Vol. ()

A Quasi-Bayesian Perspective to Online Clustering

Le Li

Université d'Angers & iAdvize
e-mail: le@iadvize.com

Benjamin Guedj[*]

Inria
Modal project-team, Lille - Nord Europe research center
France
e-mail: benjamin.guedj@inria.fr
url: https://bguedj.github.io

and

Sébastien Loustau

Lumen AI
url: http://www.lumenai.fr

Abstract: When faced with high frequency streams of data, clustering raises theoretical and algorithmic pitfalls. We introduce a new and adaptive online clustering algorithm relying on a quasi-Bayesian approach, with a dynamic (*i.e.*, time-dependent) estimation of the (unknown and changing) number of clusters. We prove that our approach is supported by minimax regret bounds. We also provide an RJMCMC-flavored implementation (called PACBO, see https://cran.r-project.org/web/packages/PACBO/index.html) for which we give a convergence guarantee. Finally, numerical experiments illustrate the potential of our procedure.

MSC 2010 subject classifications: Primary 62L12; secondary 62C10, 62C20, 62L20.
Keywords and phrases: Online clustering, Quasi-Bayesian learning, Minimax regret bounds, Reversible Jump Markov Chain Monte Carlo.

Contents

[*]Corresponding author

1. Introduction

Online learning has been extensively studied these last decades in game theory and statistics (see Cesa-Bianchi and Lugosi, 2006, and references therein). The problem can be described as a sequential game: a blackbox reveals at each time t some $z_t \in \mathcal{Z}$. Then, the forecaster predicts the next value based on the past observations and possibly other available information. In the present work we will consider the scenario in which the sequence (z_t) is not assumed to be a realization of some stochastic process. One of the well known problem in online learning that happened to attract a lot of interest is prediction with expert advice. In this setting, the forecaster has access to a set $\{f_{e,t} \in \mathcal{D} : e \in \mathcal{E}\}$ of experts' predictions, where $f_{e,t}$ is the prediction of expert e at time t, $\mathcal{D}$ is a decision space which is assumed to be a convex subset of vector space and $\mathcal{E}$ is a finite set of experts (such as deterministic physical models, or stochastic decisions). Predictions made by the forecaster and experts are assessed with a loss function $\ell : \mathcal{D} \times \mathcal{Z} \longrightarrow \mathbb{R}_+$. The goal is to build a sequence $\hat{z}_1, \ldots, \hat{z}_T$ (denoted by $(\hat{z}_t)_{1:T}$ in the sequel) of predictions which are nearly as good as the best expert's predictions in the first T time rounds, *i.e.*, satisfying uniformly over any sequence (z_t) the following regret bound

$$\sum_{t=1}^{T} \ell(\hat{z}_t, z_t) - \min_{e \in \mathcal{E}} \left\{ \sum_{t=1}^{T} \ell\left(f_{e,t}, z_t\right) \right\} \leq \Delta_T(\mathcal{E}),$$

where $\Delta_T(\mathcal{E})$ is a remainder term. This term should be as small as possible and in particular sublinear in T. When $\mathcal{E}$ is finite, and the loss is bounded in $[0,1]$ and convex in its first argument, an optimal $\Delta_T(\mathcal{E}) = \sqrt{(T/2)\log|\mathcal{E}|}$ is given by Theorem 2.2 of Cesa-Bianchi and Lugosi (2006). The optimal forecaster is then obtained by forming the exponentially weighted average of all experts. For similar results, we refer the reader to Littlestone and Warmuth (1994) and Cesa-Bianchi et al. (1997).

Online learning techniques have also been applied to the regression framework. In particular, sequential ridge regression has been studied by Vovk (2001). For any $t = 1,\dots,T$, we now assume that $z_t = (x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$. At each time t, the forecaster gives a prediction $\hat{y}_t$ of y_t, using only newly revealed side information x_t and past observations $(x_s, y_s)_{1:(t-1)}$. Let $\langle \cdot, \cdot \rangle$ denote the scalar product in $\mathbb{R}^d$. A possible goal is to build a forecaster whose performance is nearly as good as the best linear forecaster $f_\theta \colon x \mapsto \langle \theta, x \rangle$, *i.e.*, such that uniformly over all sequences $(x_t, y_t)_{1:T}$,

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \ell(\langle \theta, x_t \rangle, y_t) \right\} \le \Delta_T(d), \tag{1}$$

where $\Delta_T(d)$ is a remainder term. This setting has been addressed by numerous contributions to the literature. In particular, Azoury and Warmuth (2001) and Vovk (2001) each provide an algorithm close to the ridge regression with a remainder term $\Delta_T(d) = \mathcal{O}(d \log T)$. Other authors have investigated the Gradient-Descent algorithm (Cesa-Bianchi et al., 1996; Kivinen and Warmuth, 1997) and the Exponentiated Gradient Forecasters (Cesa-Bianchi, 1999; Kivinen and Warmuth, 1997). Gerchinovitz (2011) extended the linear form $\langle u, x_t \rangle$ in (1) to $\langle u, \varphi(x_t) \rangle = \sum_{j=1}^{d} u_j \varphi_j(x_t)$, where $\varphi = (\varphi_1, \dots, \varphi_d)$ is a dictionary of base forecasters. In the so-called high dimensional setting ($d \gg T$), a sparsity regret bound with a remainder term $\Delta_T(d)$ growing logarithmically with d and T is proved by Gerchinovitz (2011, Proposition 3.1).

The purpose of the present work is to generalize the aforecited framework to the clustering problem, which has attracted attention from the machine learning and streaming communities. As an example, Guha et al. (2003), Barbakh and Fyfe (2008) and Liberty et al. (2016) study the so-called data streaming clustering problem. It amounts to clustering online data to a fixed number of groups in a single pass, or a small number of passes, while using little memory. From a machine learning perspective, Choromanska and Monteleoni (2012) aggregate online clustering algorithms, with a fixed number K of centers. The present paper investigates a more general setting since we aim to perform online clustering with a varying number K_t of centers. To the best of our knowledge, this is the first attempt of the sort in the literature. Let us stress that our approach only requires an upper bound p to K_t, which can be either a constant or an increasing function of the time horizon T.

Our approach strongly relies on a quasi-Bayesian methodology. The use of quasi-Bayesian estimators is especially advocated by the PAC-Bayesian theory which originates in the machine learning community in the late 1990s, in the seminal works of Shawe-Taylor and Williamson (1997) and McAllester (1999a,b) (see also Seeger, 2002, 2003). In the statistical learning community, the PAC-Bayesian approach has been extensively developed by Catoni (2004, 2007), Audibert (2004) and Alquier (2006), and later on adapted to the high dimensional setting Dalalyan and Tsybakov (2007, 2008), Alquier and Lounici (2011), Alquier and Biau (2013), Guedj and Alquier (2013), Guedj and Robbiano (2017) and Alquier and Guedj (2017). In a parallel effort, the online learning community has contributed to the PAC-Bayesian theory in the online regression setting (Kivinen and Warmuth, 1999). Audibert (2009) and Gerchinovitz (2011) have been the first attempts to merge both lines of research. Note that our approach is *quasi-Bayesian* rather than PAC-Bayesian, since we derive regret bounds (on quasi-Bayesian predictors) instead of PAC oracle inequalities.

Our main contribution is to generalize algorithms suited for supervised learning to the unsupervised setting. Our online clustering algorithm is adaptive in the sense that it does not require the knowledge of the time horizon T to be used and studied. The regret bounds that we obtain have a remainder term of magnitude $\sqrt{T \log T}$ and we prove that they are asymptotically minimax optimal.

The quasi-posterior which we derive is a complex distribution and direct sampling is not available. In Bayesian and quasi-Bayesian frameworks, the use of Markov Chain Monte Carlo (MCMC) algorithms is a popular way to compute estimates from posterior or quasi-posterior distributions. We refer to the comprehensive monograph Robert and Casella (2004) for an introduction to MCMC methods. For its ability to cope with transdimensional moves, we focus on the Reversible Jump MCMC algorithm from Green (1995), coupled with ideas from the Subspace Carlin and Chib algorithm proposed by Dellaportas et al. (2002) and Petralias and Dellaportas (2013). MCMC procedures for quasi-Bayesian predictors were firstly considered by Catoni (2004) and Dalalyan and Tsybakov (2012). Alquier and Biau (2013), Guedj and Alquier (2013) and Guedj and Robbiano (2017) are the first to have investigated the RJMCMC and Subspace Carlin and Chib techniques and we show in the present paper that this scheme is well suited to the clustering problem.

The paper is organised as follows. Section 2 introduces our notation and our online clustering procedure. Section 3 contains our mathematical claims, consisting in regret bounds for our online clustering algorithm. Remainder terms which are sublinear in T are obtained for a model selection-flavored prior. We also prove that these remainder terms are minimax optimal. We then discuss in Section 4 the practical implementation of our method, which relies on an adaptation of the RJMCMC algorithm to our setting. In particular, we prove its convergence towards the target quasi-posterior. The performance of the resulting algorithm, called PACBO, is evaluated on synthetic data. For the sake of clarity, proofs are postponed to Section 5. Finally, Appendix A contains an extension of our work to the case of a multivariate Student prior along with additional numerical experiments.

