter m, which diverge with m (as indicated by the unbounded Chernoff information), distinguishing between the two hypotheses at each iteration m becomes feasible with just a *single* data point Y sampled from f_m or g_m. The following lemma, whose proof is given in Appendix A.1, provides the optimal error rate of this hypothesis problem. This lemma cannot be directly derived from existing results, as the pdfs f and g are not fixed but vary with m.

Lemma 2 *Given two sequences of pdfs (f_m) and (g_m) indexed by a parameter $m \in \mathbb{Z}_+$, consider the two hypotheses $H_0\colon Y \sim f_m$ and $H_1\colon Y \sim g_m$. Let $\phi^{\mathrm{MLE}}(Y) \;=\; \mathbb{1}\{f_m(Y) < g_m(Y)\}$ and define the worst-case error of $\phi\colon Y \mapsto \phi(Y) \in \{0,1\}$ by*

$$r(\phi) = \max\{\mathbb{P}\left(\phi(Y)=0 \mid H_1\right), \mathbb{P}\left(\phi(Y)=1 \mid H_0\right)\}.$$

Then, $\inf_\phi r(\phi) = r(\phi^{\mathrm{MLE}})$. Furthermore, if $\mathrm{Chernoff}(f_m, g_m,) = \omega(1)$, we have

$$\log r(\phi^{\mathrm{MLE}}) \;=\; -(1+o(1))\mathrm{Chernoff}(f_m, g_m).$$

Otherwise, if $\mathrm{Chernoff}(f_m, g_m) = O(1)$, we have $r(\phi^{\mathrm{MLE}}) \geq c$ for some constant $c > 0$.

A direct consequence of Lemma 2 is that the classification rule $\hat{z}_i = \arg\max_{m\in[k]} f_m(X_i)$ for all $i \in [n]$ yields an error rate of $\exp(-(1+o(1))\mathrm{Chernoff}(\mathcal{F}))$. Hence, if one has access to the true probability distributions $f_1, \cdots, f_k$ composing the mixture, then the lower bound given in Theorem 1 is tight. In most practical settings, the true probability distributions are unknown. The following section demonstrates how the minimax error rate can still be achieved.

3. Clustering error of iterative algorithms on parametric mixture models

3.1. Parametric mixture model

In this section, we consider the recovery of the clusters of parametric mixture models. More precisely, we let $\mathcal{P}_\Theta = \{f_\theta, \theta \in \Theta\}$ be a family of pdfs parameterised by a subset $\Theta \subset \mathbb{R}^p$. For m points $Y_1, \cdots, Y_m$ sampled from f_θ, we denote by $\hat{\theta}(\{Y_1, \cdots, Y_m\})$ an estimator of θ. We let $z^* \in [k]^n$ be the vector of cluster assignments, and $\theta_1, \cdots, \theta_k \in \Theta$. Conditioned on z^*, the n observed data points $(X_1, \cdots, X_n)$ are independently sampled, such that

$$X_i \,|\, z_i^* = a \;\sim\; f_{\theta_a}. \tag{3.1}$$

A natural estimator $\hat{z}^{\mathrm{oracle}}$ of z^* that uses the knowledge of the true parameters of the model $(\theta_a)_{a\in[k]}$ is given by

$$\forall i \in [n]: \quad \hat{z}_i^{\mathrm{oracle}} \;=\; \underset{a\in[k]}{\arg\max} \log f_{\theta_a}(X_i). \tag{3.2}$$

When the model parameters $(\theta_a)_{a\in[k]}$ are unknown, Algorithm 1 provides an iterative scheme for estimating the cluster assignment z^*. This algorithm sequentially performs the estimation and clustering stages. The goal of this section is to provide general bounds on the error made by Algorithm 1. Following the same proof strategy as previous works on iterative algorithms (Gao and Zhang, 2022), we first decompose the error into two terms and then provide general conditions under which these terms can be upper-bounded.

Algorithm 1 Clustering parametric mixture models.

Input: Set of n data points $(X_1, \cdots, X_n) \in \mathcal{X}^n$, parametric family $\mathcal{P}_\Theta = \{f_\theta, \theta \in \Theta\}$ of pdfs, number of clusters k, number of iteration $t_{\max}$, initial clustering $\hat{z}^{(0)} \in [k]^n$.

Output: Predicted clusters $\hat{z} \in [k]^n$.

For $t = 1 \cdots t_{\max}$ **do**

1. For $a = 1, \cdots, k$, let $\hat{\theta}_a^{(t)} = \hat{\theta}\left(\{X_i \colon z_i^{(t-1)} = a\}\right)$ be an estimate of θ_a;
2. For $i = 1, \cdots, n$ let $\hat{z}_i^{(t)} = \arg\max_{a\in[k]} \log f_{\hat{\theta}_a^{(t)}}(X_i)$.

Return: $\hat{z} = \hat{z}^{(t_{\max})}$.

3.1.1. Decomposition of the error term

Let us introduce $\ell_a(x) = \log f_{\theta_a}(x)$ and $\hat{\ell}_a^{(t)}(x) = \log f_{\hat{\theta}_a^{(t)}}(x)$. We show in Appendix B.1 that we can upper-bound the error $\mathrm{loss}(z^*, \hat{z}^{(t)})$ of Algorithm 1 made at step t as

$$\mathrm{loss}\left(z^*, \hat{z}^{(t)}\right) \;\le\; \xi_{\mathrm{ideal}}(\delta) + \xi_{\mathrm{excess}}^{(t)}(\delta), \tag{3.3}$$

where $\delta > 0$ and

$$\xi_{\mathrm{ideal}}(\delta) \;=\; \sum_{i\in[n]} \sum_{b\in[k]\backslash\{z_i^*\}} \mathbb{1}\left\{\ell_{z_i^*}(X_i) - \ell_b(X_i) < \delta\right\}, \tag{3.4}$$

$$\xi_{\mathrm{excess}}^{(t)}(\delta) \;=\; 2\delta^{-1} \sum_{i\in[n]} \sum_{b\in[k]\backslash\{z_i^*\}} \mathbb{1}\left\{\hat{z}_i^{(t)} = b\right\} \max_{a\in[k]} \left|\hat{\ell}_a^{(t)}(X_i) - \ell_a(X_i)\right|. \tag{3.5}$$

When $\delta = 0$, the ideal error $\xi_{\text{ideal}}(0)$ is an upper bound on the error done by one step of Algorithm 1 that uses the correct parameters $\theta_1^*, \cdots, \theta_k^*$ and not the estimated ones. Studying $\xi_{\text{ideal}}(\delta)$ instead of $\xi_{\text{ideal}}(0)$ gives us some room to control the excess error $\xi_{\text{excess}}^{(t)}(\delta)$ made by estimating the model parameters. The value of δ must be small enough so that $\xi_{\text{ideal}}(\delta)$ has the same asymptotic behaviour as $\xi_{\text{ideal}}(0)$, but large enough so that $\xi_{\text{excess}}^{(t)}(\delta)$ remains small. The following lemma motivates the choice of $\delta = o(\text{Chernoff}(\mathcal{F}))$.

Lemma 3 *Consider a family $\mathcal{F} = (f_1, \cdots, f_k)$ of pdf. Suppose* $\text{Chernoff}(\mathcal{F}) = \omega(1)$ *and let* $\delta = o(\text{Chernoff}(\mathcal{F}))$. *Then, with a probability of at least* $1 - e^{-\sqrt{\text{Chernoff}(\mathcal{F})}}$,

$$\xi_{\text{ideal}}(\delta) \;\leq\; nke^{-(1+o(1))\text{Chernoff}(\mathcal{F})}.$$

3.1.2. CONDITIONS FOR RECOVERY

After Lemma 3, the last remaining step to upper-bound the loss is to control the excess error term. Because the estimates $\hat{z}^{(t)}$ are data dependent, we have to establish that, starting from *any* $\hat{z}^{(0)}$ with a loss small enough, the excess error after one step is upper bounded by a nicely behaved quantity. More precisely, denote z_{new} the clustering obtained after one step of Algorithm 1 starting from some arbitrary initial configuration $z_{\text{old}} \in [k]^n$, and define the following event

$$\mathcal{E}_{\tau,\delta,c,c'} \;=\Big\{ \text{loss}(z^*, z_{\text{old}}) \leq nk^{-1}\tau \;\text{ implies }\; \xi_{\text{excess}}(\delta) \leq c\,\text{loss}(z^*, z_{\text{new}}) + c'\,\text{loss}(z^*, z_{\text{old}}) \Big\},$$

where τ, δ, c, c' are determined later. The following condition states that the event $\mathcal{E}_{\tau,\delta,c,c'}$ holds with probability $1 - o(1)$ (with respect to the data sampling process) for a certain choice of τ, δ.

Condition 1 *Assume there exists $\tau = \Omega(1)$, $\delta = o(\text{Chernoff}(\mathcal{F}))$ and constants $c, c' \in (0,1)$ with $c' < 1 - c$ such that $\mathbb{P}\left(\mathcal{E}_{\tau,\delta,c,c'}\right) \geq 1 - o(1)$.*

Assume $\text{loss}(z^*, \hat{z}^{(0)}) \leq nk^{-1}\tau$. Conditionally on the high probability event $\mathcal{E}_{\tau,\delta,c,c'}$, we establish (by induction and by combining the error decomposition (3.3) with Lemma 3) that

$$\text{loss}\left(z^*, \hat{z}^{(t)}\right) \;\leq\; \frac{nk}{1-c} e^{-(1+o(1))\text{Chernoff}(\mathcal{F})} + \frac{c'}{1-c}\text{loss}\left(z^*, \hat{z}^{(t-1)}\right), \tag{3.6}$$

as long as we can ensure that $\text{loss}\left(z^*, \hat{z}^{(t)}\right) \leq nk^{-1}\tau$ at every step $t \geq 0$ for the *same* $\tau = O(1)$. We can now state the following lemma, whose proof is given in Appendix B.3.

Lemma 4 *Let $\theta_1, \cdots, \theta_k \in \Theta$ and $\mathcal{F} = (f_{\theta_1}, \cdots, f_{\theta_k})$. Let $\tau = \Omega(1)$ such that Condition 1 holds and* $\text{Chernoff}(\mathcal{F}) = \omega(\log(k^2\tau))$. *Let $\hat{z}^{(t)}$ be the output of Algorithm 1 after t steps. We have*

$$\forall t \geq \left\lfloor \log\left(\frac{1-c}{c'}\right) \log n \right\rfloor : \qquad n^{-1}\text{loss}\left(z^*, \hat{z}^{(t)}\right) \;\leq\; e^{-(1+o(1))\text{Chernoff}(\mathcal{F})}.$$

Lemma 4 establishes that Algorithm 1 achieves the minimax rate of recovering z^* with respect to the loss function $n^{-1}\text{loss}(z^*, z)$ after at most $\Theta(\log n)$ iterations when Condition 1 is verified. In the next two sections, we show that this condition holds for specific parametric families $\mathcal{P}(\Theta)$.

3.2. Clustering Laplace mixture models

A real-valued random variable Y has a (1-dimensional) Laplace distribution with location $\mu \in \mathbb{R}$ and scale $\sigma > 0$ if its pdf is $g_{(\mu,\sigma)}(x) = \frac{1}{2\sigma} \exp\left(-\sigma^{-1}|x-\mu|\right)$. We denote such a rv by $Y \sim \mathrm{Lap}(\mu, \sigma)$. In this section, we suppose that the n observed data points $X_1, \cdots, X_n$ belong to $\mathbb{R}^d$ and are generated from the mixture model (1.1) such that for every i, the d coordinates of X_i are independently generated and follow a Laplace distribution, *i.e.,*

$$\forall \ell \in [d] \colon X_{i\ell} \;=\; \mu_{z_i^* \ell} + \sigma_{z_i^* \ell} \epsilon_{i\ell} \quad \text{where } \epsilon_{i\ell} \;\sim\; \mathrm{Lap}(0,1), \tag{3.7}$$

where for all $a \in [k]$ we have $\mu_a \in \mathbb{R}^d$ and $\sigma_a \in (0,\infty)^d$. Equivalently, we can rewrite the mixture (3.7) as a mixture of the parametric family indexed over $\Theta = \mathbb{R}^d \times (0,\infty)^d$ defined by

$$\mathcal{P}(\Theta) \;=\; \left\{ f_\theta(x) = \prod_{\ell=1}^{d} \sigma_\ell^{-1} g_{(0,1)}\left(\frac{x_\ell - \mu_\ell}{\sigma_\ell}\right), \theta = (\mu, \sigma) \right\}.$$

Given a sample $Y_1, \cdots, Y_m$ of a 1-dimensional Laplace distribution, we estimate the location and the scale by

$$\hat{\mu}(Y_1, \cdots, Y_m) \;=\; m^{-1} \sum_{i=1}^{m} Y_i, \quad \text{and} \quad \hat{\sigma}(Y_1, \cdots, Y_m) \;=\; m^{-1} \sum_{i=1}^{m} \left| Y_i - \hat{\mu}(Y_1, \cdots, Y_m) \right|.$$

For simplicity of the exposition of the theorem, we assume that the locations $\mu_{a\ell}$ depend on n, but the scales $\sigma_{a\ell}$ are constant. We denote $\Delta_{\mu,\infty} = \max_{a \neq b \in [k]} \|\mu_a - \mu_b\|_\infty$ the maximum distance between the cluster centres. The following theorem establishes bounds on the recovery of Laplace mixture models. The proof is given in Appendix D.

Theorem 5 *Let $X_1, \cdots, X_n$ be generated from a Laplace mixture model as defined in (3.7). Suppose that $k \log^2(dk) = o(n)$ and $\min_{a\in[k]} \sum_{i\in[n]} \mathbb{1}\{z_i^* = a\} \geq \alpha n k^{-1}$ for some $\alpha > 0$ (independent of n). Assume that $\Delta_{\mu,\infty} = O\left(d^{-1}\mathrm{Chernoff}(\mathcal{F})\right)$ and $\mathrm{Chernoff}(\mathcal{F}) = \omega(d\sqrt{k}(1 + \frac{\Delta_{\mu,\infty}}{\sqrt{nk^{-1}}}))$. Let $\hat{z}^{(t)}$ be the output of Algorithm 1 after t steps, and suppose that the initialization verifies $\mathrm{loss}(z^*, \hat{z}^{(0)}) = o\left(nk^{-1}\Delta_{\mu,\infty}^{-1}\right)$. Then, with probability of at least $1 - o(1)$, it holds*

$$n^{-1}\mathrm{loss}\left(z^*, \hat{z}^{(t)}\right) \;\leq\; e^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})} \quad \forall t \geq \lfloor c \log n \rfloor,$$

for any arbitrary constant $c > 0$.

While we adopt a similar error decomposition approach as in prior works on clustering sub-Gaussian mixtures (Chen and Zhang, 2021; Gao and Zhang, 2022), our analysis of the individual error terms is different due to the sub-exponential nature of the data. This is done in Appendix C.

The conditions $\Delta_{\mu,\infty} = O\left(d^{-1}\mathrm{Chernoff}(\mathcal{F})\right)$ and $\mathrm{loss}(z^*, \hat{z}^{(0)}) = o\left(nk^{-1}\Delta_{\mu,\infty}^{-1}\right)$ in Theorem 5 impose that the quantity $\Delta_{\mu,\infty}$ should not be too large. This might seem counter-intuitive at first. In fact, $\Delta_{\mu,\infty}$ is the *maximum* distance between the cluster centres, and therefore a large $\Delta_{\mu,\infty}$ should not impact the difficulty of recovery. But the first step of Algorithm 1 estimates the quantities $\hat{\mu}_a^{(1)}$ by taking a sample mean based on the initial prediction $\hat{z}^{(0)}$. Because (i) the sample mean is not a robust estimator and (ii) mistakes made by the initial clustering are arbitrary, those mistakes may have an enormous impact on the mean estimation if $\Delta_{\mu,\infty}$ is arbitrarily large. Similar conditions, albeit usually involving $\Delta_{\mu,2} = \max_{a\neq b} \|\mu_a - \mu_b\|_2$, already appear in the study of k-means algorithm for (sub)gaussian mixture models. We refer the reader to (Lu and Zhou, 2016, Section A.5) for a counter-example showing that such a condition is necessary when studying worst-case scenarios. Gao and Zhang (2022) avoid such an extra condition on $\Delta_{\mu,2}$, but at the expense of using a different loss function: $\text{w-loss}\,(z^*, \hat{z}) = \sum_{a=1}^{k} \sum_{\substack{b=1\\ b\neq a}}^{k} \|\mu_a - \mu_b\|^2 \mathbb{1}\{z_i^* = a, \hat{z}_i = b\}$. This new loss function imposes a heavier penalty on mistakes made between clusters having a large $\|\mu_a - \mu_b\|^2$. Assuming that $\text{w-loss}\left(z^*, \hat{z}^{(0)}\right) = o(nk^{-1} \min_{a\neq b} \|\mu_a - \mu_b\|^2)$, Gao and Zhang (2022) establish that Lloyd's algorithm attains the optimal error rate in isotropic Gaussian mixture models. The assumption of $\text{w-loss}\left(z^*, \hat{z}^{(0)}\right)$ is stronger than the assumption on $\text{loss}\left(z^*, \hat{z}^{(0)}\right)$, as the former automatically rules out settings in which too many mistakes are made across cluster pairs that have a large $\|\mu_a - \mu_b\|$[1].