2. A quasi-Bayesian perspective to online clustering

Let $(x_t)_{1:T}$ be a sequence of data, where $x_t \in \mathbb{R}^d$. Our goal is to learn a time-dependent parameter K_t and a partition of the observed points into K_t cells, for any $t = 1, \dots, T$. To this aim, the output of our algorithm at time t is a vector $\hat{\mathbf{c}}_t = (\hat{c}_{t,1}, \hat{c}_{t,2}, \dots, \hat{c}_{t,K_t})$ of K_t centers in $\mathbb{R}^{dK_t}$, depending on the past information $(x_s)_{1:(t-1)}$ and $(\hat{\mathbf{c}}_s)_{1:(t-1)}$. A partition is then created by assigning any point in $\mathbb{R}^d$ to its closest center. When x_t is newly revealed, the instantaneous loss is computed as

$$\ell(\hat{\mathbf{c}}_t, x_t) = \min_{1 \le k \le K_t} |\hat{c}_{t,k} - x_t|_2^2, \tag{2}$$

where $|\cdot|_2$ is the ℓ_2-norm in $\mathbb{R}^d$. In what follows, we investigate regret bounds for cumulative losses. Given a measurable space Θ (embedded with its Borel σ-algebra), we let $\mathscr{P}(\Theta)$ denote the set of probability distributions on Θ, and for some reference measure ν, we let $\mathscr{P}_\nu(\Theta)$ be the set of probability distributions absolutely continuous with respect to ν. For any probability distributions $\rho, \pi \in \mathscr{P}(\Theta)$, the Kullback-Leibler divergence $\mathscr{K}(\rho, \pi)$ is defined as

$$\mathscr{K}(\rho,\pi) = \begin{cases} \int_\Theta \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}\rho & \text{when } \rho \in \mathscr{P}_\pi(\Theta), \\ +\infty & \text{otherwise.} \end{cases}$$

Note that for any bounded measurable function $h: \Theta \to \mathbb{R}$ and any probability distribution $\rho \in \mathscr{P}(\Theta)$ such that $\mathscr{K}(\rho, \pi) < +\infty$,

$$-\log \int_\Theta \exp(-h)\mathrm{d}\pi = \inf_{\rho \in \mathscr{P}(\Theta)} \left\{ \int_\Theta h \mathrm{d}\rho + \mathscr{K}(\rho, \pi) \right\}. \tag{3}$$

This result, which may be found in Csiszár (1975) and Catoni (2004, Equation 5.2.1), is critical to our scheme of proofs. Further, the infimum is achieved at the so-called Gibbs quasi-posterior $\hat{\rho}$, defined by

$$\mathrm{d}\hat{\rho} = \frac{\exp(-h)}{\int \exp(-h)\mathrm{d}\pi}\mathrm{d}\pi.$$

We now introduce the notation to our online clustering setting. Let $\mathscr{C} = \cup_{k=1}^p \mathbb{R}^{dk}$ for some integer $p \ge 1$. We denote by q a discrete probability distribution on the set $[\![1,p]\!] := \{1, \dots, p\}$. For any $k \in [\![1,p]\!]$, let π_k denote a probability distribution on $\mathbb{R}^{dk}$. For any vector of cluster centers $\mathbf{c} \in \mathscr{C}$, we define $\pi(\mathbf{c})$ as

$$\pi(\mathbf{c}) = \sum_{k \in [\![1,p]\!]} q(k) \mathbb{1}_{\{\mathbf{c} \in \mathbb{R}^{dk}\}} \pi_k(\mathbf{c}). \tag{4}$$

Note that (4) may be seen as a distribution over the set of Voronoi partitions of $\mathbb{R}^d$: any $\mathbf{c} \in \mathscr{C}$ corresponds to a Voronoi partition of $\mathbb{R}^d$ with at most p cells. In the sequel, we denote by $\mathbf{c} \in \mathscr{C}$ either a vector of centers or its associated Voronoi partition of $\mathbb{R}^d$ if no confusion arises, and we denote by $\pi \in \mathscr{P}(\mathscr{C})$ a prior over $\mathscr{C}$.

Let $\lambda > 0$ be some (inverse temperature) parameter. At each time t, we observe x_t and a random partition $\hat{\mathbf{c}}_{t+1} \in \mathscr{C}$ is sampled from the Gibbs quasi-posterior

$$\mathrm{d}\hat{\rho}_{t+1}(\mathbf{c}) \propto \exp\big(-\lambda S_t(\mathbf{c})\big)\mathrm{d}\pi(\mathbf{c}). \tag{5}$$

This quasi-posterior distribution will allow us to sample partitions with respect to the prior π defined in (4) and bent to fit past observations through the following cumulative loss

$$S_t(\mathbf{c}) = S_{t-1}(\mathbf{c}) + \ell(\mathbf{c}, x_t) + \frac{\lambda}{2}\big(\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)\big)^2,$$

where the latter one is a variance term. It is essential to make the *online variance inequality* hold true for general loss ℓ with quasi-posterior distribution, *i.e.,* no constraint such as the convexity or boundedness is imposed on ℓ (as discussed in Audibert, 2009, Section 4.2). $S_t(\mathbf{c})$ consists in the cumulative loss of $\mathbf{c}$ in the first t rounds and a term that controls the variance of the next prediction. Note that since $(x_t)_{1:T}$ is deterministic, no likelihood is attached to our approach, hence the terms "quasi-posterior" for $\hat{\rho}_{t+1}$ and "quasi-Bayesian" for our global method. The resulting estimate is a realization of $\hat{\rho}_{t+1}$ with a random number K_t of cells. This scheme is described in Algorithm 1. Note that this algorithm is an instantiation of Audibert's online SeqRand algorithm (Audibert, 2009, Section 4) to the special case of the loss defined in (2). However SeqRand does not account for adaptive rates $\lambda = \lambda_t$, as discussed in the next section.

Algorithm 1 The quasi-Bayesian online clustering algorithm

1: **Input parameters**: $p > 0, \pi \in \mathscr{P}(\mathscr{C})$, $\lambda > 0$ and $S_0 \equiv 0$
2: **Initialization**: Draw $\hat{\mathbf{c}}_1 \sim \pi = \hat{\rho}_1$
3: **For** $t \in [\![1, T]\!]$
4: Get the data x_t
5: Draw $\hat{\mathbf{c}}_{t+1} \sim \hat{\rho}_{t+1}(\mathbf{c})$ where $\mathrm{d}\hat{\rho}_{t+1}(\mathbf{c}) \propto \exp\big(-\lambda S_t(\mathbf{c})\big)\mathrm{d}\pi(\mathbf{c})$, and

$$S_t(\mathbf{c}) = S_{t-1}(\mathbf{c}) + \ell(\mathbf{c}, x_t) + \frac{\lambda}{2}\big(\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)\big)^2.$$

6: **End for**

3. Minimax regret bounds

Let $\mathbb{E}_{\mathbf{c}\sim\nu}$ stands for the expectation with respect to the distribution ν of $\mathbf{c}$ (abbreviated as $\mathbb{E}_\nu$ where no confusion is possible). We start with the following pivotal result.

Proposition 1. *For any sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$, for any prior distribution $\pi \in \mathscr{P}(\mathscr{C})$ and any $\lambda > 0$, the procedure described in Algorithm 1 satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{\rho \in \mathscr{P}_\pi(\mathscr{C})} \left\{ \mathbb{E}_{\mathbf{c}\sim\rho}\left[\sum_{t=1}^{T} \ell(\mathbf{c}, x_t)\right] + \frac{\mathscr{K}(\rho, \pi)}{\lambda} \right.$$
$$\left. + \frac{\lambda}{2}\mathbb{E}_{(\hat{\rho}_1, \ldots, \hat{\rho}_T)}\mathbb{E}_{\mathbf{c}\sim\rho}\sum_{t=1}^{T}[\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2 \right\}.$$

Proposition 1 is a straightforward consequence of Audibert (2009, Theorem 4.6) applied to the loss function defined in (2), the partitions $\mathscr{C}$, and any prior $\pi \in \mathscr{P}(\mathscr{C})$.

3.1. Preliminary regret bounds

In the following, we instantiate the regret bound introduced in Proposition 1. Distribution q in (4) is chosen as the following discrete distribution on the set $[\![1,p]\!]$

$$q(k) = \frac{\exp(-\eta k)}{\sum_{i=1}^{p} \exp(-\eta i)}, \quad \eta \geq 0. \tag{6}$$

When $\eta > 0$, the larger the number of cells k, the smaller the probability mass. Further, π_k in (4) is chosen as a product of k independent uniform distributions on ℓ_2-balls in $\mathbb{R}^d$:

$$\mathrm{d}\pi_k(\mathbf{c},R) = \left(\frac{\Gamma\left(\frac{d}{2}+1\right)}{\pi^{\frac{d}{2}}}\right)^k \frac{1}{(2R)^{dk}} \left\{\prod_{j=1}^{k} \mathbb{1}_{\{B_d(2R)\}}(c_j)\right\} \mathrm{d}\mathbf{c}, \tag{7}$$

where $R > 0$, Γ is the Gamma function and

$$B_d(r) = \left\{x \in \mathbb{R}^d,\, |x|_2 \leq r\right\} \tag{8}$$

is an ℓ_2-ball in $\mathbb{R}^d$, centered in $0 \in \mathbb{R}^d$ with radius $r > 0$. Finally, for any $k \in [\![1,p]\!]$ and any $R > 0$, let

$$\mathscr{C}(k,R) = \left\{\mathbf{c} = (c_j)_{j=1,\dots,k} \in \mathbb{R}^{dk}, \text{ such that } |c_j|_2 \leq R \quad \forall j\right\}.$$

Corollary 1. *For any sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$ and any $p \geq 1$, consider π defined by (4), (6) and (7) with $\eta \geq 0$ and $R \geq \max_{t=1,\dots,T} |x_t|_2$. If $\lambda \geq (d+2)/(2TR^2)$, the procedure described in Algorithm 1 satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\dots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{k \in [\![1,p]\!]} \left\{ \inf_{\mathbf{c} \in \mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \frac{dk}{2\lambda} \log\left(\frac{8R^2\lambda T}{d+2}\right) + \frac{\eta}{\lambda} k \right\} + \left(\frac{\log p}{\lambda} + \frac{d}{2\lambda} + \frac{81\lambda T R^4}{2}\right),$$

Note that $\inf_{\mathbf{c} \in \mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t)$ is a non-increasing function of the number k of cells while the penalty is linearly increasing with k. Small values for λ (or equivalently, large values for R) lead to small values for k. The additional term induced by the complexity of $\mathscr{C} = \bigcup_{k=1,\dots,p} \mathbb{R}^{dk}$ is $\log p$. A reasonable choice of λ would be such that $d/\lambda \log(\lambda T R^2/d+2)$ and $\lambda T R^4$ are of the same order in T. The calibration $\lambda = (d+2)\sqrt{\log T}/(2\sqrt{T}R^2)$ yields a sublinear remainder term in the following corollary.

Corollary 2. *Under the previous notation with* $\lambda = (d+2)\sqrt{\log T}/2\sqrt{T}R^2$*,* $R \geq \max_{t=1,\ldots,T} |x_t|_2$ *and* $T > 2$*, the procedure described in Algorithm 1 satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{k\in[\![1,p]\!]} \left\{ \inf_{\mathbf{c}\in\mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \frac{2(d+\eta)R^2}{d+2} k\sqrt{T\log T} \right\} + \left(\frac{2R^2 \log p}{d+2} + \frac{dR^2}{d+2} + \frac{81(d+2)R^2}{4} \right) \sqrt{T\log T}. \quad (9)$$

Remark 1. *If we assume* T *and* R *are constants, the reason that* λ *is chosen to be of order of magnitude of* d *here, rather than of* $\sqrt{d}$*, is to guarantee that it satisfies the condition* $\lambda \geq (d+2)/2TR^2$ *in Corollary 1. However, if* T *is sufficiently large,* e.g., $T \geq (d+2)^2/d$*, then the choice* $\lambda = \sqrt{d\log T}/2\sqrt{T}R^2$ *will satisfy the condition and will make the right hand side of the above inequality grow linearly in* $\sqrt{d}$ *while keeping the order of magnitude for* T *and* R*.*

Let us assume that the sequence $x_1, \ldots, x_T$ is generated from a distribution with $k^\star \in [\![1,p]\!]$ clusters. We then define the expected cumulative loss (ECL) and oracle cumulative loss (OCL) as

$$\mathrm{ECL} = \sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t),$$

$$\mathrm{OCL} = \inf_{\mathbf{c}\in\mathscr{C}(k^\star,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t).$$

Then Corollary 2 yields

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c}\in\mathscr{C}(k^\star,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) \leq Jk^\star\sqrt{T\log T}, \quad (10)$$

where J is a constant depending on d, R and $\log p$. In (10) the regret of our randomized procedure, defined as the difference between ECL and OCL is sublinear in T. However, whenever $k^\star > p$, we can deduce from Corollary 2 that

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c}\in\mathscr{C}(k^\star,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) \leq \inf_{k\in[\![1,p]\!]} \left\{ \inf_{\mathbf{c}\in\mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) - \inf_{\mathbf{c}\in\mathscr{C}(k^\star,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \frac{2(d+\eta)R^2}{d+2} k\sqrt{T\log T} \right\} + \left(\frac{R^2(2\log p + d)}{d+2} + \frac{81(d+2)R^2}{4} \right) \sqrt{T\log T},$$

where $\inf_{\mathbf{c}\in\mathscr{C}(k^\star,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t)$ is the oracle cumulative loss (*i.e.,* OCL) with $k^\star$ clusters.

If there exists a $k \in [\![1,p]\!]$ such that $\inf_{\mathbf{c}\in\mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t)$ is close to OCL, then our ECL is also close to OCL up to a term of order $k\sqrt{T\log T}$. However, if no such

k exists, then the term $\frac{2(d+\eta)R^2}{d+2}k\sqrt{T\log T}$ starts to dominate, hence the quality of bound is deteriorated.

Finally, note that the dependency in k inside the braces on the right-hand side of (9) may be improved by choosing $\lambda = (d+2)\sqrt{p\log T}/2\sqrt{T}R^2$ in Corollary 2. This allows to achieve the optimal rate $\sqrt{k}$ instead of k, since $k/\sqrt{p} \le \sqrt{k}$ for any $k \in [\![1,p]\!]$. However, this makes the last term in Corollary 2 of order of $\sqrt{pT\log T}$. Note that the effort to make the regret bound grow in $\sqrt{k}$, rather than $\sqrt{p}$ for $k \in [\![1,p]\!]$ may be achieved by using a similar strategy to the one of Wintenberger (2017), which introduces a recursive aggregation procedure with distinct learning rates for each expert in a finite set. Those learning rates are computed with a second order refinement of losses (or a linearized version when the loss is convex in its second argument) for each expert, at each time round. The regret of his strategy with respect to best aggregation of M finite experts is of the order of $\log M\sqrt{T}\log\log T$. However, the context for this procedure is not the same as ours, as we resort to the Gibbs quasi-posterior which is defined on $\mathscr{C}$, a continuous set. In addition, we focus on a single temperature parameter λ for the sake of computational complexity since the second order refinement requires the computation of the expectation of loss with respect to each expert in a finite set while, in our case, the "expert set" (*i.e.,* $\mathscr{C}$) is continuous, leading to the tedious computation of second order refinement.

3.2. Adaptive regret bounds

The time horizon T is usually unknown, prompting us to choose a time-dependent inverse temperature parameter $\lambda = \lambda_t$. We thus propose a generalization of Algorithm 1, described in Algorithm 2.

Algorithm 2 The adaptive quasi-Bayesian online clustering algorithm

1: **Input parameters**: $p > 0, \pi \in \mathscr{P}(\mathscr{C})$, $(\lambda_t)_{0:T} > 0$ and $S_0 \equiv 0$
2: **Initialization**: Draw $\hat{\mathbf{c}}_1 \sim \pi = \hat{\rho}_1$
3: **For** $t \in [\![1,T]\!]$
4: Get the data x_t
5: Draw $\hat{\mathbf{c}}_{t+1} \sim \hat{\rho}_{t+1}(\mathbf{c})$ where $\mathrm{d}\hat{\rho}_{t+1}(\mathbf{c}) \propto \exp\big(-\lambda_t S_t(\mathbf{c})\big)\mathrm{d}\pi(\mathbf{c})$, and

$$S_t(\mathbf{c}) = S_{t-1}(\mathbf{c}) + \ell(\mathbf{c},x_t) + \frac{\lambda_{t-1}}{2}\big(\ell(\mathbf{c},x_t) - \ell(\hat{\mathbf{c}}_t,x_t)\big)^2.$$

6: **End for**

This adaptive algorithm is supported by the following more involved regret bound.

Theorem 1. *For any sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$, any prior distribution π on $\mathscr{C}$, if $(\lambda_t)_{0:T}$ is a non-increasing sequence of positive numbers, then the procedure described in Algorithm 2 satisfies*

$$\sum_{t=1}^{T}\mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots,\hat{\rho}_t)}\ell(\hat{\mathbf{c}}_t,x_t) \le \inf_{\rho\in\mathscr{P}_\pi(\mathscr{C})}\left\{\mathbb{E}_{\mathbf{c}\sim\rho}\left[\sum_{t=1}^{T}\ell(\mathbf{c},x_t)\right] + \frac{\mathscr{K}(\rho,\pi)}{\lambda_T}\right.$$

$$+\mathbb{E}_{(\hat{\rho}_1,\ldots,\hat{\rho}_T)}\mathbb{E}_{\mathbf{c}\sim\rho}\left[\sum_{t=1}^{T}\frac{\lambda_{t-1}}{2}[\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t)]^2\right]\Bigg\}.$$

If λ is chosen in Proposition 1 as $\lambda=\lambda_T$, the only difference between Proposition 1 and Theorem 1 lies on the last term of the regret bound. This term will be larger in the adaptive setting than in the simpler non-adaptive setting since $(\lambda_t)_{0:T}$ is non-increasing. In other words, here is the price to pay for the adaptivity of our algorithm. However, a suitable choice of λ_t allows, again, for a refined result.

Corollary 3. *For any deterministic sequence $(x_t)_{1:T}\in\mathbb{R}^{dT}$, if q and π_k in (4) are taken respectively as in (6) and (7) with $\eta\geq 0$ and $R\geq\max_{t=1,\ldots,T}|x_t|_2$, if $\lambda_t=(d+2)\sqrt{\log t}/\left(2\sqrt{t}R^2\right)$ for any $t\in[\![1,T]\!]$ and $\lambda_0=1$, then for $T\geq 5$ the procedure described in Algorithm 2 satisfies*

$$\sum_{t=1}^{T}\mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots,\hat{\rho}_t)}\ell(\hat{\mathbf{c}}_t,x_t)\leq\inf_{k\in[\![1,p]\!]}\left\{\inf_{\mathbf{c}\in\mathscr{C}(k,R)}\sum_{t=1}^{T}\ell(\mathbf{c},x_t)+\frac{2(d+\eta)R^2}{d+2}k\sqrt{T\log T}\right\}$$
$$+\left(\frac{2R^2\log p}{d+2}+\frac{dR^2}{d+2}+\frac{81(d+2)R^2}{2}\right)\sqrt{T\log T}.$$

Therefore, the price to pay for not knowing the time horizon T (which is a much more realistic assumption for online learning) is a multiplicative factor 2 in front of the term $\frac{81(d+2)R^2}{4}\sqrt{T\log T}$. This does not degrade the rate of convergence $\sqrt{T\log T}$.

In the next corollary, we use the doubling trick (Cesa-Bianchi and Lugosi, 2006, Section 2.3, also appearing in Cesa-Bianchi et al., 2007) to show how can we overcome the difficulty when a priori bound R on the ℓ_2-norm of sequence $(x_t)_{1:T}$ is unknown.

Let us first denote by $R_0=1$, and for $t\geq 1$

$$R_t=\max_{s=1,\ldots,t}2^{\lceil\log_2(|x_s|_2)\rceil},$$

where $\lceil x\rceil$ represents the least integer greater than or equal to $x\in\mathbb{R}$. It is easy to see that $(R_t)_{t\geq1}$ is non-decreasing and satisfies for any $t\geq 1$

$$\max_{s=1,\ldots,t}|x_s|_2\leq R_t\leq 2\max_{s=1,\ldots,t}|x_s|_2.$$

We call epoch r, $r=0,1,\ldots$, the sequence $(t_{r-1}+1,t_{r-1}+2,\ldots,t_r)$ of time steps where the last step t_r is the time step $t=t_r$ when $R_t>R_{t_{r-1}}$ take places for the first time (we set conventionally $t_{-1}=0$). Within each epoch $r\geq 0$, *i.e.*, for $t\in[t_{r-1}+1,\ldots,t_r]$, let

$$\lambda_{r,t}=\frac{(d+2)\sqrt{\log t}}{2\sqrt{t}R_{t_{r-1}}^2}.$$

Let **Alg-R** be a prediction algorithm that runs Algorithm 2 in each epoch r with parameter $\lambda_{r,t}$, then we have the following result.

Corollary 4. *For any deterministic sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$, if q and π_k in (4) are taken respectively as in (6) and (7) with $\eta \geq 0$, the regret of algorithm* **Alg-R** *satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{k \in [\![1,p]\!]} \left\{ \inf_{\mathbf{c} \in \mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \frac{56(d+\eta)R^2}{3(d+2)} k \sqrt{T \log T} \right\} + \frac{28}{3} \left(\frac{2R^2 \log p}{d+2} + \frac{dR^2}{d+2} + \frac{81(d+2)R^2}{2} \right) \sqrt{T \log T} + \frac{112}{3} R^2,$$

where $R = \max_{t=1,\ldots,T} |x_t|_2$.