Finally, we note that in previous literature, the difficulty of clustering is expressed by a small signal-to-noise ratio, instead of a small Chernoff information. In many cases, the two are related, but as we saw in the introduction, the signal-to-noise ratio might take a different expression depending on the model considered. This is also the case for the Laplace mixture model. For example, if each dimension has a unique scale across the k clusters (*i.e.,* $\sigma_{1\ell} = \cdots = \sigma_{k\ell} = \sigma_\ell$), we have (see detailed computations in Appendix D.3)

$$\text{Chernoff}(\mathcal{F}) \;=\; (1 + o(1)) \min_{a\neq b} \sum_{\ell=1}^{d} \frac{|\mu_{a\ell} - \mu_{b\ell}|}{\sigma_\ell}.$$

This can be rewritten as

$$\text{Chernoff}(\mathcal{F}) \;=\; (1 + o(1)) \min_{a\neq b} \|\Sigma^{-1}(\mu_a - \mu_b)\|_1,$$

where Σ is the diagonal matrix whose elements are $\sigma_1, \cdots, \sigma_\ell$. This quantity $\|\Sigma^{-1}(\mu_a - \mu_b)\|_1$ can be interpreted as an SNR. If we further restrict $\sigma_1 = \cdots = \sigma_\ell = \sigma$ (isotropic Laplace mixture model), we obtain

$$\text{Chernoff}(\mathcal{F}) \;=\; (1 + o(1)) \frac{\min_{a\neq b} \|\mu_a - \mu_b\|_1}{\sigma}.$$

For this isotropic Laplace mixture model, the error rate involves the ℓ^1 distance, instead of the more traditional ℓ^2 distance used in the Gaussian mixture model (see (1.5)).

3.3. Bregman hard clustering of exponential family mixture models

A set of pdf $\mathcal{P}_\psi(\Theta) = \{p_\theta, \theta \in \Theta\}$ form an exponential family if each pdf p_θ (defined with respect to a common reference measure ν) can be expressed as

$$p_{\psi,\theta}(y) \;=\; h(y) e^{<u(y),\theta> - \psi(\theta)}, \tag{3.8}$$

where $h(\cdot)$ is the carrier measure, $u(\cdot)$ is the sufficient statistics, $\psi(\theta) = \log \int h(y) e^{<u(y),\theta>} d\nu(y)$ is the log-normalizer (also called the cumulant function), and θ is the natural parameter belonging

1. More precisely, we notice that $\min_{a\neq b} \|\mu_a - \mu_b\|^2 \text{loss}(z^*, z) \leq \text{w-loss}(z^*, z)$ for all $z \in [k]^n$. Therefore, the condition $\text{w-loss}\left(z^*, \hat{z}^{(0)}\right) = o(nk^{-1} \min_{a\neq b} \|\mu_a - \mu_b\|^2)$ implies $\text{loss}\left(z^*, \hat{z}^{(0)}\right) = o(nk^{-1})$, but the converse does not hold.

to the space $\Theta = \{\theta \in \mathbb{R}^p \colon \psi(\theta) < \infty\}$. We assume that Θ is open so that $\mathcal{P}(\Theta)$ forms a regular exponential family, and that u is a minimal sufficient statistics[2]. Among important properties of regular exponential families, we recall that ψ is a differentiable and strongly convex function which verifies $\mathbb{E}_{Y\sim p_{\psi,\theta}}\left[u(Y)\right] = \nabla\psi(\theta)$ (Banerjee et al., 2005, Sections 4.1 and 4.2).

We consider a family of k pdf $\mathcal{F} = \{f_{\theta_1}, \cdots, f_{\theta_k}\}$ belonging to the *same* exponential family, such as each f_{θ_a} can be written as

$$f_{\theta_a}(x_1, \cdots, x_d) \;=\; \prod_{\ell=1}^{d} h_\ell(x_\ell) e^{< u_\ell(x_\ell), \theta_{a\ell} > - \psi_\ell(\theta_{a\ell})}. \tag{3.9}$$

In other words, each coordinate $X_{i\ell}$ of X_i is sampled from an exponential family with sufficient statistics u_ℓ, cumulant function ψ_ℓ, and parameter $\theta_{z_i^*\ell}$. We note that, for each coordinate ℓ, different clusters share the *same* the sufficient statistic u_ℓ and cumulant function ψ_ℓ, but have *different* parameters $\theta_{1\ell}, \cdots, \theta_{k\ell}$. Moreover, we assume that u_ℓ is a function from $\mathbb{R}$ to $\mathbb{R}$, but our results extend naturally if $u_\ell \colon \mathbb{R} \to \mathbb{R}^p$.

For any convex, differentiable function $\varphi\colon \Theta \to \mathbb{R}$, we define its *Legendre transform* as $\varphi^*(y) = \sup_{\theta\in\Theta}\{< \theta, y > - \varphi(\theta)\}$. The *Bregman divergence* $\mathrm{Breg}_\varphi(\cdot\|\cdot)$ with generator φ is defined by

$$\mathrm{Breg}_\varphi(x\|y) \;=\; \varphi(x) - \varphi(y) - (x-y)^T \nabla\varphi(y).$$

The pdf $p_{\psi,\theta}$ defined in (3.8) can be rewritten as (Banerjee et al., 2005, Theorem 4)

$$p_{\psi,\theta}(y) \;=\; b_\psi(y) e^{-\mathrm{Breg}_{\psi^*}(x,\mu)},$$

where $b_\psi(\cdot)$ is independent of θ. Therefore, for any $f_{\theta_a} \in \mathcal{P}(\Theta^d)$ we have

$$f_{\theta_a}(x) \;=\; b(x) e^{-\sum_{\ell=1}^{L} \mathrm{Breg}_{\psi_\ell^*}(u_\ell(x_\ell), \mu_{a\ell})},$$

where $\mu_{a\ell} = \mathbb{E}_{X\sim f_{\theta_a}}\left[u_\ell(X_\ell)\right] = \nabla\psi_\ell(\theta_{a\ell})$. Therefore, for a mixture model for the parametric family (3.8), we can reformulate Algorithm 1 as Algorithm 2. As in Section 3.2, we also define $\Delta_{\mu,\infty} = \max_{1\le a\ne b\le k} \|\mu_a - \mu_b\|_\infty$.

The following theorem, whose proof is provided in Appendix E, shows that Algorithm 2 is rate-optimal if correctly initialised.

Theorem 6 *Let $X_1, \cdots, X_n$ be generated from a mixture model of exponential family as defined in (3.9), and such that $u_\ell(X_{i\ell})$ is sub-exponential. Suppose that $k\log^2(dk) = o(n)$ and $\min_{a\in[k]}\sum_{i\in[n]} \mathbb{1}\{z_i^* = a\} \ge \alpha n k^{-1}$ for some constant $\alpha > 0$. Suppose that $\nabla^2\psi^*(\mu_{a\ell}) = \Theta(1)$, $\Delta_{\mu,\infty} = O(d^{-1}\mathrm{Chernoff}(\mathcal{F}))$ and $\mathrm{Chernoff}(\mathcal{F}) = \omega(d\sqrt{k}(1 + \frac{\Delta_{\mu,\infty}}{\sqrt{nk^{-1}}}))$. Let $\hat{z}^{(t)}$ be the output of Algorithm 2 after t steps, where the initialisation verifies $\mathrm{loss}(z^*, \hat{z}^{(0)}) = o\left(nk^{-1}\Delta_{\mu,\infty}^{-1}\right)$. Then, with a probability of at least $1 - o(1)$, it holds*

$$n^{-1}\mathrm{loss}\left(z^*, \hat{z}^{(t)}\right) \;\le\; e^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})} \quad \forall t \ge \lfloor c\log n \rfloor,$$

for any arbitrary constant $c > 0$.

2. A sufficient statistic u is said to be minimal if for any other sufficient statistic $\tilde{u}$, there exists a measurable function φ such that $u = \varphi(\tilde{u})$.

Algorithm 2 Bregman hard clustering (Banerjee et al., 2005)

Input: Set of n data points $(X_1, \cdots, X_n) \in \mathbb{R}^{n \times d}$, sufficient statistics $u_1, \cdots, u_d \colon \mathbb{R} \to \mathbb{R}$, convex functions $\psi_1^*, \cdots, \psi_p^* \colon \mathbb{R} \to \mathbb{R}$, number of clusters k, number of iteration $t_{\max}$, initial clustering $\hat{z}^{(0)} \in [k]^n$.

Output: Predicted clusters $\hat{z} \in [k]^n$.

For $t = 1 \cdots t_{\max}$ **do**

1. For $a = 1, \cdots, k$ and $\ell = 1, \cdots, d$, let $\hat{\mu}_{a\ell} = \frac{\sum_i \mathbb{1}\left\{\hat{z}_i^{(t-1)} = a\right\} u_\ell(X_{i\ell})}{\sum_i \mathbb{1}\left\{\hat{z}_i^{(t-1)} = a\right\}}$;
2. For $i = 1, \cdots, n$ let $\hat{z}_i^{(t)} = \underset{a \in [k]}{\arg\min} \sum_{\ell=1}^{L} \mathrm{Breg}_{\psi_\ell^*}(u_\ell(x_\ell), \hat{\mu}_{a\ell})$.

Return: $\hat{z} = \hat{z}^{(t_{\max})}$.

We need the technical condition $\nabla^2 \psi_\ell^*(\mu_{a\ell}) = \Theta(1)$ to control the term $\mathrm{Breg}_{\psi_\ell^*}(\mu_{a\ell}, \hat{\mu}_{a\ell})$ when $\hat{\mu}_{a\ell}$ is an estimate of $\mu_{a\ell}$. This condition is verified in many models of interest (such as Poisson, Negative Binomial, Exponential, or Gaussian mixture models). For example, for Poisson distributions, we have $\psi^*(x) = x \log x - 1$ and hence $\nabla^2 \psi^*(x) = x^{-1}$ is a $\Theta(1)$ if we assume that the intensities of the Poisson pdf forming the mixture are all lower-bounded.

The assumption that $u_\ell(X_{i\ell})$ is sub-exponential can be verified even if $X_{i\ell}$ has a heavier tail than exponential. For example, if $X_{i\ell}$ is log-normal, then $u_\ell(X_{i\ell}) = \log(X_{i\ell})$ is Gaussian and hence has sub-exponential tails. Pareto distribution provides another interesting example: if $X_{i\ell}$ is Pareto distributed with shape α and scale $x_m = 1$, then $\log X$ is exponentially distributed with mean α^{-1}.

Finally, we notice that, except for particular cases (such as Gaussian mixture models), the quantity $\mathrm{Chernoff}(\mathcal{F})$ does not have a nice closed-form expression, and we can not easily define an SNR in those models. An important example of such a quantity already appearing in the literature is the *Chernoff-Hellinger divergence*, originally defined in Stochastic Block Models (Abbe and Sandon, 2015; Dreveton et al., 2023), and appearing in the study of Poisson mixture models, as shown in the following example.

Example 2 (Poisson mixture model) *Consider the family $\mathcal{F} = \{f_{\theta_1}, \cdots, f_{\theta_k}\}$ of multi-variate Poisson distributions, defined by $f_{\theta_a}(x) = \prod_{\ell=1}^d \frac{\theta_{a\ell}^{x_\ell}}{x_\ell!} e^{-\theta_{a\ell}}$ for $x \in \mathbb{Z}_+^d$ and $\theta_a \in \mathbb{R}_+^d$. Then,* $\mathrm{Chernoff}(\mathcal{F}) = \min_{1 \le a \ne b \le k} \sup_{t \in (0,1)} \sum_{\ell=1}^d \left(t\theta_{a\ell} + (1-t)\theta_{b\ell} - \theta_{a\ell}^t \theta_{b\ell}^{1-t} \right)$.

4. Discussion

4.1. Initialisation

In the literature, initialisation is commonly accomplished through *spectral methods*, an umbrella term denoting a dimension reduction via spectral decomposition followed by clustering. Here, we perform the dimension reduction through the Singular Value Decomposition (SVD) of a well-chosen matrix Y, and the clustering is done by finding an $(1+\epsilon)$-approximation of a k-means problem.

1. Let $Y = \sum_{\ell=1}^{p \wedge n} s_\ell v_\ell w_\ell^T$ with $s_1 \ge s_2 \ge \cdots \ge s_{p \wedge n}$ be the SVD decomposition of $Y \in \mathbb{R}^{p \times n}$. Let $V = [v_1, \cdots, v_k] \in \mathbb{R}^{p \times k}$ and define $\hat{M} = VY \in \mathbb{R}^{k \times n}$.

2. Return $\hat{z}^{(0)}$, an $(1+\epsilon)$ approximation of $\underset{\substack{z \in k^n \\ \check{\mu} \in \mathbb{R}^{d \times k}}}{\arg\min} \sum_{i=1}^n \|\hat{M}_{\cdot i} - \check{\mu}_{z_i}\|_2^2$, where $\hat{M}_{\cdot i}$ is the i-th column of $\hat{M}$ (Kumar et al., 2004).

For the Laplace mixture model, we apply the SVD directly on $Y = X$, while for an exponential family mixture, we apply it on the matrix Y obtained such that $Y_{i\ell} = u_\ell(X_{i\ell})$ for all $i \in [n], \ell \in [d]$. The following lemma ensures that the error made by this initialisation is $o(nk^{-1}\Delta_{\mu,\infty}^{-1})$, as required by Theorems 5 and 6.

Lemma 7 *Define $\delta_{\mu,2} = \min_{1 \le a \neq b \le k} \|\mu_a - \mu_b\|_2$. Let $\hat{z}^{(0)}$ be the clustering obtained by the initialisation described above, with ϵ being defined in step 2. Assume $\min_{a \in [k]} \sum_{i \in [n]} \mathbb{1}\{z_i^* = a\} \ge \alpha n k^{-1}$ for some constant $\alpha > 0$ and $\Delta_{\mu,\infty} = o\left(\frac{\delta_{\mu,2}^2}{(1+\epsilon)k^2(1+\frac{d}{n})}\right)$. Then*

$$\mathrm{loss}\left(z^*, \hat{z}^{(0)}\right) \;=\; o\left(nk^{-1}\Delta_{\mu,\infty}^{-1}\right).$$

The proof of Lemma 7 follows the same steps as in the proof of (Gao and Zhang, 2022, Proposition 4.1), the only modification being a different choice of the loss function. The central argument in the proof of (Gao and Zhang, 2022, Proposition 4.1) is that $\|Y - \mathbb{E}Y\|_2 = O(\sqrt{n+d})$ with probability at least $1 - e^{-Cn}$ for some $C > 0$ when Y has independent Gaussian entries. In our setting, Y is a random matrix with independent, sub-exponential random entries, and hence its concentrate (see for example (Bandeira and van Handel, 2016, Corollary 3.5) and (Dai et al., 2023)).