Note that the price to pay for making our algorithm adaptive to unknown bound R is a multiplicative term $\frac{28}{3}$ and an additional $\frac{112}{3}R^2$ in the regret bound.

3.3. Minimax regret

This section is devoted to the study of the minimax optimality of our approach. The regret bound in Corollary 3 has a rate $\sqrt{T \log T}$, which is not a surprising result. Indeed, many online learning problems give rise to similar bounds depending also on the properties of the loss function. However, in the online clustering setting, it is legitimate to wonder wether the upper bound is tight, and more generally if there exists other algorithms which provide smaller regrets. The sequel answers both questions in a minimax sense.

Let us first denote by $|\mathbf{c}|$ the number of cells for a partition $\mathbf{c} \in \mathscr{C}$. We also introduce the following assumption.

Assumption $\mathscr{H}(s)$: *Let $R > 0$ and $T \in \mathbb{N}^*$. For a given $s \in [\![1,p]\!]$, we assume that the number of cells $\left|\mathbf{c}^\star_{T,R}\right|$ for partition $\mathbf{c}^\star_{T,R}$ defined by the following*

$$\mathbf{c}^\star_{T,R} = \operatorname*{arg\,min}_{\mathbf{c} \in \cup_{k=1}^{p} \mathscr{C}(k,R)} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + |\mathbf{c}| \sqrt{T \log T} \right\}.$$

equals to s, i.e.,$\left|\mathbf{c}^\star_{T,R}\right| = s$.

Note that several partitions may achieve the minimum. In that case, we adopt the convention that $\mathbf{c}^\star_{T,R}$ is any such partition with the smallest number of cells. Assumption $\mathscr{H}(s)$ means that $(x_t)_{1:T}$ could be well summarized by s cells since the infimum is reached for the partition $\mathbf{c}^\star_{T,R}$. We introduce the set

$$\omega_{s,R} = \left\{ (x_t) \text{ such that } \mathscr{H}(s) \text{ holds} \right\} \subseteq \mathbb{R}^{dT}.$$

For Algorithm 2, we have from Corollary 3 that

$$\sup_{(x_t) \in \omega_{s,R}} \left\{ \sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c} \in \mathscr{C}(s,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) \right\} \leq c_1 \times s \sqrt{T \log T},$$

where c_1 is a constant depending on R, d, p (recall that they are respectively the bound on the ℓ_2-norm of sequence $(x_t)_{1:T}$, the dimension of the data point and the maximum number of cells allowed for clustering).

Then for any $s \in \mathbb{N}^*$, $R > 0$, our goal is to obtain a lower bound of the form

$$\inf_{(\hat{\rho}_t)} \sup_{(x_t)\in\omega_{s,R}} \left\{ \sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\dots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c}\in\mathscr{C}(s,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) \right\} \geq c_2 \times s\sqrt{T\log T},$$

where c_2 is some constant satisfying $c_2 \leq c_1$.

The first infimum is taken over all distributions $(\hat{\rho}_t)_{1:T}$ whose support is $\cup_{k=1}^{p} \prod_{j=1}^{k} B_d(2R)$, where $B_d(2R)$ is defined in (8). Next, we obtain

$$\begin{aligned}
&\inf_{(\hat{\rho}_t)} \sup_{(x_t)\in\omega_{s,R}} \left\{ \sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\dots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c}\in\mathscr{C}(s,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) \right\} \\
&\qquad \geq \inf_{(\hat{\rho}_t)} \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\dots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, X_t) - \inf_{\mathbf{c}\in\mathscr{C}(s,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, X_t) \right\} \mathbb{1}_{\{(X_t)\in\omega_{s,R}\}}, \qquad (11)
\end{aligned}$$

where X_t, $t = 1,\dots,T$ are i.i.d with distribution μ defined on $\mathbb{R}^d$ and μ^T stands for the joint distribution of $(X_1,\dots,X_T)$. Unfortunately, in (11), since the infimum is taken over any distribution $(\hat{\rho}_t)$, there is no restriction on the number of cells of each partition $\hat{\mathbf{c}}_t$. Then, the left hand side of (11) could be arbitrarily small or even negative and the lower bound does not match the upper bound of Corollary 3. To handle this, we need to introduce a penalized term which accounts for the number of cells of each partition to the loss function ℓ. The upcoming theorem provides minimax results for an augmented value $\mathscr{V}_T(s)$ defined as

$$\mathscr{V}_T(s) = \inf_{(\hat{\rho}_t)} \sup_{(x_t)\in\omega_{s,R}} \left\{ \sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\dots,\hat{\rho}_t)} \left(\ell(\hat{\mathbf{c}}_t, x_t) + \frac{\sqrt{\log T}}{\sqrt{T}} |\hat{\mathbf{c}}_t| \right) - \inf_{\mathbf{c}\in\mathscr{C}(s,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) \right\}. \qquad (12)$$

In (12), we add a term which penalizes the number of cells of each partition. To capture the asymptotic behavior of $\mathscr{V}_T(s)$, we derive an upper bound for the penalized loss in (12). This is done in the following theorem which combines an upper and lower bound for the regret, hence proving that it is minimax optimal.

Theorem 2. *Let $s \in \mathbb{N}^*$, $R > 0$ such that*

$$2 \leq s \leq \left\lfloor \left(\frac{RT^{\frac{1}{4}}}{6\log T^{\frac{1}{4}}} \right)^{\frac{d}{d+1}} \right\rfloor, \qquad (13)$$

where $\lfloor x \rfloor$ represents the largest integer that is smaller than x. If T satisfies $T^{\frac{d+2}{2}} \geq 8R^{2d}\sqrt{\log T}$, then

$$s\sqrt{T\log T}\left(1 - \frac{2}{T}\left[1 + \frac{s-1}{2s^2}\right]\right) \leq \mathscr{V}_T(s) \leq \text{const.} \times s\sqrt{T\log T}. \qquad (14)$$

The lower bound on $\mathcal{V}_T(s)/T$ is asymptotically of order $\sqrt{\log T}/\sqrt{T}$. Note that Bartlett et al. (1998) obtained the less satisfying rate $1/\sqrt{T}$, however holding with no restriction on the number of cells retained in the partition whereas our claim has to comply with (13). This is the price to pay for our additional $\sqrt{\log T}$ factor. Note however that this price is mild as s is no longer upper bounded whenever T or R grow to $+\infty$, casting our procedure onto the online setting where the time horizon is not assumed finite and the number of clusters is evolving along time.

As a conclusion to the theoretical part of the manuscript, let us summarize our results. Regret bounds for Algorithm 1 are produced for our specific choice of prior π (Corollary 1) and with an involved choice of λ (Corollary 2). For the adaptive version Algorithm 2, the pivotal result is Theorem 1, which is instantiated for our prior in Corollary 3. Finally, the lower bound is stated in Theorem 2, proving that our regret bounds are minimax whenever the number of cells retained in the partition satisfies (13). We now move to the implementation of our approach.

4. The PACBO algorithm

Since direct sampling from the Gibbs quasi-posterior is usually not possible, we focus on a stochastic approximation in this section, called PACBO (available in the companion eponym R package from Li, 2016). Both implementation and convergence (towards the Gibbs quasi-posterior) of this scheme are discussed. This section also includes a short numerical experiment on synthetic data to illustrate the potential of PACBO compared to other popular clustering methods.

4.1. Structure and links with RJMCMC

In Algorithm 1 and Algorithm 2, it is required to sample at each t from the Gibbs quasi-posterior $\hat{\rho}_t$. Since $\hat{\rho}_t$ is defined on the massive and complex-structured space $\mathscr{C}$ (let us recall that $\mathscr{C}$ is a union of heterogeneous spaces), direct sampling from $\hat{\rho}_t$ is not an option and is much rather an algorithmic challenge. Our approach consists in approximating $\hat{\rho}_t$ through MCMC under the constraint of favouring local moves of the Markov chain. To do it, we will use resort to Reversible Jump MCMC (Green, 1995), adapted with ideas from the Subspace Carlin and Chib algorithm proposed by Dellaportas et al. (2002) and Petralias and Dellaportas (2013). Since sampling from $\hat{\rho}_t$ is similar for any $t = 1, \dots, T$, the time index t is now omitted for the sake of brevity.

Let $(k^{(n)}, \mathbf{c}^{(n)})_{0\le n\le N}$, $N \ge 1$ be the states of the Markov Chain of interest of length N, where $k^{(n)} \in [\![1, p]\!]$ and $\mathbf{c}^{(n)} \in \mathbb{R}^{dk^{(n)}}$. At each RJMCMC iteration, only local moves are possible from the current state $(k^{(n)}, \mathbf{c}^{(n)})$ to a proposal state $(k', \mathbf{c}')$, in the sense that the proposal state should only differ from the current state by at most one covariate. Hence, $\mathbf{c}^{(n)} \in \mathbb{R}^{dk^{(n)}}$ and $\mathbf{c}' \in \mathbb{R}^{dk'}$ may be in different spaces ($k' \neq k^{(n)}$). Two auxiliary vectors $v_1 \in \mathbb{R}^{d_1}$ and $v_2 \in \mathbb{R}^{d_2}$ with $d_1, d_2 \ge 1$ are needed to compensate for this dimensional difference, *i.e.*, satisfying the dimension matching condition introduced by Green (1995)

$$dk^{(n)} + d_1 = dk' + d_2,$$

such that the pairs $(v_1, \mathbf{c}^{(n)})$ and $(v_2, \mathbf{c}')$ are of analogous dimension. This condition is a preliminary to the detailed balance condition that ensures that the Gibbs quasi-posterior $\hat{\rho}_t$ is the invariant distribution of the Markov chain. The structure of PACBO is presented in Figure 1.

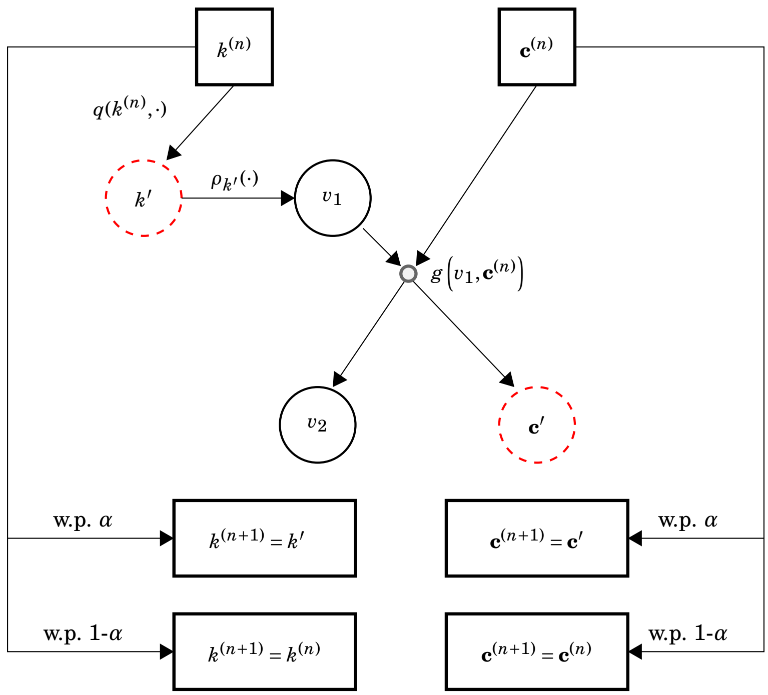


Figure 1: General structure of PACBO.

Let $\rho_{k'}(\cdot, \mathfrak{c}_{k'}, \tau_{k'})$ denote the multivariate Student distribution on $\mathbb{R}^{dk'}$

$$\rho_{k'}(\mathbf{c}, \mathfrak{c}_{k'}, \tau_{k'}) = \prod_{j=1}^{k'} \left\{ C^{-1}_{\tau_{k'}} \left(1 + \frac{|c_j - \mathfrak{c}_{k',j}|_2^2}{6\tau_{k'}^2} \right)^{-\frac{3+d}{2}} \right\} \mathrm{d}\mathbf{c}, \tag{15}$$

where $C^{-1}_{\tau_{k'}}$ denotes a normalizing constant. Let us now detail the proposal mechanism. First, a local move from $k^{(n)}$ to k' is proposed by choosing $k' \in [\![k^{(n)} - 1, k^{(n)} + 1]\!]$ with probability $q(k^{(n)}, \cdot)$. Next, choosing $d_1 = dk'$, $d_2 = dk^{(n)}$, we sample v_1 from $\rho_{k'}$ in (15). Finally, the pair $(v_2, \mathbf{c}')$ is obtained by

$$(v_2, \mathbf{c}') = g\left(v_1, \mathbf{c}^{(n)}\right),$$

where $g : (x, y) \in \mathbb{R}^{dk'} \times \mathbb{R}^{dk^{(n)}} \mapsto (y, x) \in \mathbb{R}^{dk^{(n)}} \times \mathbb{R}^{dk'}$ is a one-to-one, first order derivative mapping. The resulting RJMCMC acceptance probability is

$$\begin{aligned}\alpha\left[\left(k^{(n)}, \mathbf{c}^{(n)}\right), \left(k', \mathbf{c}'\right)\right] &= \min\left\{1, \frac{\hat{\rho}_t(\mathbf{c}')q(k', k^{(n)})\rho_{k^{(n)}}(v_2)}{\hat{\rho}_t(\mathbf{c}^{(n)})q(k^{(n)}, k')\rho_{k'}(v_1)}\left|\frac{\partial g\left(v_1, \mathbf{c}^{(n)}\right)}{\partial v_1 \partial \mathbf{c}^{(n)}}\right|\right\}, \\ &= \min\left\{1, \frac{\hat{\rho}_t(\mathbf{c}')q(k', k^{(n)})\rho_{k^{(n)}}(\mathbf{c}^{(n)}, \mathfrak{c}_{k^{(n)}}, \tau_{k^{(n)}})}{\hat{\rho}_t(\mathbf{c}^{(n)})q(k^{(n)}, k')\rho_{k'}(\mathbf{c}', \mathfrak{c}_{k'}, \tau_{k'})}\right\},\end{aligned}$$

since the determinant of the Jacobian matrix of g is 1. The resulting PACBO algorithm is described in Algorithm 3.

Algorithm 3 PACBO

1: **Initialization**: $(\lambda_t)_{1:T}$
2: **For** $t \in [\![1, T]\!]$
3: **Initialization**: $\left(k^{(0)}, \mathbf{c}^{(0)}\right) \in [\![1, p]\!] \times \mathbb{R}^{dk^{(0)}}$. Typically $k^{(0)}$ is set to $k^{(N)}$ from iteration $t-1$ ($k^{(0)} = 1$ at iteration $t = 1$).
4: **For** $n \in [\![1, N-1]\!]$
5: Sample $k' \in \left[\!\left[\max(1, k^{(n)} - 1), \min(p, k^{(n)} + 1)\right]\!\right]$ from $q(k^{(n)}, \cdot) = \frac{1}{3}$.
6: Let $\mathfrak{c}' \leftarrow$ standard k'-means output trained on $(x_s)_{1:(t-1)}$.
7: Let $\tau' = 1/\sqrt{pt}$.
8: Sample $v_1 \sim \rho_{k'}(\cdot, \mathfrak{c}_{k'}, \tau_{k'})$.
9: Let $(v_2, \mathbf{c}') = g(v_1, \mathbf{c}^{(n)})$.
10: Accept the move $(k^{(n)}, \mathbf{c}^{(n)}) = (k', \mathbf{c}')$ with probability

$$\begin{aligned}\alpha\left[(k^{(n)}, \mathbf{c}^{(n)}), (k', \mathbf{c}')\right] &= \min\left\{1, \frac{\hat{\rho}_t(\mathbf{c}')q(k', k^{(n)})\rho_{k^{(n)}}(v_2, \mathfrak{c}_{k^{(n)}}, \tau_{k^{(n)}})}{\hat{\rho}_t(\mathbf{c}^{(n)})q(k^{(n)}, k')\rho_{k'}(v_1, \mathfrak{c}_{k'}, \tau_{k'})}\left|\frac{\partial g(v_1, \mathbf{c}^{(n)})}{\partial v_1 \partial \mathbf{c}^{(n)}}\right|\right\} \\ &= \min\left\{1, \frac{\hat{\rho}_t(\mathbf{c}')q(k', k^{(n)})\rho_{k^{(n)}}(\mathbf{c}^{(n)}, \mathfrak{c}_{k^{(n)}}, \tau_{k^{(n)}})}{\hat{\rho}_t(\mathbf{c}^{(n)})q(k^{(n)}, k')\rho_{k'}(\mathbf{c}', \mathfrak{c}_{k'}, \tau_{k'})}\right\}\end{aligned}$$

11: Else $(k^{(n+1)}, \mathbf{c}^{(n+1)}) = (k^{(n)}, \mathbf{c}^{(n)})$.
12: **End for**
13: Let $\hat{\mathbf{c}}_t = \mathbf{c}^{(N)}$.
14: **End for**

4.2. Convergence of PACBO *towards the Gibbs quasi-posterior*

We prove that Algorithm 3 builds a Markov chain whose invariant distribution is precisely the Gibbs quasi-posterior as N goes to $+\infty$. To do so, we need to prove that the chain is $\hat{\rho}_t$-irreducible, aperiodic and Harris recurrent, see Robert and Casella (2004, Theorem 6.51) and Roberts and Rosenthal (2006, Theorem 20).

Recall that at each RJMCMC iteration in Algorithm 3, the chain is said to propose a "between model move" if $k' \neq k^{(n)}$ and a "within model move" if $k' = k^{(n)}$ and $\mathbf{c}' \neq \mathbf{c}^{(n)}$. The following result gives a sufficient condition for the chain to be Harris recurrent.

Lemma 1. *Let D be the event that no "within-model move" is ever accepted and $\mathscr{E}$ be the support of $\hat{\rho}_t$. Then the chain generated by Algorithm 3 satisfies*

$$\mathbb{P}\left[D|\left(k^{(0)},\mathbf{c}^{(0)}\right)=(k,\mathbf{c})\right]=0,$$

for any $k \in [\![1,p]\!]$ and $\mathbf{c} \in \mathbb{R}^{dk} \cap \mathscr{E}$.

Lemma 1 states that the chain must eventually accept a "within-model move". It remains true for other choices of $q(k^{(n)},\cdot)$ in Algorithm 3, provided that the stationarity of $\hat{\rho}_t$ is preserved.

Theorem 3. *Let $\mathscr{E}$ denote the support of $\hat{\rho}_t$. Then for any $\mathbf{c}^{(0)} \in \mathscr{E}$, the chain $\left(\mathbf{c}^{(n)}\right)_{1:N}$ generated by Algorithm 3 is $\hat{\rho}_t$-irreducible, aperiodic and Harris recurrent.*

Theorem 3 legitimates our approximation PACBO to perform online clustering, since it asymptotically mimics the behavior of the computationally unavailable $\hat{\rho}_t$. To the best of our knowledge, this kind of guarantee is original in the PAC-Bayesian literature.

Finally, let us stress that obtaining an explicit rate of convergence is beyond the scope of the present work. However, in most cases the chain converges rather quickly in practice, as illustrated by Figure 2. At time t, we advocate for setting $k^{(0)}$ as $k^{(N)}$ from round $t-1$, as a warm start.

4.3. Numerical study

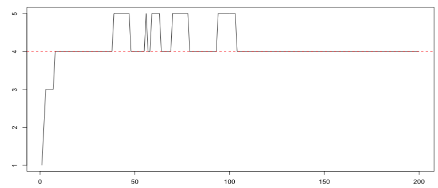
This section is devoted to the illustration of the potential of our quasi-Bayesian approach on synthetic data. Let us stress that all experiments are reproducible, thanks to the PACBO R package (Li, 2016). We do not claim to be exhaustive here but rather show the (good) behavior of our implementation on a toy example.