Finally, Lemma 7 requires an additional assumption on $\delta_{\mu,2}$. While we might be able to get rid of this extra technical condition, we also notice that this condition is verified in interesting regimes. We refer the reader to the detailed example of the Laplace mixture in Section D.3, for which $\delta_{\mu,2}^2 = \Theta(d\Delta_{\mu,\infty}^2)$ and $\mathrm{Chernoff}(\mathcal{F}) = \Theta(d\Delta_{\mu,\infty})$. In this regime, the extra condition in Lemma 7 becomes $\mathrm{Chernoff}(\mathcal{F}) = \omega((1+\epsilon)k^2(1+dn^{-1}))$, which is weak if $d = o(n)$.

4.2. Discussion and future work

There has been a recent surge in interest in establishing the error rates of various clustering algorithms in (sub-Gaussian) mixture models. In this section, we provide a concise overview of some of the latest and most pertinent works in this area, as well as directions for future work.

Robustness to model specification, perturbed samples, and heavier tails Due to its simplicity and inclusion in popular libraries like *scikit-learn*, the standard Lloyd's algorithm often serves as the default choice for clustering tasks. While its optimality has been demonstrated for clustering isotropic Gaussian mixture models, its performance on other mixture models has not been studied. More generally, Theorems 5 and 6 demonstrate that iterative algorithms are rate-optimal when the parametric family underlying the mixture distributions is known. But what happens under model misspecification? For instance, what error rate can we expect to achieve if we cluster a mixture of negative binomial distributions using the Bregman divergence associated with the Poisson distribution? As a first result in this direction, (Jana et al., 2023, Theorem 1) establishes that employing the ℓ^1 distance instead of the squared ℓ^2 distance for clustering a mixture of isotropic Gaussian yields an error rate of *at least* $\exp(-(2+C)^{-1}\mathrm{SNR}_{\text{isotropic}}^2)$, where $C > 0$, which is larger than the optimal rate of $\exp(-2^{-1}\mathrm{SNR}_{\text{isotropic}}^2)$.

Another important type of robustness lies in the observation of perturbed samples. Suppose that $\{X_i\}_{i\in[n]}$ is generated from a mixture model, but we observe a perturbed sample $\{\tilde{X}_i\}_{i\in[n]}$ with $\tilde{X}_i = X_i + e_i$, where the noise terms $\{e_i\}_{i\in[n]}$ verify $\|e_i\| \le \epsilon$. For sub-Gaussian X_i, (Patel et al., 2023, Theorem 4.1) establish that the mis-clustering rate of Lloyd's algorithm on this model is *at most* $\exp\left(-4^{-1}\mathrm{SNR}^2_{\text{isotropic}} \min\{1, 2\epsilon\}\right)$.

Because the mean is notoriously non-robust to outliers (Tukey, 1960; Huber, 1964), another strategy to ensure the robustness of the iterative method is to estimate the cluster means by a robust location estimator, such as the coordinate-wise median (Jana et al., 2023), the geometric median (Godichon-Baggioni and Robin, 2024), or trimmed estimators (Cuesta-Albertos et al., 1997; García-Escudero et al., 2008; Brécheteau et al., 2021). Furthermore, robust estimators might become necessary for handling distributions with tails heavier than sub-exponential.

High dimension regime When d can grow arbitrarily large, Ndaoud (2022) showed that the optimal error rate for clustering mixture of isotropic Gaussians with $k = 2$ clusters is no longer $\exp(-2^{-1}\mathrm{SNR}^2_{\text{isotropic}})$ but becomes $\exp\left(\Theta\left(\frac{\mathrm{SNR}^4_{\text{isotropic}}}{\mathrm{SNR}^2_{\text{isotropic}}+dn^{-1}}\right)\right)$. An extension to $k = \Theta(1)$ clusters is studied in Chen and Yang (2021). The theoretical analysis of both of these works heavily relies on the Gaussian assumption, and it remains open to extend such results to other mixture models. The key challenge is that in a mixture of two isotropic Gaussians $\frac{1}{2}\mathcal{N}(\mu_1, I_d) + \frac{1}{2}\mathcal{N}(\mu_2, I_d)$ where $d \gg n$, the dimension of the parameters of the distributions ($\mu_1, \mu_2 \in \mathbb{R}^d$) is much larger than the number of data points n. This creates a discrepancy between the minimax error rates of algorithms with and without access to the true centres μ_1, μ_2 (Ndaoud, 2022). Exploring this phenomenon for models beyond the mixture of two isotropic Gaussians is a crucial avenue for future research.

(Semi)-supervised extensions Once the unsupervised error rate of various mixture models is well understood, researchers can also examine the supervised error rate of classification (Li et al., 2017; Minsker et al., 2021). An intriguing perspective emerges when extending these analyses to a semi-supervised setting, aiming to ascertain whether a small amount of labelled data can notably diminish the clustering error rate (Lelarge and Miolane, 2019; Tifrea et al., 2023).

References

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688. IEEE, 2015.

Konstantin Avrachenkov, Maximilien Dreveton, and Lasse Leskelä. Community recovery in non-binary and temporal stochastic block models. *arXiv preprint arXiv:2008.04790*, 2022.

Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479 – 2506, 2016.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, Joydeep Ghosh, and John Lafferty. Clustering with Bregman divergences. *Journal of machine learning research*, 6(10), 2005.

Charles Bouveyron and Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014.

Claire Brécheteau, Aurélie Fischer, and Clément Levrard. Robust Bregman clustering. *The Annals of Statistics*, 49(3):1679–1701, 2021.

Xiaohui Chen and Yun Yang. Cutoff for exact recovery of Gaussian mixture models. *IEEE Transactions on Information Theory*, 67(6):4223–4238, 2021.

Xin Chen and Anderson Y. Zhang. Optimal clustering in anisotropic Gaussian mixture models. *arXiv preprint arXiv:2101.05402*, 2021.

Juan Antonio Cuesta-Albertos, Alfonso Gordaliza, and Carlos Matrán. Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576, 1997.

Guozheng Dai, Zhonggen Su, and Hanchao Wang. Tail bounds on the spectral norm of sub-exponential random matrices. *Random Matrices: Theory and Applications*, 2023. doi: 10.1142/S2010326323500132.

Maximilien Dreveton, Felipe Schreiber Fernandes, and Daniel R Figueiredo. Exact recovery and Bregman hard clustering of node-attributed stochastic block model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Chao Gao and Anderson Y. Zhang. Iterative algorithm for discrete structure recovery. *The Annals of Statistics*, 50(2):1066 – 1094, 2022.

Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153 – 2185, 2018.

Luis A. García-Escudero, Alfonso Gordaliza, Carlos Matrán, and Agustin Mayo-Iscar. A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3):1324 – 1345, 2008.

Manuel Gil, Fady Alajaji, and Tamas Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.

Antoine Godichon-Baggioni and Stéphane Robin. A robust model-based clustering based on the geometric median and the median covariation matrix. *Statistics and Computing*, 34(1):55, 2024.

Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637–640, 2014.

Ashraful Haque, Jessica Engel, Sarah A Teichmann, and Tapio Lönnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine*, 9(1):1–12, 2017.

Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

Soham Jana, Kun Yang, and Sanjeev Kulkarni. Adversarially robust clustering with optimality guarantees. *arXiv preprint arXiv:2306.09977*, 2023.

Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 454–462. IEEE, 2004.

Marc Lelarge and Léo Miolane. Asymptotic Bayes risk for Gaussian mixture in a semi-supervised setting. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 639–643. IEEE, 2019.

Tianyang Li, Xinyang Yi, Constantine Carmanis, and Pradeep Ravikumar. Minimax Gaussian classification & clustering. In *Artificial Intelligence and Statistics*, pages 1–9, 2017.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Yu Lu and Harrison H Zhou. Statistical and computational guarantees of Lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

Stanislav Minsker, Mohamed Ndaoud, and Yiqiu Shen. Minimax supervised clustering in the anisotropic Gaussian mixture model: A new take on robust interpolation. *arXiv preprint arXiv:2111.07041*, 2021.

Mohamed Ndaoud. Sharp optimal recovery in the two component Gaussian mixture model. *The Annals of Statistics*, 50(4):2096–2126, 2022.

Dhruv Patel, Hui Shen, Shankar Bhamidi, Yufeng Liu, and Vladas Pipiras. Consistency of Lloyd's algorithm under perturbations. *arXiv preprint arXiv:2309.00578*, 2023.

Alexandru Tifrea, Gizem Yüce, Amartya Sanyal, and Fanny Yang. Can semi-supervised learning use all the data effectively? A lower bound perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

John Wilder Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, Philip S Yu, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14:1–37, 2008.

Appendix A. Proof of the lower-bound

A.1. Proof of Lemma 2

We recall that, given two pdfs f and g with respect to a reference dominating measure ν, the Rényi divergence of order t between f and g is

$$\mathrm{Ren}_t(f\|g) \;=\; \frac{1}{t-1}\log\int f^t(x)g^{1-t}(x)d\nu(x).$$

Chernoff information and Rényi divergences are linked by the following relationship

$$\mathrm{Chernoff}(f,g) \;=\; \sup_{t\in(0,1)}(1-t)\mathrm{Ren}_t(f\|g).$$

The Rényi divergence is not symmetric in f and g (except for $t=2^{-1}$), but the Chernoff information is symmetric.

Proof [Proof of Lemma 2] Let $\ell(Y)=\log\frac{g_m}{f_m}(Y)$. By the definition of ϕ^{MLE} and of the worst-case risk $r(\cdot)$, we have

$$r\left(\phi^{\mathrm{MLE}}\right) \;=\; \max\left\{\mathbb{P}_{Y\sim f_m}\left(\ell(Y)>0\right),\mathbb{P}_{Y\sim g_m}\left(\ell(Y)<0\right)\right\}.$$

In the following, we establish upper and lower bounds for $\mathbb{P}_{Y\sim f_m}\left(\ell(Y)>0\right)$. A similar reasoning provides bounds for $\mathbb{P}_{Y\sim g_m}\left(\ell(Y)<0\right)$.

(i) Upper-bound. Applying Chernoff bounds, it holds for any $t\in(0,1)$

$$\mathbb{P}_{Y\sim f_m}\left(\ell(Y)>0\right) \;=\; \mathbb{P}_{Y\sim f_m}\left(e^{t\ell(Y)}>1\right) \;\le\; \mathbb{E}_{f_m}\left[e^{t\ell(Y)}\right].$$

By the definition of $\ell(Y)$, we have $\mathbb{E}_{f_m}\left[e^{t\ell(Y)}\right] \;=\; \mathbb{E}_{f_m}\left[\left(\frac{g_m}{f_m}(Y)\right)^t\right]$. By the definition of the Rényi divergence, we also have $\mathbb{E}_{f_m}\left[\left(\frac{g_m}{f_m}(Y)\right)^t\right]=e^{-(1-t)\mathrm{Ren}_t(g_m\|f_m)}$. Hence,

$$\begin{aligned}\mathbb{P}_{Y\sim f_m}\left(\ell(Y)>0\right) &\le \inf_{t\in(0,1)} e^{-(1-t)\mathrm{Ren}_t(g_m\|f_m)}\\ &= e^{-\sup_{t\in(0,1)}(1-t)\mathrm{Ren}_t(g_m\|f_m)}\\ &= e^{-\mathrm{Chernoff}(f_m,g_m)}.\end{aligned}$$

We can similarly establish that $\mathbb{P}_{Y\sim g_m}\left(\ell(Y)<0\right)\le e^{-\mathrm{Chernoff}(f_m,g_m)}$, and thus

$$\log r(\phi^{\mathrm{MLE}}) \;\le\; -\mathrm{Chernoff}(f_m,g_m).$$

(ii) Lower-bound. For any $s\ge 0$ and $t\in(0,1)$, we have

$$\begin{aligned}\mathbb{P}_{f_m}\left(\ell(Y)>0\right) &\ge \mathbb{E}_{f_m}\left[e^{-t\ell(Y)}e^{t\ell(Y)}\mathbb{1}\{0\le\ell(Y)\le s\}\right]\\ &\ge e^{-ts}\,\mathbb{E}_{f_m}\left[e^{t\ell(Y)}\mathbb{1}\{0\le\ell(Y)\le s\}\right].\end{aligned}$$

Next, we define $h_{1-t} = \frac{f_m^{1-t} g_m^t}{\int f_m^{1-t} g_m^t}$. We notice that $\int f_m^{1-t} g_m^t = e^{-(1-t)\mathrm{Ren}_t(g_m \| f_m)}$, and furthermore

$$\begin{aligned} \mathbb{E}_{f_m}\left[e^{t\ell(Y)} \mathbb{1}\{0 \le \ell(Y) \le s\} \right] &= \int f_m^{1-t}(y) g_m^t(y) \mathbb{1}\{0 \le \ell(y) \le s\} \, d\nu(y) \\ &= e^{-(1-t)\mathrm{Ren}_t(g_m \| f_m)} \mathbb{P}_{h_{1-t}}(\ell(\tilde{Y}) \in [0, s]), \end{aligned}$$

where $\tilde{Y}$ is a random variable distributed from h_{1-t}. Therefore,

$$\begin{aligned} \mathbb{P}_{f_m}\left(\ell(Y) > 0 \right) &\ge e^{-ts} e^{-(1-t)\mathrm{Ren}_t(g_m \| f_m)} \mathbb{P}_{h_{1-t}}(\ell(\tilde{Y}) \in [0, s]) \\ &\ge e^{-s} e^{-\mathrm{Chernoff}(f_m, g_m)} \mathbb{P}_{h_{1-t}}(\ell(\tilde{Y}) \in [0, s]), \end{aligned}$$

where we used $e^{-ts} \ge e^{-s}$ and $(1-t)\mathrm{Ren}_t(g_m \| f_m) \le \mathrm{Chernoff}(f_m \| g_m)$ because $t \in (0, 1)$. Since the previous inequality is valid for any t, we obtain by taking $t = \frac{1}{2}$ that

$$\mathbb{P}_{f_m}\left(\ell(Y) > 0 \right) \ge e^{-s} e^{-\mathrm{Chernoff}(f_m, g_m)} \mathbb{P}_{h_{1/2}}(\ell(\tilde{Y}) \in [0, s]).$$

Next, we notice that $\mathbb{P}_{h_{1/2}}(\ell(\tilde{Y}) \in [0, s]) = \mathbb{P}_{h_{1/2}}(\ell(\tilde{Y}) \in [-s, 0])$. Let $s = \sqrt{2\mathbb{E}_{h_{1/2}}\left[\ell(\tilde{Y})^2 \right]}$. Then, Chebyshev's inequality implies that $\mathbb{P}_{h_{1/2}}(|\ell(\tilde{Y})| > s) \le s^{-2} \mathbb{E}_{h_{1/2}}\left[\ell(\tilde{Y})^2 \right] \le \frac{1}{2}$. ■

A.2. Proof of Theorem 1

This section is devoted to the proof of Theorem 1, which provides a lower bound of the minimax loss $\inf_{\hat{z}} \sup_{z \in [k]^n} \mathrm{loss}(z, \hat{z})$. Without loss of generality, we assume that the minimum in the definition of $\mathrm{Chernoff}(\mathcal{F})$ in (2.1) is achieved at $a = 1$ and $b = 2$, that is

$$\mathrm{Chernoff}(f_1, f_2) = \min_{1 \le a \ne b \le k} \mathrm{Chernoff}(f_a, f_b).$$

For any cluster membership vector $z \in [k]^n$, we denote for all $a \in [k]$ the set of all indices $i \in [n]$ belonging to cluster a by

$$\Gamma_a(z) = \{ i \in [n] \colon z_i = a \}.$$

Recall from Definition (1.2) that $\mathrm{loss}(z^*, \hat{z}) = \min_{\tau \in \mathrm{Sym}(k)} \mathrm{Ham}(z^*, \tau \circ \hat{z})$. In particular, the loss function involves a minimum over all permutations $\tau \in \mathrm{Sym}(k)$, making it hard to study directly. But, we can get rid of this minimum in the definition of the loss if we deal with vectors having a loss small enough, because in that case the min is attained by a unique minimiser. We state without proof the following lemma from Avrachenkov et al. (2022).