4.3.1. Calibration of parameters and mixing properties

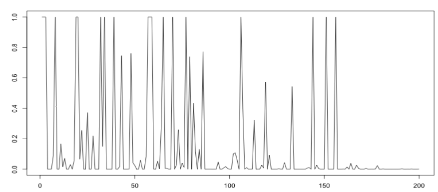
We set R to be the maximum ℓ_2-norm of the observations. Note that a too small value will yield acceptance ratios to be close to zero and will degrade the mixing of the chain. As advised by the theory, we advise to set $\lambda_t = 0.6 \times (d+2)\sqrt{\log t}/(2\sqrt{t})$. Recall that large values will enforce the quasi-posterior to account more for past data, whereas small values make the quasi-posterior alike the prior. We illustrate in Figure 2 the mixing behavior of PACBO. The convergence occurs quickly, and the default length of the RJMCMC runs is set to 500 in the PACBO package: this was a ceiling value in all our simulations.

4.3.2. Batch clustering setting

A large variety of methods have been proposed in the literature for selecting the number k of clusters in batch clustering (see Gordon, 1999; Milligan and Cooper,



(a) Number of clusters.



(b) Acceptance probability.

Figure 2: Typical RJMCMC output in PACBO. (a) $k^{(n)}_{1:N}$, number of clusters along the 200 iterations. The true number of clusters (set to 4 in this example) is indicated by a dashed red line (b) acceptance probability α along the 200 iterations, exhibiting its mixing behavior.

1985, for a survey). These methods may be of local or global nature. For local methods, at each step, each cluster is either merged with another one, split in two or remains. Global methods evaluate the empirical distortion of any clustering as a function of the number k of cells over the whole dataset, and select the minimizer of this distortion. The rule of Hartigan (1975) is a well-known representative of local methods. Popular global methods include the works of Calinski and Harabasz (1974), Krzanowski and Lai (1988) and Kaufman and Rousseeuw (1990), where functions based on the empirical distortion or on the average of within-cluster dispersion of each point are constructed and the optimal number of clusters is the maximizer of these functions. In addition, the Gap Statistic (Tibshirani et al., 2001) compares the change in within-cluster dispersion with the one expected under an appropriate reference null distribution. More recently, CAPUSHE (CAlibrating Penalty Using Slope Heuristics) introduced by Fischer (2011) and Baudry et al. (2012) addresses the problem from the penalized model selection perspective, in the form of two methods: DDSE (Data-Driven Slope Estimation) and Djump (Dimension jump). R packages implementing those methods are used with their default parameters in our simulations.

In this section, we compare PACBO to the aforecited methods in a batch setting with $n = 200$ observations simulated from the following 4 models.

Model 1 (1 group in dimension 5)**.** *Observations are sampled from a uniform distribution on the unit hypercube in* $\mathbb{R}^5$.

Model 2 (4 Gaussian groups in dimension 2)**.** *Observations are sampled from 4 bivariate Gaussian distributions with identity covariance matrix, whose mean vectors are respectively* $(0,0),(-2,-1),(0,4),(3,1)$. *Each observation is uniformly drawn from one of the four groups.*

Model 3 (7 Gaussian groups in dimension 50)**.** *Observations are sampled from 7 multivariate Gaussian distributions in* $\mathbb{R}^{50}$ *with identity covariance matrix, whose mean vectors are chosen randomly according to an uniform distribution on* $[-10,10]^{50}$. *Each observation is uniformly drawn from one of the seven groups.*

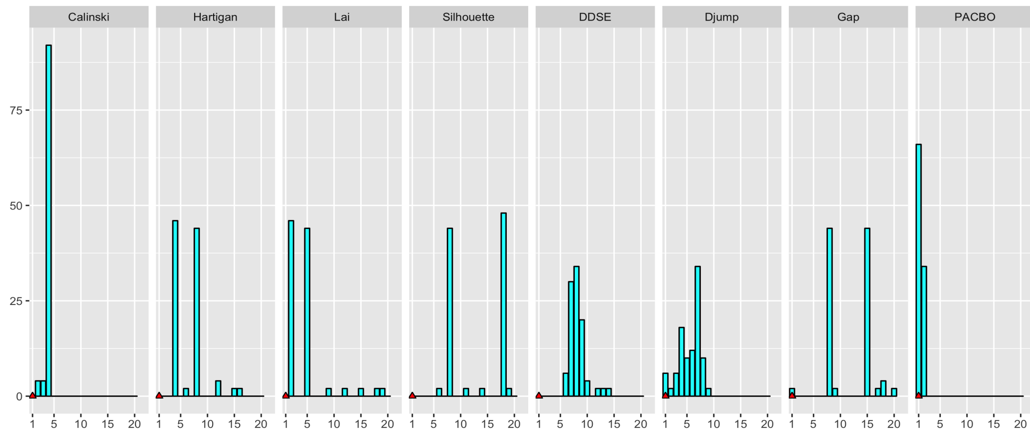
Model 4 (3 lognormal groups in dimension 3)**.** *Observations are sampled from 3 multivariate lognormal distributions in* $\mathbb{R}^3$ *with identity covariance matrix, whose mean vectors are respectively* $(1,1,1),(6,5,7),(10,9,11)$. *Each observation is uniformly drawn from one of the three groups.*

Figure 3 and Figure 4 present the percentage of the estimated number of cells k on 50 realizations of the 4 aforementioned models, for 8 methods including PACBO. In each graph, the red dot indicates the real number of groups. The methods used for selecting k are presented on the top of each panel, where DDSE (Data-Driven Slope Estimation) and Djump (Dimension jump) are the two methods introduced in CAPUSHE (Baudry et al., 2012). The maximum number of cells is set to 20.

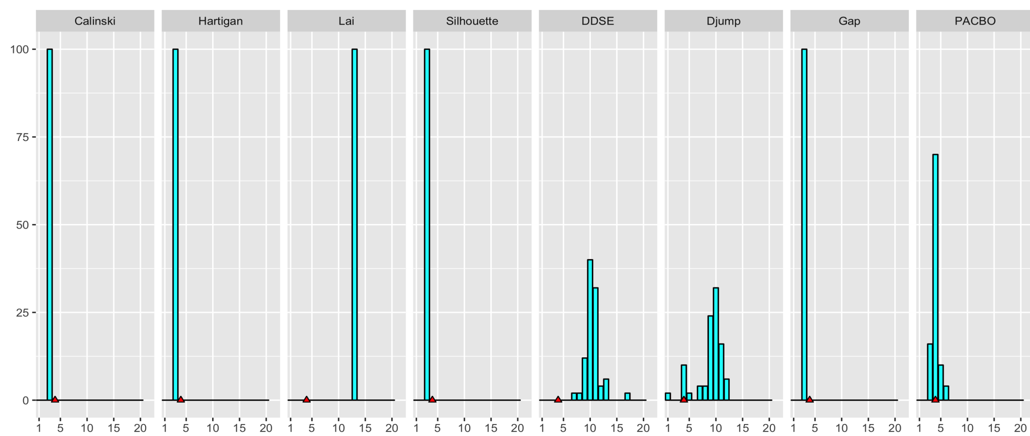
For Model 1 PACBO outperforms all competitors, since it selects the correct number of cells in almost 70% of our simulations, when all other methods barely find it (Figure 3a).
For Model 2 Calinski, Hartigan, Silhouette and Gap underestimate the number of cells by identifying 3 groups. Djump finds the true value $k=4$ less than 10%. PACBO identifies 4 groups in 60% of our runs (Figure 3b).
For Model 3 PACBO is one of the two best methods, together with Gap (Figure 4a).
For Model 4 where 3 groups of observations are generated from a heavy-tailed distribution, we consider a variant of PACBO with the ℓ_1-norm in $\mathbb{R}^d$, *i.e.*, we replace the loss in (2) by $\ell(\hat{\mathbf{c}}_t, x_t) = \min_{1\le k\le K_t} |\hat{c}_{t,k} - x_t|_1$. Figure 4b shows that most methods perform poorly, to the notable exception of this PACBO(ℓ_1).

4.3.3. *Online clustering setting*

In the last part, we have compared, in the batch setting, our method with 7 other methods on different datasets. However let us stress here that none of the aforementioned methods is specifically designed for *online* clustering. Indeed, to the best of our knowledge PACBO is the sole procedure that explicitly takes advantage of the *sequential nature* of data. For that reason, we present below the behavior and a comparison of running times between PACBO and the aforementioned methods, on the following synthetic online clustering toy example.

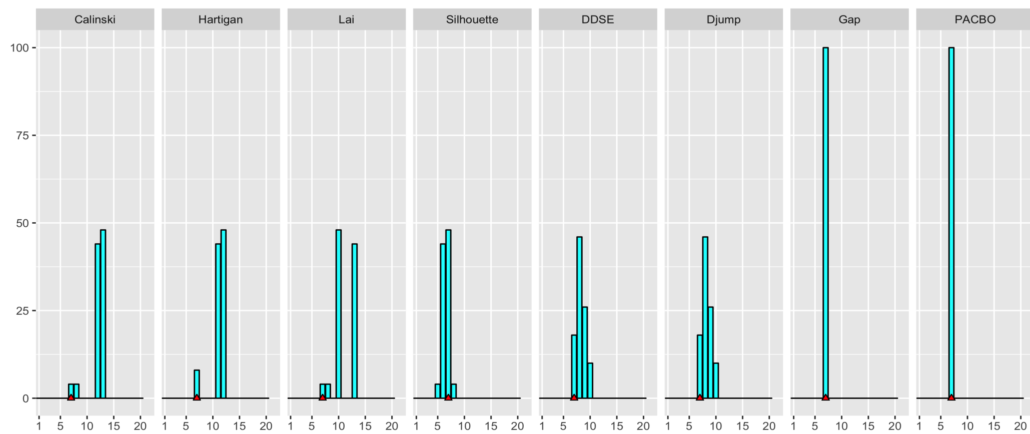


(a) Model 1.

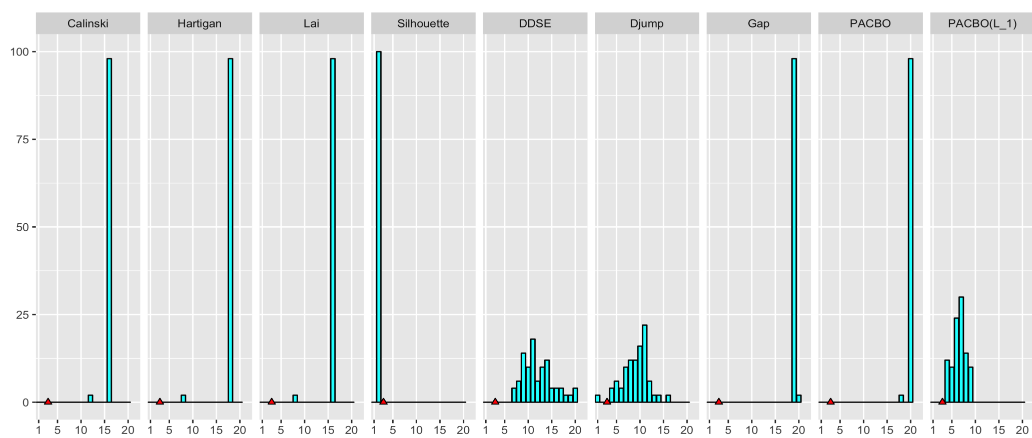


(b) Model 2.

Figure 3: Histograms of the estimated number of cells on 50 realizations. The red mark indicates the true number of cells.



(a) Model 3.



(b) Model 4.

Figure 4: Histograms of the estimated number of cells on 50 realizations. The red mark indicates the true number of cells.

Model 5 (10 mixed groups in dimension 2)**.** *Observations* $(x_t)_{t=1,\dots,T=200}$ *are simulated in the following way: define firstly for each* $t \in [\![1,T]\!]$ *a pair* $(c_{1,t}, c_{2,t}) \in \mathbb{R}^2$*, where* $c_{1,t} = -\frac{5}{2}\pi + \frac{5\pi}{9}\left(\lfloor\frac{t-1}{20}\rfloor - 1\right)$ *and* $c_{2,t} = 5\sin(c_{1,t})$*. Then for* $t \in [\![1,100]\!]$*,* x_t *is sampled from a uniform distribution on the unit cube in* $\mathbb{R}^2$*, centered at* $(c_{x,t}, c_{y,t})$*. For* $t \in [\![101,200]\!]$*,* x_t *is generated by a bivariate Gaussian distribution, centered at* $(c_{x,t}, c_{y,t})$ *with identity covariance matrix.*

In this online setting, the true number $k_t^\star$ of groups will augment of 1 unit every 20 time steps to eventually reach 10 (and the maximal number of clusters is set to 20 for all methods). Figure 5a shows ECL for PACBO and OCL along with 95% confidence intervals computed on 100 realizations with $T = 200$ observations, with $\lambda_t = 0.6 \times (d+2)/2\sqrt{t}$ and $R = 15$ (so that all observations are in the ℓ_2-ball $B_2(R)$. Jumps in the ECL occur when new clusters of data are observed. Since PACBO outputs a partition based only on the past observations, the instantaneous loss is larger whenever a new cluster appears. However PACBO quickly identifies the new cluster. This is also supported by Figure 5b which represents the true and estimated numbers of clusters.

In addition we also count the number of correct estimations of the true number $k_t^\star$ of clusters. Table 1 contains its mean (and standard deviation, on 100 repetitions) for PACBO and its seven competitors. PACBO has the largest mean by a significant margin and identifies the correct number of clusters of about 120 observations out of 200.

Calinski	Hartigan	Lai	Silhouette	DDSE	Djump	Gap	PACBO
34.92 (8.24)	63.72 (4.81)	52.23 (4.64)	72.44 (4.39)	22.73 (4.17)	38.38 (6.21)	56.73 (14.38)	**119.95 (7.08)**

TABLE 1

Mean and standard deviation of correct estimations of the true number of clusters.

Next, we compare the running times of PACBO and its competitors, in the online setting. At each time $t = 1,\dots,200$, we measure the running time of each method. Table 2 presents the mean (and standard deviation) on 100 repetitions of the total running times. The superiority of PACBO is a straightforward consequence of the fact that it adapts to the *sequential nature* of data, whereas all other methods conduct a batch clustering at each time step.

Calinski	Hartigan	Lai	Silhouette	DDSE	Djump	Gap	PACBO
46.86 (5.66)	39.27 (2.75)	52.07 (3.53)	118.44 (1.98)	33.85 (6.82)	33.85 (6.82)	207.55 (2.72)	**28.13 (4.06)**

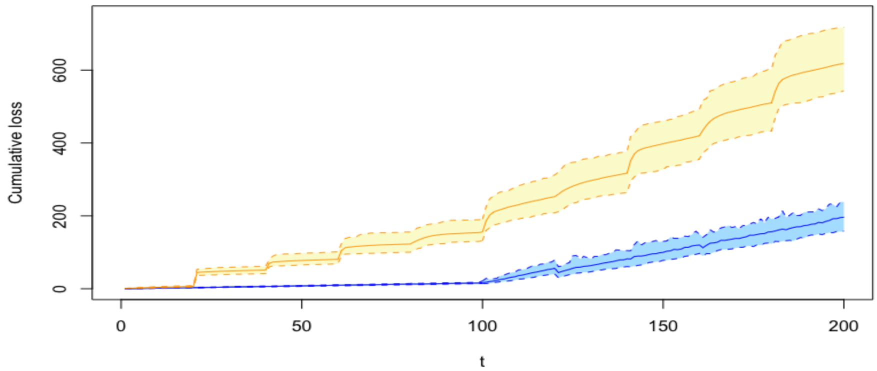
TABLE 2

Mean (and standard deviation) of total running time (in seconds).

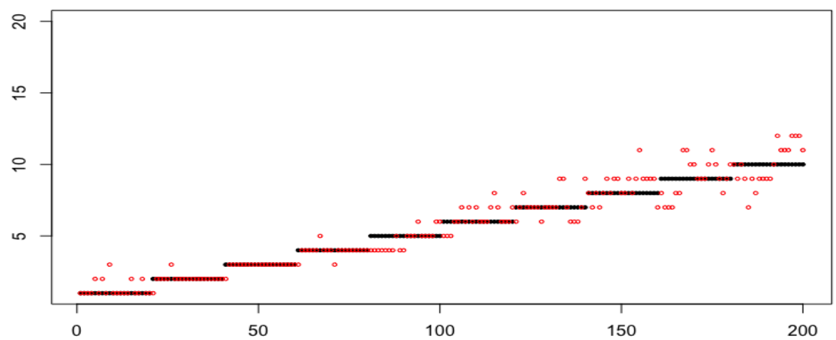
For the sake of completion, Appendix A contains an instance of the performance of all methods to estimate the true number of clusters.

5. Proofs

This section contains the proofs to all original results claimed in Section 3 and Section 4.



(a) ECL (yellow line) and OCL (blue line) as function of t, with 95% confidence intervals (dashed line).



(b) Estimated number of cells (red dots) by PACBO as a function of t. Black lines represent the true number of cells.

Figure 5: Performance of PACBO.

5.1. *Proof of Corollary 1*

Let us first introduce some notation. For any $k \in [\![1, p]\!]$ and $R > 0$, let

$$\mathscr{C}(k,R) = \left\{ \mathbf{c} = (c_j)_{j=1,\dots,k} \in \mathbb{R}^{dk} : |c_j|_2 \leq R, \forall j \right\},$$
$$\Xi(k,R) = \left\{ \xi = (\xi_j)_{j=1,\dots,k} \in \mathbb{R}^{k} : 0 < \xi_j \leq R, \forall j \right\}.$$

We denote by $\rho_k(\mathbf{c},\mathfrak{c},\xi)$ the density consisting in the product of k independent uniform distributions on ℓ_2-balls in $\mathbb{R}^d$, namely,

$$\mathrm{d}\rho_k(\mathbf{c},\mathfrak{c},\xi)=\prod_{j=1}^{k}\left\{\frac{\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}}}\left(\frac{1}{\xi_j}\right)^d \mathbb{1}_{\{B_d(\mathfrak{c}_j,\xi_j)\}}(c_j)\right\}\mathrm{d}\mathbf{c},$$

where $\mathfrak{c}\in\mathscr{C}(k,R)$, $\xi\in\Xi(k,R)$ and $B_d(\mathfrak{c}_j,\xi_j)$ is an ℓ_2-ball in $\mathbb{R}^d$, centered in $\mathfrak{c}_j$ with radius ξ_j. In the following, we will shorten $\rho_k(\mathbf{c},\mathfrak{c},\xi)$ to ρ_k when no confusion can arise. The proof relies on choosing a specific ρ in Proposition 1. For any $k\in[\![1,p]\!]$, $\mathfrak{c}\in\mathscr{C}(k,R)$ and $\xi\in\Xi(k,R)$, let $\rho=\rho_k\mathbb{1}_{\{\mathbf{c}\in\mathbb{R}^{dk}\}}$. Then ρ is a well-defined distribution on $\mathscr{C}$ and belongs to $\mathscr{P}_\pi(\mathscr{C})$. Proposition 1 yields

$$\sum_{t=1}^{T}\mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)}\ell(\hat{\mathbf{c}}_t,x_t)\le\inf_{k\in[\![1,p]\!]}\inf_{\substack{\rho\in\mathscr{P}_\pi(\mathscr{C})\\ \rho=\rho_k\mathbb{1}_{\{\mathbf{c}\in\mathbb{R}^{dk}\}}}}\left\{\mathbb{E}_{\mathbf{c}\sim\rho}\sum_{t=1}^{T}[\ell(\mathbf{c},x_t)]+\frac{\mathscr{K}(\rho,\pi)}{\lambda}\right.$$
$$\left.+\frac{\lambda}{2}\mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_T)}\mathbb{E}_{\mathbf{c}\sim\rho}\sum_{t=1}^{T}[\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t)]^2\right\}. \qquad (16)$$

For any $\rho=\rho_k\mathbb{1}_{\{\mathbf{c}\in\mathbb{R}^{dk}\}}$, the first term on the right-hand side of (16) satisfies

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_{\mathbf{c}\sim\rho}[\ell(\mathbf{c},x_t)]&=\sum_{t=1}^{T}\mathbb{E}_{\mathbf{c}\sim\rho_k}[\ell(\mathbf{c},x_t)]\\&\le\sum_{t=1}^{T}\min_{j=1,\dots,k}\left\{\mathbb{E}_{\mathbf{c}\sim\rho_k}\left[|c_j-\mathfrak{c}_j|_2^2\right]+|\mathfrak{c}_j-x_t|_2^2\right\}\\&=\sum_{t=1}^{T}\min_{j=1,\dots,k}\left\{\frac{d}{d+2}\xi_j^2+|\mathfrak{c}_j-x_t|_2^2\right\}\\&\le\frac{dT}{d+2}\max_{j=1,\dots,k}\xi_j^2+\sum_{t=1}^{T}\ell(\mathfrak{c},x_t).\end{aligned}\qquad (17)$$