Lemma 8 (Lemma C.5 in Avrachenkov et al. (2022)) *Let $z_1, z_2 \in [k]^n$ such that $\mathrm{Ham}(z_1, \tau^* \circ z_2) < \frac{1}{2} \min_{a \in [k]} |\Gamma_a(z_1)|$ for some $\tau^* \in \mathrm{Sym}(k)$. Then τ^* is the unique minimiser of $\tau \in \mathrm{Sym}(k) \mapsto \mathrm{Ham}(z_1, \tau \circ z_2)$.*

Following the same proof strategy as previous works on clustering block models (Gao et al., 2018; Chen and Zhang, 2021), we define a clustering problem over a subset of $[k]^n$ to avoid the issues of label permutations. Let $\alpha > 0$ be an arbitrary constant independent of n. We define $\mathcal{Z} = \mathcal{Z}_{n,k} \subset [k]^n$ the set of vectors such that all clusters have size at least $\alpha n k^{-1}$:

$$\mathcal{Z} \;=\; \left\{ z \in [k]^n \colon \; |\Gamma_a(z)| \;\ge\; \alpha n k^{-1} \quad \text{ for all } a \in [k] \right\}.$$

Let $z^* \in \mathcal{Z}$. For every cluster $a \in [k]$, collect the indices of the $|\Gamma_a(z^*)| - \frac{\alpha n}{5k}$ smallest indices i's in $\Gamma_a(z^*) = \{ i \in [n] \colon z_i^* = a \}$ into a set T_a. Let $T = T_1 \cup T_2 \cup_{a=3}^k \Gamma_a(z^*)$ and define a new parameter space $\tilde{\mathcal{Z}}$

$$\tilde{\mathcal{Z}} \;=\; \left\{ z \in \mathcal{Z} \colon z_i = z_i^* \text{ for all } i \in T \quad \text{ and } \quad z_i \in \{1,2\} \text{ if } i \in T^c \right\}.$$

Because $T^c = T_1^c \cup T_2^c$, this new space $\tilde{\mathcal{Z}}$ is composed of all cluster labelling z that only differs from z^* on the indices i's that do not belong to T_1 or T_2. By construction of $\tilde{\mathcal{Z}}$, we have for any $z, z' \in \tilde{\mathcal{Z}}$

$$\mathrm{Ham}(z, z') \;=\; \sum_{i=1}^n \mathbb{1}\{ z_i \ne z_i' \} \;\le\; |T^c| \;=\; 2 \, \frac{\alpha n}{5k}.$$

Because $z \in \tilde{\mathcal{Z}} \subset \mathcal{Z}$, we have by definition of $\mathcal{Z}$ that $\min_{a \in [k]} |\Gamma_a(z)| > \alpha n k^{-1}$. Therefore, the previous inequality ensure that $\mathrm{Ham}(z, z') < 2^{-1} \min_{a \in [k]} |\Gamma_a(z)|$ for all $z, z' \in \mathcal{Z}$. We can thus apply Lemma 8 to establish that

$$\forall z, z' \in \tilde{\mathcal{Z}} : \quad \mathrm{loss}(z, z') \;=\; \mathrm{Ham}(z, z') \;=\; \sum_{i \in T^c} \mathbb{1}\{ z_i \ne z_i' \}. \tag{A.1}$$

Because $\tilde{\mathcal{Z}} \subset \mathcal{Z} \subset [k]^n$, we also have

$$\inf_{\hat{z}} \sup_{z \in [k]^n} \mathbb{E}_z \mathrm{loss}(z, \hat{z}) \;\ge\; \inf_{\hat{z}} \sup_{z \in \tilde{\mathcal{Z}}} \mathbb{E}_z \mathrm{loss}(z, \hat{z}) \;=\; \inf_{\hat{z}} \sup_{z \in \tilde{\mathcal{Z}}} \mathbb{E}_z \mathrm{Ham}(z, \hat{z}),$$

where the equality follows from (A.1). Bounding the minimax risk by the Bayes risk leads to

$$\inf_{\hat{z}} \sup_{z \in \tilde{\mathcal{Z}}} \mathbb{E}_z \left[\mathrm{Ham}(z, \hat{z}) \right] \;\ge\; \inf_{\hat{z}} \frac{1}{|\tilde{\mathcal{Z}}|} \sum_{z \in \tilde{\mathcal{Z}}} \mathbb{E}_z \left[\mathrm{Ham}(z, \hat{z}) \right].$$

Moreover,

$$\begin{aligned} \inf_{\hat{z}} \frac{1}{|\tilde{\mathcal{Z}}|} \sum_{z \in \tilde{\mathcal{Z}}} \mathbb{E}_z \mathrm{Ham}(z, \hat{z}) \;&=\; \inf_{\hat{z}_1, \cdots, \hat{z}_n} \frac{1}{|\tilde{\mathcal{Z}}|} \sum_{z \in \tilde{\mathcal{Z}}} \sum_{i \in T^c} \mathbb{P}_z(\hat{z}_i \ne z_i) \\ &=\; \sum_{i \in T^c} \inf_{\hat{z}_i} \frac{1}{|\tilde{\mathcal{Z}}|} \sum_{z \in \tilde{\mathcal{Z}}} \mathbb{P}_z(\hat{z}_i \ne z_i). \end{aligned}$$

Therefore, we can conclude from these previous inequalities that

$$\inf_{\hat{z}} \sup_{z \in [k]^n} \mathbb{E}_z \mathrm{loss}(z, \hat{z}) \;\ge\; \sum_{i \in T^c} \inf_{\hat{z}_i} \frac{1}{|\tilde{\mathcal{Z}}|} \sum_{z \in \tilde{\mathcal{Z}}} \mathbb{P}_z(\hat{z}_i \ne z_i). \tag{A.2}$$

Fix $i \in T^c$ and define $\tilde{\mathcal{Z}}_a^{(i)} = \{z \in \tilde{\mathcal{Z}} \colon z_i = a\}$ for $a \in \{1,2\}$. We observe that $\tilde{\mathcal{Z}}_1^{(i)} \cup \tilde{\mathcal{Z}}_2^{(i)} = \tilde{\mathcal{Z}}$ and that $\tilde{\mathcal{Z}}_1^{(i)} \cap \tilde{\mathcal{Z}}_2^{(i)} = \emptyset$. Let $f \colon \tilde{\mathcal{Z}}_1^{(i)} \to \tilde{\mathcal{Z}}_2^{(i)}$ such that for any $z \in \tilde{\mathcal{Z}}_1^{(i)}$ we have $f(z) \in \tilde{\mathcal{Z}}_2^{(i)}$ defined by

$$(f(z))_j \;=\; \begin{cases} z_j & \text{if } j \neq i, \\ 2 & \text{if } j = i. \end{cases}$$

The function f defines a one-to-one mapping from $\tilde{\mathcal{Z}}_1^{(i)}$ to $\tilde{\mathcal{Z}}_2^{(i)}$. Because these two sets partition $\tilde{\mathcal{Z}}$, we have $|\tilde{\mathcal{Z}}_1^{(i)}| = 2^{-1}|\tilde{\mathcal{Z}}|$. Moreover,

$$\begin{aligned}
\inf_{\hat{z}_i} \frac{1}{|\tilde{\mathcal{Z}}|} \sum_{z \in \tilde{\mathcal{Z}}} \mathbb{P}_z(\hat{z}_i \neq z_i) \;&=\; \inf_{\hat{z}_i} \frac{1}{|\tilde{\mathcal{Z}}|} \Big(\sum_{z \in \tilde{\mathcal{Z}}_1^{(i)}} \mathbb{P}_z\left(\hat{z}_i \neq 1\right) + \sum_{z \in \tilde{\mathcal{Z}}_2^{(i)}} \mathbb{P}_z\left(\hat{z}_i \neq 2\right) \Big) \\
&=\; \inf_{\hat{z}_i} \frac{1}{|\tilde{\mathcal{Z}}|} \sum_{z \in \tilde{\mathcal{Z}}_1^{(i)}} \left(\mathbb{P}_z\left(\hat{z}_i \neq 1\right) + \mathbb{P}_{f(z)}\left(\hat{z}_i \neq 2\right) \right) \\
&\geq\; \frac{1}{|\tilde{\mathcal{Z}}|} \sum_{z \in \tilde{\mathcal{Z}}_1^{(i)}} \inf_{\hat{z}_i} \left(\mathbb{P}_z\left(\hat{z}_i \neq 1\right) + \mathbb{P}_{f(z)}\left(\hat{z}_i \neq 2\right) \right). \qquad \text{(A.3)}
\end{aligned}$$

We are now reduced to the problem of estimating $\hat{z}_i$, and the best estimator for this task is

$$\hat{z}_i^{\mathrm{MLE}} \;=\; \begin{cases} 1 & \text{if } f_1(X_i) > f_2(X_i), \\ 2 & \text{otherwise.} \end{cases}$$

In other words, we are in the setting of Lemma 2, where we observe a single sample X_i and want to discriminate between $H_0 \colon X_i \sim f_1$ and $H_1 \colon X_i \sim f_2$. Because $\mathrm{Chernoff}(\mathcal{F}) = \omega(\log k)$, Lemma 2 ensures that

$$\inf_{\hat{z}_i} \left(\mathbb{P}_z\left(\hat{z}_i \neq 1\right) + \mathbb{P}_{f(z)}\left(\hat{z}_i \neq 2\right) \right) \;\geq\; e^{-(1+o(1))\mathrm{Chernoff}(f_1,f_2)},$$

and thus

$$\inf_{\hat{z}_i} \frac{1}{|\tilde{\mathcal{Z}}|} \sum_{z \in \tilde{\mathcal{Z}}} \mathbb{P}_z(\hat{z}_i \neq z_i) \;\geq\; \frac{|\tilde{\mathcal{Z}}_1|}{|\tilde{\mathcal{Z}}|} e^{-(1+o(1))\mathrm{Chernoff}(f_1,f_2)} \;=\; \frac{1}{2} e^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})}.$$

Going back to inequality with (A.2) leads to

$$\begin{aligned}
\inf_{\hat{z}} \sup_{z \in [k]^n} \mathbb{E}_z \mathrm{loss}(z, \hat{z}) \;&\geq\; \sum_{i \in T^c} \inf_{\hat{z}_i} \frac{1}{|\tilde{\mathcal{Z}}|} \sum_{z \in \tilde{\mathcal{Z}}} \mathbb{P}_z(\hat{z}_i \neq z_i) \\
&\geq\; \frac{|T^c|}{2} e^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})} \\
&=\; \frac{\alpha n}{5k} e^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})},
\end{aligned}$$

where the last line uses $|T^c| = \frac{2\alpha n}{5k}$. We finish the proof by using the assumption $\mathrm{Chernoff}(\mathcal{F}) = \omega(\log k)$.

Appendix B. Proofs of Section 3.1

B.1. Decomposition of the error

We first notice that

$$\operatorname{loss}\left(z^*, \hat{z}^{(t)}\right) \;=\; \sum_{i\in[n]} \mathbb{1}\left\{z_i^{(t)} \neq z_i^*\right\} \;=\; \sum_{i\in[n]} \sum_{b\in[k]\backslash\{z_i^*\}} \mathbb{1}\left\{z_i^{(t)} = b\right\}. \tag{B.1}$$

Combining

$$\mathbb{1}\left\{z_i^{(t)} = b\right\} = \left(\mathbb{1}\left\{z_i^{(t)} = b\right\}\right)^2$$

and

$$\mathbb{1}\left\{z_i^{(t)} = b\right\} \;=\; \mathbb{1}\left\{\forall a \in [k]\backslash\{b\} \colon \hat{\ell}_b^{(t)}(X_i) > \hat{\ell}_a^{(t)}(X_i)\right\} \;\le\; \mathbb{1}\left\{\hat{\ell}_b^{(t)}(X_i) > \hat{\ell}_{z_i^*}^{(t)}(X_i)\right\}$$

with (B.1), we obtain

$$\operatorname{loss}\left(z^*, \hat{z}^{(t)}\right) \;\le\; \sum_{i\in[n]} \sum_{b\in[k]\backslash\{z_i^*\}} \mathbb{1}\left\{z_i^{(t)} = b\right\} \mathbb{1}\left\{\hat{\ell}_b^{(t)}(X_i) > \hat{\ell}_{z_i^*}^{(t)}(X_i)\right\}. \tag{B.2}$$

Let us study the term $\mathbb{1}\left\{\hat{\ell}_b^{(t)}(X_i) > \hat{\ell}_{z_i^*}^{(t)}(X_i)\right\}$. For any $\delta > 0$, we have

$$\begin{aligned}
&\mathbb{1}\left\{\hat{\ell}_b^{(t)}(X_i) > \hat{\ell}_{z_i^*}^{(t)}(X_i)\right\} \\
&\le \mathbb{1}\left\{\ell_{z_i^*}(X_i) - \ell_b(X_i) < \delta\right\} + \mathbb{1}\left\{\delta < \hat{\ell}_b^{(t)}(X_i) - \ell_b(X_i) + \ell_{z_i^*}(X_i) - \hat{\ell}_{z_i^*}^{(t)}(X_i)\right\}.
\end{aligned}$$

Using $\mathbb{1}\{1 \le x + y\} \le \mathbb{1}\{1 \le |x| + |y|\} \le |x| + |y|$ for any $x, y \in \mathbb{R}$, we can further upper-bound the two terms appearing in the right-hand side of the last inequality by

$$\begin{aligned}
&\mathbb{1}\left\{\delta < \hat{\ell}_b^{(t)}(X_i) - \ell_b(X_i) + \ell_{z_i^*}(X_i) - \hat{\ell}_{z_i^*}^{(t)}(X_i)\right\} \\
&\le\; \delta^{-1}\left(\left|\hat{\ell}_b^{(t)}(X_i) - \ell_b(X_i)\right| + \left|\ell_{z_i^*}(X_i) - \hat{\ell}_{z_i^*}^{(t)}(X_i)\right|\right) \\
&\le\; 2\delta^{-1} \max_{a\in[k]} \left|\hat{\ell}_a^{(t)}(X_i) - \ell_a(X_i)\right|.
\end{aligned}$$

Therefore, we obtain using Expression (B.2)

$$\operatorname{loss}\left(z^*, \hat{z}^{(t)}\right) \;\le\; \xi_{\text{ideal}}(\delta) + \xi_{\text{excess}}^{(t)}(\delta),$$

where

$$\begin{aligned}
\xi_{\text{ideal}}(\delta) &\;=\; \sum_{i\in[n]} \sum_{b\in[k]\backslash\{z_i^*\}} \mathbb{1}\left\{\ell_{z_i^*}(X_i) - \ell_b(X_i) < \delta\right\}, \\
\xi_{\text{excess}}^{(t)}(\delta) &\;=\; 2\delta^{-1} \sum_{i\in[n]} \sum_{b\in[k]\backslash\{z_i^*\}} \mathbb{1}\left\{z_i^{(t)} = b\right\} \max_{a\in[k]} \left|\hat{\ell}_a^{(t)}(X_i) - \ell_a(X_i)\right|.
\end{aligned}$$

B.2. Study of the ideal error

Proof [Proof of Lemma 3] Taking the expectations in (3.4), we have

$$\mathbb{E}\xi_{\text{ideal}}(\delta) \;=\; \sum_{i\in[n]}\sum_{b\in[k]\setminus\{z_i^*\}} \mathbb{P}\left(\ell_b(X_i)-\ell_{z_i^*}(X_i)>\delta\right).$$

Chernoff's bound (see the proof of Lemma 2) yields that

$$\mathbb{P}\left(\ell_b(X_i)-\ell_{z_i^*}(X_i)>\delta\right) \;\le\; e^{\delta-(1+o(1))\mathrm{Chernoff}(f_{z_i^*},f_b)},$$

and therefore

$$\mathbb{E}\xi_{\text{ideal}}(\delta) \;\le\; nke^{\delta-(1+o(1))\mathrm{Chernoff}(\mathcal{F})}.$$

Now, because $\delta = o\left(\mathrm{Chernoff}(\mathcal{F})\right)$, Markov's inequality implies that

$$\mathbb{P}\left(\xi_{\text{ideal}}(\delta) > \mathbb{E}\xi_{\text{ideal}}(\delta)e^{\sqrt{\mathrm{Chernoff}(\mathcal{F})}}\right) \;\le\; e^{-\sqrt{\mathrm{Chernoff}(\mathcal{F})}}.$$

Therefore with probability of at least $1-e^{-\sqrt{\mathrm{Chernoff}(\mathcal{F})}}$ we have

$$\xi_{\text{ideal}}(\delta) \;\le\; \mathbb{E}\xi_{\text{ideal}}(\delta)e^{\sqrt{\mathrm{Chernoff}(\mathcal{F})}} \;\le\; nke^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})},$$

because $\mathrm{Chernoff}(\mathcal{F}) = \omega(1)$. ■

B.3. Proof of Lemma 4

Proof [Proof of Lemma 4] The assumption $\mathrm{Chernoff}(\mathcal{F}) = \omega(\log(k^2\tau))$ combined with $\mathrm{loss}(z^*,\hat{z}^{(0)}) = o(nk^{-1}\tau)$ and Condition 1 ensures that $\mathrm{loss}(z^*,\hat{z}^{(t)}) = o(nk^{-1}\tau)$ for every $t\ge 0$. Therefore,

$$\begin{aligned} n^{-1}\mathrm{loss}\left(z^*,z^{(t)}\right) \;&\le\; \left(\sum_{\tau=0}^{t}\left(\frac{c'}{1-c}\right)^{\tau}\right)e^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})} + \left(\frac{c'}{1-c}\right)^{t} \\ &\le\; e^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})} + \left(\frac{c'}{1-c}\right)^{t}, \end{aligned}$$

where the second inequality holds because $\sum_{\tau=0}^{\infty}\left(\frac{c'}{1-c}\right)^{\tau} = O(1)$ and $\mathrm{Chernoff}(\mathcal{F}) = \omega(1)$. Because $n^{-1}\mathrm{loss}\left(z^*,z^{(t)}\right)$ takes value in the set $\{jn^{-1}, j\in\{0,\cdots,n\}\}$, the term $\left(\frac{c'}{1-c}\right)^{t}$ is negligible if $\left(\frac{c'}{1-c}\right)^{t} = o(n^{-1})$, which occurs whenever $t \ge \left\lfloor \log\left(\frac{1-c}{c'}\right)\log n\right\rfloor$. ■

Appendix C. Mean and scale estimations of sub-exponential random variables

C.1. Large deviations of a sum of sub-exponential random variables

Let $Y_1, \cdots, Y_n$ be independent, zero-mean, sub-exponential random variables. For any subset of indices $S \subseteq [n]$, Bernstein's inequality (Vershynin, 2018, Theorem 2.8.2) ensures that

$$\mathbb{P}\left(\left| \frac{1}{\sqrt{|S|}} \sum_{i \in S} Y_i \right| \geq t \right) \;\leq\; \begin{cases} 2\exp(-ct^2) & \text{if } t < C\sqrt{|S|}, \\ 2\exp(-t\sqrt{|S|}) & \text{if } t \geq C\sqrt{|S|}. \end{cases} \tag{C.1}$$

for some positive constants c and C. The following lemma establishes a uniform upper bound on the quantity $\frac{1}{\sqrt{|S|}} \sum_{i \in S} Y_i$ over all the sets $S \subset [n]$ of size smaller than s.

Lemma 9 *Let $Y_1, \cdots, Y_n$ be independent, zero-mean, sub-exponential random variables. Let C be the constant in (C.1). For any $s = \omega(\log^6 n)$, we have*

$$\max_{\substack{S \subset [n] \\ |S| \leq s}} \frac{1}{\sqrt{|S|}} \left| \sum_{i \in S} Y_i \right| \;\leq\; C\sqrt{s}$$

with probability at least $1 - 6e^{-C\sqrt{s}/2}$.

Proof We will use (C.1) with $t = C\sqrt{s}$. By a union bound, we have

$$\begin{aligned} \mathbb{P}\left(\max_{\substack{S \subset [n] \\ |S| \leq s}} \frac{1}{\sqrt{|S|}} \left| \sum_{i \in S} Y_i \right| \geq t \right) \;&\leq\; \sum_{\ell=1}^{\lfloor s \rfloor} \mathbb{P}\left(\max_{S \subset [n] \colon |S| = \ell} \frac{1}{\sqrt{|S|}} \left| \sum_{i \in S} Y_i \right| \geq t \right) \\ &\leq\; \sum_{\ell=1}^{\lfloor s \rfloor} \binom{n}{\ell} 2\exp\left(-t\sqrt{\ell}\right). \end{aligned}$$

For any integers a, b verifying $1 \leq a < b \leq n$, we define

$$S_n(a,b) \;=\; \sum_{\ell=a}^{b} \binom{n}{\ell} 2\exp(-t\sqrt{\ell}).$$

Let us chose a sequence β verifying $\beta = \omega(\log^2 n)$ and $\beta^2 \log^2 n = o(s)$. Such a choice is possible because $s = \omega(\log^6 n)$[3]. We then split the sum $S_n(1, \lfloor s \rfloor)$ into three terms as follows:

$$S_n(1, \lfloor s \rfloor) \;=\; S_n\left(1, \left\lfloor \sqrt{s/\beta} \right\rfloor\right) + S_n\left(\left\lfloor \sqrt{s/\beta} \right\rfloor + 1, \left\lfloor \sqrt{s\beta} \right\rfloor\right) + S_n\left(\left\lfloor \sqrt{s\beta} \right\rfloor + 1, \lfloor s \rfloor\right),$$

and we show that each of these three terms is less than $2e^{-C\sqrt{s}/2}$.

For ease of notation, we drop the $\lfloor \cdot \rfloor$. We also recall the inequality $\binom{n}{\ell} \leq \left(\frac{en}{\ell}\right)^{\ell}$, yielding $\binom{n}{\ell} \leq (en)^{\sqrt{s/\beta}}$ for all $\ell \leq \sqrt{s/\beta}$.

3. Since $s = \omega(\log^6 n)$, there exists a diverging sequence $\omega_n = \omega(1)$ such that $s = \omega_n \log^6 n$. We then let $\beta = \omega_n^{1/4} \log^2 n$.

(i) Recalling that $t = C\sqrt{s}$, we have for all $1 \le \ell \le \sqrt{s/\beta}$

$$\binom{n}{\ell} 2\exp(-t\sqrt{\ell}) \;\le\; 2e^{-C\sqrt{s}}\,(en)^{\sqrt{s/\beta}}\,,$$

and therefore

$$\begin{aligned} S_n\left(1, \sqrt{s/\beta}\right) \;&\le\; 2e^{-C\sqrt{s}}\sqrt{\frac{s}{\beta}}\,(en)^{\sqrt{s/\beta}} \\ &=\; 2e^{-C\sqrt{s}\left(1 - \frac{\log(s/\beta)}{2C\sqrt{s}} - \frac{\log(en)}{C\sqrt{\beta}}\right)}. \end{aligned}$$

Because $\log n = o(\sqrt{\beta})$ (since $\beta = \omega(\log^2 n)$) and $\log(s/\beta) = o(\sqrt{s})$ (since $\beta^2 \log^2 n = o(s)$), we have for n large enough,

$$S_n\left(1, \sqrt{s/\beta}\right) \;\le\; 2e^{-\frac{C}{2}\sqrt{s}}.$$

(ii) Similarly,

$$\begin{aligned} S_n\left(\sqrt{s/\beta}+1, \sqrt{s\beta}\right) \;&\le\; 2e^{-Cs\beta^{-1/2}}\sqrt{s\beta}\,(en)^{\sqrt{s\beta}} \\ &=\; 2e^{-C\frac{s}{\sqrt{\beta}}\left(1 - \frac{\sqrt{\beta}\log(s\beta)}{2Cs} - \frac{\beta\log(en)}{C\sqrt{s}}\right)}. \end{aligned}$$

With our choice of β, we have $\sqrt{\beta}\log(s\beta) = o(s)$ and $\beta\log n = o(\sqrt{s})$, and therefore for n large enough,

$$S_n\left(\sqrt{s/\beta}+1, \sqrt{s\beta}\right) \;\le\; 2e^{-\frac{C}{2}\frac{s}{\sqrt{\beta}}} \;\le\; 2e^{-\frac{C}{2}\sqrt{s}},$$

because $\sqrt{s/\beta} = \omega(1)$.

(iii) Finally,

$$S_n\left(\sqrt{s\beta}+1, s\right) \;\le\; 2e^{-Cs\beta^{1/2}}s\,(en)^s \;\le\; 2e^{-Cs\beta^{1/2}\left(1 - \frac{\log s}{Cs\sqrt{\beta}} - \frac{\log(en)}{C\sqrt{\beta}}\right)} \;\le\; 2e^{-\frac{C}{2}\sqrt{s}}$$

for n large enough. This concludes the proof. ■

C.2. Quality of the estimates $\hat{\mu}_{a\ell}$ and $\hat{\sigma}_{a\ell}$

In this section, we upper bound the quantity $\|\hat{\mu}_a - \mu_a\|$. For any cluster labeling $z \in [k]^n$ and cluster $a \in [k]$, let $\Gamma_a(z) = \{i \in [n] \colon z_i = a\}$. Recall that the empirical location $\hat{\mu}_{a\ell}(z)$ and scale $\hat{\sigma}_{a\ell}(z)$ estimated from z are defined by

$$\begin{aligned} \hat{\mu}_{a\ell}(z) \;&=\; \frac{1}{|\Gamma_a(z)|}\sum_{i\in n} \mathbb{1}\{z_i = a\} X_{i\ell} \\ \hat{\sigma}_{a\ell}(z) \;&=\; \frac{1}{|\Gamma_a(z)|}\sum_{i\in n} \mathbb{1}\{z_i = a\} \left| X_{i\ell} - \hat{\mu}_{a\ell}(z) \right|. \end{aligned}$$

Finally, let $\bar{\mu}_{a\ell}(z) = \mathbb{E}\hat{\mu}_{a\ell}(z)$ and $\xi_{a\ell}(z) = \hat{\mu}_{a\ell}(z) - \bar{\mu}_{a\ell}(z)$.

Proposition 10 *Let* $\Delta_{\mu,\infty} = \max_{\substack{a,b\in[k] \\ \ell\in[d]}} |\mu_{a\ell} - \mu_{b\ell}|$. *Assume* $\min_{a\in[k]} |\Gamma_a(z^*)| \ge \alpha n k^{-1}$ *for some constant* $\alpha > 0$. *For* $z \in [k]^n$ *and some constant* C, *define*

$$\mathcal{E}_\mu(z) = \left\{ \max_{\substack{a\in[k] \\ \ell\in[d]}} |\hat{\mu}_{a\ell}(z) - \mu_{a\ell}| \le \frac{2k\Delta_{\mu,\infty}}{\alpha n} \mathrm{Ham}(z^*, z) + C' \sqrt{\frac{k\mathrm{Ham}(z^*, z)}{n}} + \sqrt{\frac{k}{n}} \right\}$$

with $C' = 2\sqrt{\frac{2}{\alpha}} \left(C + \sqrt{2}\right)$. *There exists a constant* $C > 0$ *such that the event* $\bigcap_{\mathrm{Ham}(z^*,z)\le\frac{\alpha n}{2k}} \mathcal{E}_\mu(z)$ *holds with probability at least* $1 - kd\left(4e^{-\alpha\sqrt{\frac{n}{k}}} + 6e^{-\frac{C}{2}\sqrt{\frac{\alpha n}{2k}}}\right)$.

Proof Let $a \in [k]$ and $\ell \in [d]$. A first triangle inequality leads to

$$|\hat{\mu}_{a\ell}(z) - \mu_{a\ell}| \le |\hat{\mu}_{a\ell}(z) - \hat{\mu}_{a\ell}(z^*)| + |\hat{\mu}_{a\ell}(z^*) - \mu_{a\ell}|. \tag{C.2}$$

Moreover, another triangle inequality yields that

$$\begin{aligned} |\hat{\mu}_{a\ell}(z) - \hat{\mu}_{a\ell}(z^*)| &= |\hat{\mu}_{a\ell}(z) - \bar{\mu}_{a\ell}(z) + \bar{\mu}_{a\ell}(z) - \bar{\mu}_{a\ell}(z^*) + \bar{\mu}_{a\ell}(z^*) - \hat{\mu}_{a\ell}(z^*)| \\ &\le |\bar{\mu}_{a\ell}(z) - \bar{\mu}_{a\ell}(z^*)| + |\xi_{a\ell}(z) - \xi_{a\ell}(z^*)|. \end{aligned} \tag{C.3}$$

Therefore, combining (C.2) and (C.3) gives

$$|\hat{\mu}_{a\ell}(z) - \mu_{a\ell}| \le |\hat{\mu}_{a\ell}(z^*) - \mu_{a\ell}| + |\bar{\mu}_{a\ell}(z) - \bar{\mu}_{a\ell}(z^*)| + |\xi_{a\ell}(z) - \xi_{a\ell}(z^*)|. \tag{C.4}$$

We will now upper–bound separately the three terms appearing on the right-hand side of (C.4).

(i) Bounding $|\hat{\mu}_{a\ell}(z^*) - \mu_{a\ell}|$. Let $\epsilon_{i\ell} = X_{i\ell} - \mu_{z_i^*\ell}$. We have

$$|\hat{\mu}_{a\ell}(z^*) - \mu_{a\ell}| = \frac{1}{|\Gamma_a(z^*)|} \left| \sum_{i\in\Gamma_a(z^*)} \epsilon_{i\ell} \right|.$$

By concentration of sub-exponential random variables (see Equation (C.1)), we have

$$\mathbb{P}\left(\frac{1}{|\Gamma_a(z^*)|} \left| \sum_{i\in[n]} \mathbb{1}\{z_i^* = a\}\epsilon_{i\ell} \right| \ge t \right) \le 2e^{-t|\Gamma_a(z^*)|}.$$

Because $|\Gamma_a(z^*)| \ge \alpha n/k$, this implies that with $t = \sqrt{k}$

$$\frac{1}{|\Gamma_a(z^*)|} \left| \sum_{i\in[n]} \mathbb{1}\{z_i^* = a\}\epsilon_{i\ell} \right| \le \sqrt{\frac{k}{n}} \tag{C.5}$$

with probability at least $1 - 2e^{-\alpha\sqrt{n/k}}$. Therefore, a union bound over $a \in [k]$ and $\ell \in [d]$ ensures that

$$\max_{\substack{a\in[k] \\ \ell\in[d]}} |\hat{\mu}_{a\ell}(z^*) - \mu_{a\ell}| \le \sqrt{\frac{k}{n}}$$

with probability at least $1 - 2dke^{-\alpha\sqrt{n/k}}$.

(ii) Bounding $|\bar{\mu}_{a\ell}(z) - \bar{\mu}_{a\ell}(z^*)|$. Since $\bar{\mu}_{a\ell}(z^*) = \mu_{a\ell}$ and $\sum_{i\in[n]} \sum_{b\in[k]} \mathbb{1}\{z_i = a, z_i^* = b\} = |\Gamma_a(z)|$,

$$\bar{\mu}_{a\ell}(z) - \bar{\mu}_{a\ell}(z^*) \; = \; \frac{1}{|\Gamma_a(z)|} \sum_{i\in[n]} \sum_{b\in[k]} \mathbb{1}\{z_i = a, z_i^* = b\} \left(\mu_{b\ell} - \mu_{a\ell}\right).$$

Hence, using $|\Gamma_a(z)| \geq 2^{-1}\alpha n k^{-1}$ (Lemma 12), we obtain

$$\begin{aligned} |\bar{\mu}_{a\ell}(z) - \bar{\mu}_{a\ell}(z^*)| \; &\leq \; \frac{1}{|\Gamma_a(z)|} \sum_{i\in[n]} \sum_{b\in[k]\setminus\{a\}} \mathbb{1}\{z_i = a, z_i^* = b\} \left|\mu_{b\ell} - \mu_{a\ell}\right| \\ &\leq \; \frac{2k}{\alpha n} \Delta_{\mu,\infty} \mathrm{Ham}(z^*, z). \end{aligned}$$

(iii) Bounding $|\xi_{a\ell}(z) - \xi_{a\ell}(z^*)|$. This upper-bound is more complex to derive, and is computed in Lemma 13, which establishes that for any z verifying $\mathrm{Ham}(z^*, z) \leq 2^{-1}\alpha n k^{-1}$,

$$\max_{\substack{a\in[k] \\ \ell\in[d]}} |\xi_{a\ell}(z) - \xi_{a\ell}(z^*)| \; \leq \; 2\sqrt{\frac{2}{\alpha}} \left(C + \sqrt{2}\right) \sqrt{\frac{k\mathrm{Ham}(z^*, z)}{n}}$$

holds with probability at least $1 - 6dke^{-\frac{C}{2}\sqrt{\frac{\alpha n}{2k}}} - 2dke^{-\alpha\sqrt{\frac{n}{k}}}$.

We conclude the proof by combining the upper-bounds obtained in steps (i), (ii) and (iii) with the decomposition (C.4). ■

Proposition 11 *Let* $\Delta_{\mu,\infty} = \max_{a,b\in[k]} \|\mu_a - \mu_b\|_\infty$ *and* $\Delta_{\sigma,\infty} = \max_{a,b\in[k]} \|\sigma_a - \sigma_b\|_\infty$. *Assume* $\min_{a\in[k]} |\Gamma_a(z^*)| \geq \alpha n k^{-1}$ *for some constant* $\alpha > 0$. *For* $z \in [k]^n$ *and some constant* C*, define*

$$\mathcal{E}_\sigma(z) \; = \; \left\{ \max_{\substack{a\in[k] \\ \ell\in[d]}} |\hat{\sigma}_{a\ell}(z) - \sigma_{a\ell}| \; \leq \; 2\sqrt{\frac{k}{n}} + \frac{2k(\Delta_{\mu,\infty} + \Delta_{\sigma,\infty})}{\alpha n} \mathrm{Ham}(z^*, z) + 2C'\sqrt{\frac{k\mathrm{Ham}(z^*, z)}{n}} \right\}.$$

where $C' = 2(1 + C\sqrt{2\alpha^{-1}})$. *There exists a constant* $C > 0$ *such that the event* $\bigcap_{\mathrm{Ham}(z^*,z)\leq 2^{-1}\alpha n k^{-1}} \mathcal{E}_\sigma(z)$ *holds with probability at least* $1 - 2kd\left(4e^{-\alpha\sqrt{\frac{n}{k}}} + 6e^{-\frac{C}{2}\sqrt{\frac{\alpha n}{2k}}}\right)$.

Proof We compute, using the triangle inequality, that

$$\begin{aligned} |\hat{\sigma}_{a\ell}(z) - \sigma_{a\ell}| \; &= \; \left| \frac{1}{|\Gamma_a(z)|} \sum_{i\in\Gamma_a(z)} |X_{i\ell} - \hat{\mu}_{a\ell}| - \sigma_{a\ell} \right| \\ &\leq \; \frac{1}{|\Gamma_a(z)|} \left| \sum_{i\in\Gamma_a(z)} \left(|X_{i\ell} - \mu_{a\ell}| - \sigma_{a\ell}\right) \right| + |\hat{\mu}_{a\ell}(z) - \mu_{a\ell}|. \end{aligned}$$

Proposition 10 provides an upper bound on the second term of the right-hand side of the last inequality, $|\hat{\mu}_{a\ell}(z) - \mu_{a\ell}|$.

It also provides an upper bound on the first term of the right-hand side of the last inequality. Indeed, let $\tilde{X}_{i\ell} = |X_{i\ell} - \mu_{a\ell}|$ and $\tilde{\sigma}_{a\ell}(z) = \frac{1}{|\Gamma_a(z)|} \sum_{i \in \Gamma_a(z)} \tilde{X}_{i\ell}$. The random variables $\tilde{X}_{i\ell}$ are sub-exponential, and $\tilde{\sigma}_{a\ell}(z)$ is the sample mean computed over the subset $\{\tilde{X}_{i\ell}, i \in \Gamma_a(z)\}$. Therefore we can again apply Proposition 10 to show that

$$\max_{\substack{a \in [k] \\ \ell \in [d]}} |\tilde{\sigma}_{a\ell}(z) - \sigma_{a\ell}| \leq \sqrt{\frac{k}{n}} + \frac{2k\Delta_{\sigma,\infty}}{\alpha n} \mathrm{Ham}(z^*, z) + C' \sqrt{\frac{k\mathrm{Ham}(z^*, z)}{n}}$$

with probability at least $1 - kd \left(4e^{-\alpha\sqrt{\frac{n}{k}}} + 6e^{-\frac{C}{2}\sqrt{\frac{\alpha n}{2k}}} \right)$. ■

C.3. Additional technical lemmas

Lemma 12 *Let $z^* \in [k]^n$ such that $\min_{a \in [k]} \sum_{i \in [n]} |\Gamma_a(z^*)| \geq \alpha n/k$ for all $a \in [k]$ and for some $\alpha > 0$. Let $z \in [k]^n$ such that $\mathrm{Ham}(z, z') \leq \alpha n/(2k)$. Then*

$$\sum_{i=1}^{n} \mathbb{1}\{z_i = a \cap z_i^* = a\} \geq \frac{\alpha n}{2k}.$$

In particular, $\sum_{i=1}^{n} \mathbb{1}\{z_i = a\} \geq \alpha n/(2k)$.

Proof We have

$$\begin{aligned} \sum_{i=1}^{n} \mathbb{1}\{z_i = a \cap z_i^* = a\} &= \sum_{i=1}^{n} \mathbb{1}\{z_i^* = a\} - \sum_{i=1}^{n} \mathbb{1}\{z_i^* = a \cap z_i \neq a\} \\ &\geq \sum_{i=1}^{n} \mathbb{1}\{z_i^* = a\} - \sum_{i=1}^{n} \mathbb{1}\{z_i^* \neq z_i\} \\ &\geq \frac{\alpha n}{k} - \frac{\alpha n}{2k} \\ &= \frac{\alpha n}{2k}. \end{aligned}$$

Finally, because $\sum_{i=1}^{n} \mathbb{1}\{z_i = a\} \geq \sum_{i=1}^{n} \mathbb{1}\{z_i = a \cap z_i^* = a\}$ we also have $\sum_{i=1}^{n} \mathbb{1}\{z_i = a\} \geq \alpha n/(2k)$. ■

Lemma 13 *Let $z^* \in [k]^n$ such that $\min_{a \in [k]} |\Gamma_a(z^*)| \geq \alpha n/k$ for some constant $\alpha > 0$. For any $z \in [k]^n$ and $C \geq 1$, we define the event*

$$\mathcal{E}(z) = \left\{ \max_{\substack{a \in [k] \\ \ell \in [d]}} |\xi_{a\ell}(z) - \xi_{a\ell}(z^*)| \leq 2\sqrt{\frac{2}{\alpha}} \left(C + \sqrt{2} \right) \sqrt{\frac{k\mathrm{Ham}(z^*, z)}{n}} \right\}.$$

There exists a constant $C \geq 1$ such that the event $\bigcap_{z \colon \mathrm{Ham}(z^*, z) \leq \frac{\alpha n}{2k}} \mathcal{E}(z)$ holds with probability at least $1 - 6dke^{-\frac{C}{2}\sqrt{\frac{\alpha n}{2k}}} - 2dke^{-\alpha\sqrt{\frac{n}{k}}}$.

Proof The random variables $\epsilon_{i\ell}$ are independent, zero-mean, and sub-exponential. Therefore Lemma 9 and a union bound ensure the existence of a constant $C > 0$ such that the event

$$\mathcal{E}_1 \;=\; \left\{ \max_{\ell \in [d]} \sup_{\substack{S \subset [n] \\ |S| \le \frac{\alpha n}{2k}}} \frac{1}{\sqrt{|S|}} \left| \sum_{i \in S} \epsilon_{i\ell} \right| \;\le\; C\sqrt{n} \right\} \tag{C.6}$$

holds with a probability of at least $1 - 6de^{-\frac{C}{2}\sqrt{\frac{\alpha n}{2k}}}$. Similarly, we have established in (C.5) that the event

$$\mathcal{E}_2(a) \;=\; \left\{ \max_{\ell \in [d]} \frac{1}{|\Gamma_a(z^*)|} \left| \sum_{i \in [n]} \mathbb{1}\{z_i^* = a\}\epsilon_{i\ell} \right| \;\le\; \sqrt{\frac{k}{n}} \right\},$$

where $a \in [k]$, holds with probability at least $1 - 2de^{-\alpha\sqrt{n/k}}$. Let $\mathcal{E}_2 = \cap_{a \in [k]} \mathcal{E}_2(a)$. We have

$$\mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \ge 1 - 6de^{-C/2\sqrt{\alpha n/(2k)}} - 2dke^{-\alpha\sqrt{n/k}}. \tag{C.7}$$

In the rest of the proof, we work conditionally on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, and will show that the event $\cap_{z\colon \mathrm{Ham}(z^*,z) \le \alpha n/(2k)} \mathcal{E}(z)$ holds. Let $z \in [k]^n$ verifying $\mathrm{Ham}(z^*, z) \le \alpha n/(2k)$ and let $a \in [k]$ and $\ell \in [d]$. We have

$$\begin{aligned}
&|\xi_{a\ell}(z) - \xi_{a\ell}(z^*)| \\
&= \left| \frac{\sum_{i\in[n]} \mathbb{1}\{z_i = a\}\epsilon_{i\ell}}{\sum_{i\in[n]} \mathbb{1}\{z_i = a\}} - \frac{\sum_{i\in[n]} \mathbb{1}\{z_i^* = a\}\epsilon_{i\ell}}{\sum_{i\in[n]} \mathbb{1}\{z_i^* = a\}} \right| \\
&\le \underbrace{\left| \frac{\sum_{i\in[n]} \mathbb{1}\{z_i = a\}\epsilon_{i\ell}}{\sum_{i\in[n]} \mathbb{1}\{z_i = a\}} - \frac{\sum_{i\in[n]} \mathbb{1}\{z_i^* = a\}\epsilon_{i\ell}}{\sum_{i\in[n]} \mathbb{1}\{z_i = a\}} \right|}_{E_1} + \underbrace{\left| \frac{\sum_{i\in[n]} \mathbb{1}\{z_i^* = a\}\epsilon_{i\ell}}{\sum_{i\in[n]} \mathbb{1}\{z_i = a\}} - \frac{\sum_{i\in[n]} \mathbb{1}\{z_i^* = a\}\epsilon_{i\ell}}{\sum_{i\in[n]} \mathbb{1}\{z_i^* = a\}} \right|}_{E_2}.
\end{aligned}$$

Let us first upper bound E_1. We have

$$\begin{aligned}
E_1 \;&=\; \frac{1}{|\Gamma_a(z)|} \left| \sum_{i\in[n]} \left(\mathbb{1}\{z_i = a\} - \mathbb{1}\{z_i^* = a\}\right)\epsilon_{i\ell} \right| \\
&\le\; \underbrace{\frac{1}{|\Gamma_a(z)|} \left| \sum_{i\in[n]} \mathbb{1}\{z_i = a, z_i^* \neq a\}\epsilon_{i\ell} \right|}_{E_{11}} + \underbrace{\frac{1}{|\Gamma_a(z)|} \left| \sum_{i\in[n]} \mathbb{1}\{z_i \neq a, z_i^* = a\}\epsilon_{i\ell} \right|}_{E_{12}}.
\end{aligned}$$

Let us denote by $\Gamma_a^c(z^*) = [n] \backslash \Gamma_a(z^*)$ the complement of $\Gamma_a(z^*)$. Noticing that $|\Gamma_a(z)| \ge \alpha n/(2k)$ (because of Lemma 12) and that

$$|\Gamma_a(z) \cap \Gamma_a^c(z^*)| \;=\; \sum_{i\in[n]} \mathbb{1}\{z_i = a, z_i^* \neq a\} \;\le\; \mathrm{Ham}(z^*, z) \;\le\; \frac{\alpha n}{2k},$$

we obtain

$$\begin{aligned} E_{11} &= \frac{1}{|\Gamma_a(z)|}\sqrt{|\Gamma_a(z)\cap\Gamma_a^c(z^*)|}\cdot\frac{\left|\sum_{i\in[n]}\mathbb{1}\{z_i=a,z_i^*\neq a\}\epsilon_{i\ell}\right|}{\sqrt{|\Gamma_a(z)\cap\Gamma_a^c(z^*)|}} \\ &\le \frac{2k}{\alpha n}\sqrt{\mathrm{Ham}(z^*,z)}\sup_{\substack{S\subset[n]\\|S|\le 2^{-1}\alpha n k^{-1}}}\frac{1}{\sqrt{|S|}}\left|\sum_{i\in S}\epsilon_{i\ell}\right|. \end{aligned}$$

Conditioning E_{11} on the event $\mathcal{E}_1$, we have therefore

$$E_{11} \le \frac{2k}{\alpha n}\sqrt{\mathrm{Ham}(z^*,z)}\cdot C\sqrt{\frac{\alpha n}{2k}} = C\sqrt{\frac{2k}{\alpha n}\mathrm{Ham}(z^*,z)}.$$

Proceeding similarly, we establish the same upper bound holds for E_{12}. Therefore, conditionally on $\mathcal{E}_1$, we have

$$E_1 \le 2C\sqrt{\frac{2k}{\alpha n}\mathrm{Ham}(z^*,z)}. \tag{C.8}$$

We can now upper-bound E_2, whose expression can be recast as

$$E_2 = \underbrace{\left|\frac{\sum_{i\in[n]}\mathbb{1}\{z_i^*=a\}-\mathbb{1}\{z_i=a\}}{\sum_{i\in[n]}\mathbb{1}\{z_i=a\}}\right|}_{E_{21}}\cdot\underbrace{\left|\frac{\sum_{i\in[n]}\mathbb{1}\{z_i^*=a\}\epsilon_{i\ell}}{\sum_{i\in[n]}\mathbb{1}\{z_i^*=a\}}\right|}_{E_{22}}.$$

We have

$$\begin{aligned} \left|\sum_{i\in[n]}\mathbb{1}\{z_i^*=a\}-\mathbb{1}\{z_i=a\}\right| &= \left|\sum_{i\in[n]}\mathbb{1}\{z_i^*=a,z_i\neq a\}-\sum_{i\in[n]}\mathbb{1}\{z_i^*\neq a,z_i=a\}\right| \\ &\le \max\left\{\left|\sum_{i\in[n]}\mathbb{1}\{z_i^*=a,z_i\neq a\}\right|,\left|\sum_{i\in[n]}\mathbb{1}\{z_i^*\neq a,z_i=a\}\right|\right\} \\ &\le \mathrm{Ham}(z^*,z). \end{aligned}$$

Moreover, Lemma 12 yields that $\sum_{i\in[n]}\mathbb{1}\{z_i=a\}\ge\sum_{i\in[n]}\mathbb{1}\{z_i=a,z_i^*=a\}\ge\alpha n/(2k)$. Therefore,

$$E_{21} \le \frac{2k}{\alpha n}\mathrm{Ham}(z^*,z).$$

Finally, conditionally on the event $\mathcal{E}_2$, we have $E_{22}\le\sqrt{k/n}$. Using $\sqrt{\mathrm{Ham}(z^*,z)}\le\sqrt{\alpha n/(2k)}$, we obtain

$$E_2 \le \frac{2k}{\alpha n}\sqrt{\frac{k}{n}}\mathrm{Ham}(z^*,z) \le \frac{4}{\sqrt{2\alpha}}\sqrt{\frac{k\mathrm{Ham}(z^*,z)}{n}}. \tag{C.9}$$

We conclude the proof by combining (C.8) and (C.9) and using (C.7). ■

Appendix D. Proofs for Section 3.2 (Laplace mixture models)

D.1. Proof of Theorem 5

Following the general result on recovering parametric mixture models (Lemma 4), to prove Theorem 5, we need to show that Condition 1 holds. We will show that it does so for arbitrary c, c' (which can be taken as small as we would like).

Because $\log f_{(\mu,\sigma)}(x) = \sum_{\ell=1}^{d}\left(-\log\sigma_\ell - \left|\frac{x-\mu_\ell}{\sigma_\ell}\right|\right)$, we have for any $a \in [k]$ and $X_i \in \mathbb{R}^d$,

$$\begin{aligned}\left|\hat{\ell}_a^{(t)}(X_i) - \ell_a(X_i)\right| &= \left|\sum_{\ell=1}^{d}\left(-\log\frac{\hat{\sigma}_{a\ell}^{(t)}}{\sigma_{a\ell}} - \left|\frac{X_{i\ell}-\hat{\mu}_{a\ell}^{(t)}}{\hat{\sigma}_{a\ell}^{(t)}}\right| + \left|\frac{X_{i\ell}-\mu_{a\ell}}{\sigma_{a\ell}}\right|\right)\right| \\ &\leq \sum_{\ell=1}^{d}\left(\left|\log\frac{\hat{\sigma}_{a\ell}^{(t)}}{\sigma_{a\ell}}\right| + \left|\left|\frac{X_{i\ell}-\mu_{a\ell}}{\sigma_{a\ell}}\right| - \left|\frac{X_{i\ell}-\hat{\mu}_{a\ell}^{(t)}}{\hat{\sigma}_{a\ell}^{(t)}}\right|\right|\right).\end{aligned}$$

Moreover, using $\big||x|-|y|\big| \leq |x-y|$, we have

$$\begin{aligned}\left|\left|\frac{X_{i\ell}-\mu_{a\ell}}{\sigma_{a\ell}}\right| - \left|\frac{X_{i\ell}-\hat{\mu}_{a\ell}^{(t)}}{\hat{\sigma}_{a\ell}^{(t)}}\right|\right| &\leq \frac{\left|\hat{\sigma}_{a\ell}^{(t)}\left(X_{i\ell}-\mu_{a\ell}\right) - \sigma_{a\ell}\left(X_{i\ell}-\hat{\mu}_{a\ell}^{(t)}\right)\right|}{\sigma_{a\ell}\hat{\sigma}_{a\ell}^{(t)}} \\ &\leq \frac{\left|\hat{\sigma}_{a\ell}^{(t)}-\sigma_{a\ell}\right|}{\sigma_{a\ell}\hat{\sigma}_{a\ell}^{(t)}}\cdot|X_{i\ell}-\mu_{a\ell}| + \frac{\left|\mu_{a\ell}-\hat{\mu}_{a\ell}^{(t)}\right|}{\hat{\sigma}_{a\ell}^{(t)}} \\ &\leq \frac{\left|\hat{\sigma}_{a\ell}^{(t)}-\sigma_{a\ell}\right|}{\sigma_{a\ell}\hat{\sigma}_{a\ell}^{(t)}}\cdot\left\{\left(\left|X_{i\ell}-\mu_{z_i^*\ell}\right|\right) + \left|\mu_{z_i^*\ell}-\mu_{a\ell}\right|\right\} + \frac{\left|\mu_{a\ell}-\hat{\mu}_{a\ell}^{(t)}\right|}{\hat{\sigma}_{a\ell}^{(t)}} \\ &\leq \frac{\left|\hat{\sigma}_{a\ell}^{(t)}-\sigma_{a\ell}\right|}{\sigma_{a\ell}\hat{\sigma}_{a\ell}^{(t)}}\cdot\left(\left|X_{i\ell}-\mu_{z_i^*\ell}\right| - \sigma_{z_i^*\ell}\right) \\ &\qquad + \frac{\left|\hat{\sigma}_{a\ell}^{(t)}-\sigma_{a\ell}\right|}{\hat{\sigma}_{a\ell}^{(t)}}\left(\frac{\Delta_{\mu,\infty}}{\sigma_{a\ell}} + \max_{b,c\in[k]}\frac{\sigma_{b\ell}}{\sigma_{c\ell}}\right) + \frac{\left|\mu_{a\ell}-\hat{\mu}_{a\ell}^{(t)}\right|}{\hat{\sigma}_{a\ell}^{(t)}}.\end{aligned}$$

Combining these upper bounds with the definition of the excess error in (3.5), we obtain

$$\xi_{\text{excess}}(\delta) \;\leq\; 2\delta^{-1}\left(F\cdot\mathrm{Ham}\left(z^*,\hat{z}^{(t)}\right) + G\right),$$

where F and G are defined by

$$\begin{aligned}\tilde{F} &= \max_{a\in[k]}\sum_{\ell=1}^{d}\left|\log\frac{\hat{\sigma}_{a\ell}^{(t)}}{\sigma_{a\ell}}\right| + \frac{\left|\mu_{a\ell}-\hat{\mu}_{a\ell}^{(t)}\right| + \left|\hat{\sigma}_{a\ell}^{(t)}-\sigma_{a\ell}\right|\left(\sigma_{a\ell}^{-1}\Delta_{\mu,\infty}+m_\sigma\right)}{\hat{\sigma}_{a\ell}^{(t)}} \\ \tilde{G} &= \max_{a\in[k]}\sum_{\ell=1}^{d}\frac{\left|\hat{\sigma}_{a\ell}^{(t)}-\sigma_{a\ell}\right|}{\sigma_{a\ell}\hat{\sigma}_{a\ell}^{(t)}}\cdot\sum_{i\in[n]}\left(\left|X_{i\ell}-\mu_{z_i^*\ell}\right| - \sigma_{z_i^*\ell}\right),\end{aligned}$$

where $\Delta_{\sigma,\infty} = \max_{\substack{1\leq a\neq b\leq k\\ \ell\in[d]}}|\sigma_{a\ell}-\sigma_{b\ell}|$ and $m_\sigma = \max_{1\leq a\neq b\leq k}\frac{\sigma_{a\ell}}{\sigma_{b\ell}}$.

The quantities $\tilde{F}$ and $\tilde{G}$ are in fact functions of the true parameters μ, σ as well as the estimated ones $\hat{\mu}^{(t)}, \hat{\sigma}^{(t)}$. Because these estimations are made based on the data points X and the predicted clusters are step $t-1$, we have $\tilde{F} = F(X, \mu, \sigma, \hat{z}^{(t-1)})$ and $\tilde{G} = G(X, \mu, \sigma, \hat{z}^{(t-1)})$, where for any $z \in [k]^n$, we define

$$F(X,\mu,\sigma,z) \;=\; \max_{a\in[k]} \sum_{\ell=1}^{d} \left| \log \frac{\hat{\sigma}_{a\ell}(z)}{\sigma_{a\ell}} \right| + \frac{|\mu_{a\ell} - \hat{\mu}_{a\ell}(z)| + |\hat{\sigma}_{a\ell}(z) - \sigma_{a\ell}| \left(\sigma_{a\ell}^{-1}\Delta_{\mu,\infty} + m_\sigma\right)}{\hat{\sigma}_{a\ell}(z)}$$

$$G(X,\mu,\sigma,z) \;=\; \max_{a\in[k]} \sum_{\ell=1}^{d} \frac{|\hat{\sigma}_{a\ell}(z) - \sigma_{a\ell}|}{\sigma_{a\ell}\hat{\sigma}_{a\ell}(z)} \cdot \sum_{i\in[n]} \left(\left| X_{i\ell} - \mu_{z_i^*\ell} \right| - \sigma_{z_i^*\ell} \right).$$

The quantities F and G are analyzed in Lemmas 14 and 15. In particular, under the assumptions of Theorem 5, we have (with probability at least $1-o(1)$)

$$\tilde{F} \;=\; o\left(d\Delta_{\mu,\infty}\right) \quad \text{and} \quad \tilde{G} \;=\; O\left(\omega_n d\sqrt{k}\left(1+\frac{\Delta_{\mu,\infty}}{\sqrt{nk^{-1}}}\right)\right) \times \max\left\{\mathrm{Ham}\left(z^*, z^{(t-1)}\right), 1\right\},$$

where we are free to choose the sequence ω_n as long as $\omega_n = \omega(1)$.

Suppose $z^{(t-1)} \neq z^*$. We choose δ such that $\delta = o(\mathrm{Chernoff}(\mathcal{F}))$ and $\omega_n d\sqrt{k}(1+\frac{\Delta_{\mu,\infty}}{\sqrt{nk^{-1}}}) = o(\delta)$. Such a choice is possible. Indeed, because by assumption $\mathrm{Chernoff}(\mathcal{F}) = \omega(d\sqrt{k}(1+\frac{\Delta_{\mu,\infty}}{\sqrt{nk^{-1}}}))$, we can write $\mathrm{Chernoff}(\mathcal{F}) = d\sqrt{k}(1+\frac{\Delta_{\mu,\infty}}{\sqrt{nk^{-1}}})\tau_n$ with $\tau_n = \omega(1)$. Then, we can choose $\omega_n = \tau_n^{1/4}$ and $\delta = d\sqrt{k}(1+\frac{\Delta_{\mu,\infty}}{\sqrt{nk^{-1}}})\sqrt{\tau_n}$. With this particular choice of δ, we have $\xi_{\text{excess}}^{(t)}(\delta) = o\left(\mathrm{Ham}\left(z^*, \hat{z}^{(t)}\right)\right) + o\left(\mathrm{Ham}\left(z^*, \hat{z}^{(t-1)}\right)\right)$. Hence,

$$\xi_{\text{excess}}^{(t)}(\delta) \;\le\; c\mathrm{Ham}\left(z^*, \hat{z}^{(t)}\right) + c'\mathrm{Ham}\left(z^*, \hat{z}^{(t-1)}\right)$$

for arbitrary constants $c, c' > 0$. This establishes Condition 1.

Finally, suppose that $z^{(t-1)} = z^*$. By choosing δ as in the previous paragraph, we have $\xi_{\text{excess}}^{(t)}(\delta) = o(\mathrm{Ham}(z^*, z^{(t)}) + o(1)$ and therefore

$$(1+o(1))\mathrm{Ham}(z^*, z^{(t)}) \;\le\; ne^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})} + o(1). \tag{D.1}$$

If $ne^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})}) = o(1)$ then $\mathrm{Ham}(z^*, z^{(t)}) = o(1)$ because $\mathrm{Ham}(z^*, z^{(t)})$ is integer. Otherwise, if $ne^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})}$ is bounded away from 0, then the $o(1)$ in the right hand side of (D.1) can be absorbed by the term $ne^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})}$. This implies $\mathrm{Ham}(z^*, z^{(t)}) \le ne^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})}$, and this ends the proof.

D.2. Bounding F and G

Lemma 14 *Suppose that* $\min_{a\in[k]} |\Gamma_a(z^*)| \ge \alpha n/k$ *for some* $\alpha > 0$. *Assume also that* $\sigma_{a\ell} = \Theta(1)$ *for all* $a \in [k], \ell \in [d]$. *Let* $\epsilon > 0$ *and* $\tau = o(nk^{-1}\Delta_{\mu,\infty}^{-1})$. *For* n *large enough, we have*

$$\max_{\substack{z\in[k]^n \\ \mathrm{Ham}(z^*,z)\le\tau}} F(X,\mu,\sigma,z) \;=\; o(d\Delta_{\mu,\infty})$$

with probability at least $1 - 3kd\left(4e^{-\alpha\sqrt{\frac{n}{k}}} + 6e^{-\frac{C}{2}\sqrt{\frac{\alpha n}{2k}}}\right)$.

Proof Let $\mathcal{Z}_\tau = \{z \in [k]^n : \mathrm{Ham}(z^*, z) \leq \tau\}$. In the rest of the proof, we work conditionally on the event $\mathcal{E} = \mathcal{E}_\mu \cap \mathcal{E}_\sigma$, where

$$\mathcal{E}_\mu = \left\{ \max_{\substack{z\in\mathcal{Z}_\tau \\ a\in k \\ \ell\in[d]}} |\hat{\mu}_{a\ell}(z) - \mu_{a\ell}| \leq \sqrt{\frac{k}{n}} + \frac{2k\Delta_{\mu,\infty}}{\alpha n}\tau + C'\sqrt{\frac{k\tau}{n}} \right\},$$

$$\mathcal{E}_\sigma = \left\{ \max_{\substack{z\in\mathcal{Z}_\tau \\ a\in k \\ \ell\in[d]}} |\hat{\sigma}_{a\ell}(z) - \sigma_{a\ell}| \leq 2\sqrt{\frac{k}{n}} + \frac{2k(\Delta_{\mu,\infty}+\Delta_{\sigma,\infty})}{\alpha n}\tau + 2C'\sqrt{\frac{k\tau}{n}} \right\}.$$

where $C' = 2(1+\sqrt{2\alpha^{-1}})$. By combining Proposition 10 and 11, the event $\mathcal{E} = \mathcal{E}_\mu \cap \mathcal{E}_\sigma$ holds with a probability of at least

$$\mathbb{P}(\mathcal{E}) \geq 1 - 3kd\left(4e^{-\alpha\sqrt{\frac{n}{k}}} + 6e^{-\frac{C}{2}\sqrt{\frac{\alpha n}{2k}}}\right).$$

We will show that, conditioned on this event $\mathcal{E}$, we have $\max\limits_{z\in\mathcal{Z}_\tau} F(X,\mu,\sigma,z) = o(d\Delta_{\mu,\infty})$. We first notice that

$$\max_{z\in\mathcal{Z}_\tau} F(X,\mu,\sigma,z) \leq d\left(\max_{\substack{z\in\mathcal{Z}_\tau \\ a\in[k] \\ \ell\in[d]}} \left|\log\frac{\hat{\sigma}_{a\ell}(z)}{\sigma_{a\ell}}\right| + \frac{|\mu_{a\ell} - \hat{\mu}_{a\ell}(z)|}{\hat{\sigma}_{a\ell}(z)} + \frac{\left|\hat{\sigma}_{a\ell}^{(t)} - \sigma_{a\ell}\right|}{\hat{\sigma}_{a\ell}^{(t)}} \cdot \left(\sigma_{a\ell}^{-1}\Delta_{\mu,\infty} + m_\sigma\right)\right) \tag{D.2}$$

Moreover, using the event $\mathcal{E}_\sigma$ and $\tau = o(nk^{-1})$, we have $|\hat{\sigma}_{a\ell} - \sigma_{a\ell}| = o(1)$ for any $z \in \mathcal{Z}_\tau$. Let n be large enough so that $|\hat{\sigma}_{a\ell} - \sigma_{a\ell}|\sigma_{a\ell}^{-1} \leq 2^{-1}$ (notice this large enough n does not depend on z). Using $|\log(1+t)| \leq \frac{|t|}{1-|t|}$ for any $|t| < 1$, we have

$$\left|\log\frac{\hat{\sigma}_{a\ell}(z)}{\sigma_{a\ell}}\right| = \left|\log\left(1 + \frac{\hat{\sigma}_{a\ell}(z) - \sigma_{a\ell}}{\sigma_{a\ell}}\right)\right| \leq 2\left|\frac{\hat{\sigma}_{a\ell}(z) - \sigma_{a\ell}}{\sigma_{a\ell}}\right|.$$

Because $\sigma_{a\ell} = \Theta(1)$, this ensures that

$$\max_{\substack{z\in\mathcal{Z}_\tau \\ a\in[k] \\ \ell\in[d]}} \left|\log\frac{\hat{\sigma}_{a\ell}(z)}{\sigma_{a\ell}}\right| = o(1). \tag{D.3}$$

Similarly,

$$\max_{\substack{z\in\mathcal{Z}_\tau \\ a\in[k] \\ \ell\in[d]}} \frac{|\mu_{a\ell} - \hat{\mu}_{a\ell}(z)|}{\hat{\sigma}_{a\ell}(z)} = o(1), \tag{D.4}$$

and because $\sigma_{a\ell} = \Theta(1)$ and $m_\sigma = \Theta(1)$ we also have

$$\max_{\substack{z \in \mathcal{Z}_\tau \\ a \in [k] \\ \ell \in [d]}} \frac{\left| \hat{\sigma}_{a\ell}^{(t)} - \sigma_{a\ell} \right|}{\hat{\sigma}_{a\ell}^{(t)}} \cdot \left(\sigma_{a\ell}^{-1} \Delta_{\mu,\infty} + m_\sigma \right) \;=\; o\left(\Delta_{\mu,\infty} \right). \tag{D.5}$$

We conclude by combining (D.2) with (D.3), (D.4), and (D.5). ■

Lemma 15 *Let $\epsilon > 0$, $\tau = o(n/k)$. Let $\sigma_{\min} = \min_{a \in [k], \ell \in [d]} \sigma_{a\ell}$ and $\omega_n = \omega(1)$. Suppose that $\min_{a \in [k]} |\Gamma_a(z^*)| \ge \alpha n / k$ for some $\alpha > 0$. For n large enough, it holds*

$$\max_{\substack{z \in [k]^n \\ 1 \le \mathrm{Ham}(z^*, z) \le \tau}} \frac{G(X, \mu, \sigma, z)}{\mathrm{Ham}(z^*, z)} \;=\; O\left(d\sqrt{k}\omega_n \left(1 + \frac{\Delta_{\mu,\infty}}{\sqrt{n/k}} \right) \right)$$

and

$$G(X, \mu, \sigma, z^*) \;=\; O\left(d\sqrt{k}\omega_n \right)$$

with probability at least $1 - 2e^{-\omega_n^2} - 3kd\left(4e^{-\alpha\sqrt{\frac{n}{k}}} + 6e^{-\frac{C}{2}\sqrt{\frac{\alpha n}{2k}}} \right)$.

Proof Let $\mathcal{Z}_\tau = \{ z \in [k]^n \colon 1 \le \mathrm{Ham}(z^*, z) \le \tau \}$. Because the random variables $X_{i\ell}$ are Laplace distributed with location $\mu_{z_i^*\ell}$ and scale $\sigma_{z_i^*\ell}$, the random variables $2\sigma_{z_i^*\ell}^{-1} \left| X_{i\ell} - \mu_{z_i^*\ell} \right|$ are $\chi^2(2)$ distributed. Hence, the random variables $Y_i = \left| X_{i\ell} - \mu_{z_i^*\ell} \right| - \sigma_{z_i^*\ell}$ are sub-exponential with zero mean. Bernstein's inequality for sub-exponential random variables (see (C.1)) ensures that the event

$$\mathcal{E}_1 \;=\; \left\{ \sum_{i \in [n]} \left(\left| X_{i\ell} - \mu_{z_i^*\ell} \right| - \sigma_{z_i^*\ell} \right) \;\le\; \omega_n \sqrt{n} \right\}$$

holds with a probability of at least $1 - 2e^{-\omega_n^2}$. Moreover, let $\mathcal{E}_\mu(z)$ and $\mathcal{E}_\sigma(z)$ be the events

$$\mathcal{E}_\mu(z) \;=\; \left\{ \max_{\substack{a \in k \\ \ell \in [d]}} \left| \hat{\mu}_{a\ell}(z) - \mu_{a\ell} \right| \;\le\; \sqrt{\frac{k}{n}} + \frac{2k\Delta_{\mu,\infty}}{\alpha n} \mathrm{Ham}(z^*, z) + C' \sqrt{\frac{k\mathrm{Ham}(z^*, z)}{n}} \right\},$$

$$\mathcal{E}_\sigma(z) \;=\; \left\{ \max_{\substack{a \in k \\ \ell \in [d]}} \left| \hat{\sigma}_{a\ell}(z) - \sigma_{a\ell} \right| \;\le\; 2\sqrt{\frac{k}{n}} + \frac{2k(\Delta_{\mu,\infty} + \Delta_{\sigma,\infty})}{\alpha n} \mathrm{Ham}(z^*, z) + 2C' \sqrt{\frac{k\mathrm{Ham}(z^*, z)}{n}} \right\}.$$

By Propositions 10 and 11, the event $\mathcal{E} = \mathcal{E}_1 \bigcap_{z \in \mathcal{Z}_\tau} \left(\mathcal{E}_\mu(z) \cap \mathcal{E}_\sigma(z) \right)$ holds with probability at least $1 - 2e^{-\omega_n^2} - 3kd\left(4e^{-\alpha\sqrt{\frac{n}{k}}} + 6e^{-\frac{C}{2}\sqrt{\frac{\alpha n}{2k}}} \right)$. On this event $\mathcal{E}$, we have for all $z \in \mathcal{Z}_\tau$,

$$\frac{G(X, \mu, \sigma, z)}{\mathrm{Ham}(z^*, z)} \;\le\; \frac{d\omega_n}{\min\limits_{\substack{a \in [k] \\ \ell \in [L]}} \sigma_{a\ell}^2} \left(\frac{2\sqrt{k}}{\mathrm{Ham}(z, z^*)} + \frac{2k(\Delta_{\mu,\infty} + \Delta_{\sigma,\infty})}{\alpha\sqrt{n}} + 2C' \sqrt{\frac{k}{\mathrm{Ham}(z^*, z)}} \right),$$

and the result holds by noticing that $\mathrm{Ham}(z^*, z) \ge 1$.

To obtain the upper bound on $G(X, \mu, \sigma, z^*)$, we use the events $\mathcal{E}_1$ and $\mathcal{E}_\sigma(z^*)$. ■

D.3. Example

In this section, we consider a Laplace mixture model where for all $\ell \in [d]$ we have $\sigma_{1\ell} = \cdots = \sigma_{k\ell}$, and we simply denote this quantity by σ_ℓ. Assume also that σ_ℓ is independent of n, and that $\mu_{a\ell} = m_{a\ell}\rho_n$ where $m_{a\ell}$ are non-zero constants independent of n, verifying $m_{a\ell} \neq m_{b\ell}$ whenever $a \neq b$, and $\rho_n = \omega(1)$. The following Lemma provides the expression of the Chernoff information of this model.

Lemma 16 *Let $\mathcal{F} = \{f_1, \cdots, f_k\}$ be the mixture of Laplace distributions described above. We have* $\mathrm{Chernoff}(\mathcal{F}) = (1 + o(1)) \sum_{\ell=1}^{d} \frac{|\mu_{a\ell} - \mu_{b\ell}|}{\sigma_\ell}$.

Proof We denote by $\mathrm{Ren}_t(f\|g)$ the Rényi divergence of order t between f and g, and we recall that $\mathrm{Chernoff}(f, g) = \sup_{t\in(0,1)}(1 - t)\mathrm{Ren}_t(f\|g)$. From direct computations of Renyi divergence between Laplace distributions (see for example Gil et al. (2013)), we observe that $t \mapsto (1 - t)\mathrm{Ren}_t(f_a\|f_b)$ is maximal at $t = 1/2$, and we further have

$$\mathrm{Chernoff}(f_a, f_b) \;=\; \frac{1}{2}\sum_{\ell=1}^{d}\left(\frac{|\mu_{a\ell} - \mu_{b\ell}|}{\sigma_\ell} - 2\log\left(1 + \frac{|\mu_{a\ell} - \mu_{b\ell}|}{2\sigma_\ell}\right)\right)$$

and we conclude because $|\mu_{a\ell} - \mu_{b\ell}| = \omega(1)$ and $\sigma_\ell = \Theta(1)$. ■

Hence, we have $\Delta_{\mu,\infty} = \Theta(\rho_n)$ and $\mathrm{Chernoff}(\mathcal{F}) = \Theta(d\rho_n)$. Therefore, the conditions $\Delta_{\mu,\infty} = O(d^{-1}\mathrm{Chernoff}(\mathcal{F}))$ and $\mathrm{Chernoff}(\mathcal{F}) = \omega(d\sqrt{k})$ of Theorem 5 become $\rho_n = \omega(\sqrt{k})$. Moreover, for the initialisation, the condition of Lemma 7 are verified if $\rho_n = \omega(k^2)$.

Appendix E. Proofs for Section 3.3 (Exponential family mixture models)

E.1. Proof of Theorem 6

To prove Theorem 6, we adopt the same approach as in the proof of Theorem 5 done in Section D. To simplify the notations, we suppose that the data points X_i are sampled from a natural exponential family, that is, the sufficient statistics u in the definition of the exponential family (3.8) is the identity. The proof for a general exponential family is obtained by substituting $u(X_i)$ for X_i throughout the proof.

We first notice that for any $a \in [k]$, we have

$$\begin{aligned}
&\left|\hat{\ell}_a^{(t)}(X_i) - \ell_a(X_i)\right| \\
&= \left|\sum_{\ell=1}^{d} \mathrm{Breg}_{\psi_\ell^*}\left(X_{i\ell}, \hat{\mu}_{a\ell}^{(t)}\right) - \mathrm{Breg}_{\psi_\ell^*}\left(X_{i\ell}, \mu_{a\ell}\right)\right| \\
&= \left|\sum_{\ell=1}^{d} \mathrm{Breg}_{\psi_\ell^*}\left(\mu_{a\ell}, \hat{\mu}_{a\ell}^{(t)}\right) + \langle X_{i\ell} - \mu_{a\ell}, \nabla\psi_\ell^*(\mu_{a\ell}) - \nabla\psi_\ell^*(\hat{\mu}_{a\ell}^{(t)})\rangle\right| \\
&\le \sum_{\ell=1}^{d}\left\{\left|\mathrm{Breg}_{\psi_\ell^*}\left(\mu_{a\ell}, \hat{\mu}_{a\ell}^{(t)}\right)\right| + \left(|X_{i\ell} - \mu_{z_i^*\ell}| + |\mu_{z_i^*\ell} - \mu_{a\ell}|\right)\cdot\left|\nabla\psi_\ell^*(\mu_{a\ell}) - \nabla\psi_\ell^*\left(\hat{\mu}_{a\ell}^{(t)}\right)\right|\right\},
\end{aligned}$$

where the second equality uses Lemma 17 and the last inequality uses the triangle and Cauchy-Schwarz's inequalities. Hence, combining this bound with the definition of the excess error (3.5), we obtain

$$\xi_{\text{excess}}(\delta) \;\leq\; 2\delta^{-1}\left(F\cdot\operatorname{Ham}\left(z^*, z^{(t)}\right)+G\right)$$

where $F = F\left(X,\mu,\sigma,z^{(t-1)}\right)$ and $G = G\left(X,\mu,\sigma,z^{(t-1)}\right)$ are defined by

$$\begin{aligned} F(X,\mu,\sigma,z) \;&=\; d\max_{\substack{a\in[k]\\ \ell\in[d]}}\left\{\left|\operatorname{Breg}_{\psi_\ell^*}\left(\mu_{a\ell},\hat\mu_{a\ell}(z)\right)\right| + \left(|\mu_{z_i^*\ell}-\mu_{a\ell}|+\sigma_{z_i^*\ell}\right)\cdot\left|\nabla\psi_\ell^*(\mu_{a\ell})-\nabla\psi_\ell^*\left(\hat\mu_{a\ell}(z)\right)\right|\right\}, \\ G(X,\mu,\sigma,z) \;&=\; \sum_{\ell=1}^{d}\sum_{i\in[n]}\sum_{b\in[k]}\left(|X_{i\ell}-\mu_{z_i^*\ell}|-\sigma_{z_i^*\ell}\right)\cdot\max_{a\in[k]}\left|\nabla\psi_\ell^*(\mu_{a\ell})-\nabla\psi_\ell^*\left(\hat\mu_{a\ell}(z)\right)\right|, \end{aligned}$$

where $\sigma_{z_i^*\ell} = \mathbb{E}\left[\left|X_{i\ell}-\mu_{z_i^*\ell}\right|\right]$. The bounding of the quantities F and G is done in a very similar manner as what is done in Lemmas 14 and 15. Indeed, let $\mathcal{Z}_\tau = \{z\in[k]^n\colon \operatorname{Ham}(z^*,z)\leq\tau\}$. Because by assumption the $X_{i\ell}$ are sub-exponential, we can apply Proposition 10 to show that the event $\mathcal{E} = \bigcap_{\operatorname{Ham}(z,z^*)\leq 2^{-1}\alpha nk^{-1}}\mathcal{E}_\mu(z)$, where

$$\mathcal{E}_\mu(z) \;=\; \left\{\max_{\substack{a\in k\\ \ell\in[d]}}|\hat\mu_{a\ell}(z)-\mu_{a\ell}| \;\leq\; \sqrt{\frac{k}{n}}+\frac{2k\Delta_{\mu,\infty}}{\alpha n}\operatorname{Ham}(z^*,z)+C'\sqrt{\frac{k\operatorname{Ham}(z^*,z)}{n}}\right\}$$

holds with a probability of at least $1-kd\left(4e^{-\alpha\sqrt{nk^{-1}}}+6e^{-C\sqrt{\alpha 2^{-1}nk^{-1}}}\right)$. Under this event, we notice that $|\mu_{a\ell}-\hat\mu_{a\ell}(z)| = o(1)$ and we bound F as follows. We first use Lemma 18 to show that

$$\left|\operatorname{Breg}_{\psi_\ell^*}(\mu_{a\ell},\hat\mu_{a\ell}(z))\right| \;\leq\; \left|\nabla^2\psi_\ell^*(\mu_{a\ell})\right|\cdot|\hat\mu_{a\ell}(z)-\mu_{a\ell}|^2.$$

Similarly,

$$\begin{aligned} \left|\nabla\psi_\ell^*(\mu_{a\ell})-\nabla\psi_\ell^*\left(\hat\mu_{a\ell}(z)\right)\right| \;&\leq\; \sup_{y\in[\mu_{a\ell},\hat\mu_{a\ell}(z)]}\left|\nabla^2\psi_\ell^*(y)\right|\cdot|\hat\mu_{a\ell}(z)-\mu_{a\ell}| \\ &\leq\; 2\left|\nabla^2\psi_\ell^*(\mu_{a\ell})\right|\cdot|\hat\mu_{a\ell}(z)-\mu_{a\ell}| \end{aligned}$$

where the last line holds by continuity of $\nabla^2\psi_\ell^*$ and because $|\hat\mu_{a\ell}(z)-\mu_{a\ell}| = o(1)$. By assumption, $\left|\nabla^2\psi_\ell^*(\mu_{a\ell})\right| = O(1)$ and hence $F(X,\mu,z) = o\left(d\Delta_{\mu,\infty}\right)$.

Finally, on the event $\mathcal{E}$, we have for all $z\in\mathcal{Z}_\tau$,

$$\begin{aligned} G(X,\mu,\sigma,z) \;&\leq\; 2d\omega_n\sqrt{n}\left(\sqrt{\frac{k}{n}}+\frac{2k\Delta_{\mu,\infty}}{\alpha n}\operatorname{Ham}(z^*,z)+C'\sqrt{\frac{k\operatorname{Ham}(z^*,z)}{n}}\right)\max_{\substack{a\in[k]\\ \ell\in[d]}}|\nabla^2(\mu_{a\ell})| \\ &=\; O\left(\omega_n d\sqrt{k}\left(1+\frac{\Delta_{\mu,\infty}}{\sqrt{nk^{-1}}}\right)\right)\operatorname{Ham}(z^*,z). \end{aligned}$$

To finish the proof, we proceed in the same way as at the end of the proof of Theorem 5.

E.2. Additional Lemmas

Lemma 17 (Generalized triangle inequality for Bregman divergences) *Let $x, y, z \in \mathbb{R}^d$. Then*

$$\mathrm{Breg}_\phi(x, y) - \mathrm{Breg}_\phi(x, z) \;=\; \mathrm{Breg}_\phi(z, y) + \langle x - z, \nabla\phi(z) - \nabla\phi(y)\rangle.$$

Proof By definition of Bregman divergence,

$$\begin{aligned} \mathrm{Breg}_\phi(x, y) \;&=\; \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle, \\ \mathrm{Breg}_\phi(x, z) \;&=\; \phi(x) - \phi(z) - \langle x - z, \nabla\phi(z)\rangle, \end{aligned}$$

and the result holds by direct computation. ■

Lemma 18 *Let $\phi\colon \Theta \to \mathbb{R}$ be a convex twice continuously differentiable function defined on an open space Θ, and let $(x_t) \in \Theta^{\mathbb{Z}_+}$ be a sequence such that $\lim_{t\to\infty} x_t = x$. Then, we have for n large enough*

$$\left|\mathrm{Breg}_\phi(x, x_t)\right| \;\le\; 2\left\|\nabla^2\phi(x)\right\| \cdot \|x - x_t\|^2\,.$$

Proof Because $\mathrm{Breg}_\phi(x, x_t)$ equals the difference between ϕ evaluated at x and its first order Taylor approximation around x_t evaluated at x, we have

$$\left|\mathrm{Breg}_\phi(x, x_t)\right| \;\le\; \left\|\nabla^2\phi(x_t)\right\| \cdot \|x - x_t\|^2.$$

We finish the proof by noticing that, for t large enough, we have $\left\|\nabla^2\phi(x_t)\right\| \le 2\left\|\nabla^2\phi(x)\right\|$ by continuity of $\nabla^2\phi$. ■