Let us now compute the second term on the right-hand side of (16).

$$\begin{aligned}\mathscr{K}(\rho,\pi)&=\int_{\mathscr{C}}\log\frac{\rho(\mathbf{c})}{\pi(\mathbf{c})}\rho(\mathbf{c})\mathrm{d}\mathbf{c}\\&=\int_{\mathbb{R}^{dk}}\left(\log\frac{\rho_k(\mathbf{c})}{\pi_k(\mathbf{c})}+\log\frac{\pi_k(\mathbf{c})}{\pi(\mathbf{c})}\right)\rho_k(\mathbf{c})\mathrm{d}\mathbf{c}\\&=\mathscr{K}(\rho_k,\pi_k)+\log\frac{1}{q(k)}\\&=:A+B,\end{aligned}$$

where

$$A=\int_{\mathbb{R}^{dk}}\log\prod_{j=1}^{k}\frac{\left(\frac{1}{\xi_j}\right)^d}{\left(\frac{1}{2R}\right)^d}\rho_k(\mathbf{c})\mathrm{d}\mathbf{c}=d\sum_{j=1}^{k}\log\left(\frac{2R}{\xi_j}\right).$$

Since the function $x \mapsto (1-e^{-\eta x})/x$ is non-increasing for $x>0$ and $\eta>0$, we have

$$\begin{aligned}B &= \log\left(\frac{e^{-\eta}(1-e^{-\eta p})}{1-e^{-\eta}}e^{\eta k}\right)\\ &\le \log\left(pe^{\eta(k-1)}\right)\\ &= \eta(k-1)+\log p. \end{aligned}\tag{18}$$

When $\eta=0$, q is a uniform distribution on $[\![1,p]\!]$, and the above inequality holds as well. Then, $\mathscr{K}(\rho,\pi)/\lambda$ in (16) may be upper bounded as follows:

$$\frac{\mathscr{K}(\rho,\pi)}{\lambda} \le \frac{d}{\lambda}\sum_{j=1}^{k}\log\left(\frac{2R}{\xi_j}\right)+\frac{\eta(k-1)}{\lambda}+\frac{\log p}{\lambda}. \tag{19}$$

Finally,

$$\begin{aligned}|\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t)| &= \left|\min_{j=1,\dots,k}|c_j-x_t|_2^2-\min_{j=1,\dots,K_t}|\hat{c}_{t,j}-x_t|_2^2\right|\\ &\le \left(2R+\max_{t=1,\dots,T}|x_t|_2\right)^2 =: C_1.\end{aligned}$$

Then, the third term of the right-hand side in (16) is controlled as

$$\frac{\lambda}{2}\mathbb{E}_{(\hat{\rho}_1,\dots,\hat{\rho}_T)}\mathbb{E}_{\mathbf{c}\sim\rho_k}\sum_{t=1}^{T}[\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t)]^2 \le \frac{\lambda T}{2}C_1^2. \tag{20}$$

Combining inequalities (17), (19) and (20) gives, for any $\xi\in\Xi(k,R)$,

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\dots,\hat{\rho}_t)}\ell(\hat{\mathbf{c}}_t,x_t) \le \inf_{k\in[\![1,p]\!]}\inf_{\mathfrak{c}\in\mathscr{C}(k,R)}&\left\{\sum_{t=1}^{T}\ell(\mathfrak{c},x_t)+\frac{dT}{d+2}\max_{j=1,\dots,k}\xi_j^2\right.\\ &\left.+\frac{d}{\lambda}\sum_{j=1}^{k}\log\left(\frac{2R}{\xi_j}\right)+\frac{\eta}{\lambda}(k-1)\right\}+\frac{\lambda T}{2}C_1^2+\frac{\log p}{\lambda}.\end{aligned}$$

Under the assumption that $\lambda>(d+2)/(2TR^2)$, the global minimizer of the function

$$(\xi_1,\dots,\xi_k)\mapsto\frac{Td}{d+2}\max_{j=1,\dots,k}\xi_j^2+\frac{d}{\lambda}\sum_{j=1}^{k}\log\left(\frac{2R}{\xi_j}\right) \tag{21}$$

does not necessarily belong to $\Xi(k,R)$. A possible choice of $(\xi_j)_{1:k}\in\Xi(k,R)$ is given by

$$\xi_1^\star=\xi_2^\star=\cdots=\xi_k^\star=\sqrt{\frac{d+2}{2\lambda T}}.$$

Then (21) amounts to

$$\frac{d}{2\lambda}+\frac{dk}{2\lambda}\log\left(\frac{8R^2\lambda T}{d+2}\right).$$

Hence,

$$\sum_{t=1}^{T}\mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)}\ell(\hat{\mathbf{c}}_t,x_t)\le\inf_{k\in[\![1,p]\!]}\inf_{\mathfrak{c}\in\mathscr{C}(k,R)}\left\{\sum_{t=1}^{T}\ell(\mathfrak{c},x_t)+\frac{dk}{2\lambda}\log\left(\frac{8R^2\lambda T}{(d+2)k}\right)+\frac{\eta}{\lambda}k\right\}$$
$$+\left(\frac{\log p}{\lambda}+\frac{d}{2\lambda}+\frac{\lambda T}{2}C_1^2\right).$$

5.2. Proof of Theorem 1

The proof builds upon the online variance inequality described in Audibert (2009), *i.e.*, for any $\lambda>0$, any $\hat\rho\in\mathscr{P}_\pi(\mathscr{C})$ and any $x\in\mathbb{R}^d$,

$$\mathbb{E}_{\mathbf{c}'\sim\hat\rho}[\ell(\mathbf{c}',x)]\le-\frac{1}{\lambda}\mathbb{E}_{\mathbf{c}'\sim\hat\rho}\log\mathbb{E}_{\mathbf{c}\sim\hat\rho}\left[e^{-\lambda\left[\ell(\mathbf{c},x)+\frac{\lambda}{2}\left(\ell(\mathbf{c},x)-\ell(\mathbf{c}',x)\right)^2\right]}\right].\tag{22}$$

By (22), we have

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)}\ell(\hat{\mathbf{c}}_t,x_t)&=\sum_{t=1}^{T}\mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_{t-1})}\mathbb{E}_{\hat\rho_t}\left[\ell(\hat{\mathbf{c}}_t,x_t)\mid\hat{\mathbf{c}}_1,\dots,\hat{\mathbf{c}}_{t-1}\right]\\&\le\sum_{t=1}^{T}\mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_{t-1})}\left[-\frac{1}{\lambda_{t-1}}\mathbb{E}_{\hat{\mathbf{c}}_t\sim\hat\rho_t}\log\mathbb{E}_{\mathbf{c}\sim\hat\rho_t}\left(e^{-\lambda_{t-1}[\ell(\mathbf{c},x_t)+\frac{\lambda_{t-1}}{2}(\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t))^2]}\right)\right]\\&\le\mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_T)}\left[\sum_{t=1}^{T}-\frac{1}{\lambda_{t-1}}\log\frac{\int e^{-\lambda_{t-1}S_t(\mathbf{c})}\mathrm{d}\pi(\mathbf{c})}{\int e^{-\lambda_{t-1}S_{t-1}(\mathbf{c})}\mathrm{d}\pi(\mathbf{c})}\right]\\&=\mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_T)}\left[\sum_{t=1}^{T}-\frac{1}{\lambda_{t-1}}\log\frac{V_t}{W_{t-1}}\right]\\&=\mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_T)}\left[\sum_{t=1}^{T}\left[\frac{1}{\lambda_{t-1}}\log W_{t-1}-\frac{1}{\lambda_{t-1}}\log V_t\right]\right].\end{aligned}\tag{23}$$

Applying Jensen's inequality, for any $1\le t\le T$,

$$\begin{aligned}\frac{1}{\lambda_{t-1}}\log V_t&=\frac{1}{\lambda_{t-1}}\log\mathbb{E}_{\mathbf{c}\sim\pi}\left[\left(e^{-\lambda_tS_t(\mathbf{c})}\right)^{\frac{\lambda_{t-1}}{\lambda_t}}\right]\\&\ge\frac{1}{\lambda_{t-1}}\log\left(\mathbb{E}_{\mathbf{c}\sim\pi}\left[e^{-\lambda_tS_t(\mathbf{c})}\right]\right)^{\frac{\lambda_{t-1}}{\lambda_t}}\\&=\frac{1}{\lambda_t}\log W_t.\end{aligned}$$

Therefore, since $W_0=1$,

$$\sum_{t=1}^{T}\left[\frac{1}{\lambda_{t-1}}\log W_{t-1}-\frac{1}{\lambda_{t-1}}\log V_t\right]\le-\frac{1}{\lambda_T}\log W_T,\tag{24}$$

and by (23), (24) and the duality formula (3), we have

$$\sum_{t=1}^{T}\mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)}\ell(\hat{\mathbf{c}}_t,x_t)\le\mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_T)}\left[-\frac{1}{\lambda_T}\log\mathbb{E}_{\mathbf{c}\sim\pi}\left[e^{-\lambda_TS_T(\mathbf{c})}\right]\right]$$

$$
\begin{aligned}
&\leq -\frac{1}{\lambda_T}\log \mathbb{E}_{\mathbf{c}\sim\pi}\left[e^{-\lambda_T \mathbb{E}_{(\hat{\rho_1},\ldots,\hat{\rho}_T)} S_T(\mathbf{c})}\right] \qquad \text{(by Audibert, 2009, Lemma 3.2)}\\
&= \inf_{\rho\in\mathscr{P}_\pi(\mathscr{C})}\left\{\mathbb{E}_{\mathbf{c}\sim\rho}\left[\sum_{t=1}^{T}\ell(\mathbf{c},x_t)\right] + \mathbb{E}_{\mathbf{c}\sim\rho}\mathbb{E}_{(\hat{\rho}_1,\ldots,\hat{\rho}_T)}\left[\sum_{t=1}^{T}\frac{\lambda_{t-1}}{2}(\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t))^2\right]\right.\\
&\quad\left. + \frac{\mathscr{K}(\rho,\pi)}{\lambda_T}\right\},
\end{aligned}
$$

which achieves the proof.

5.3. *Proof of Corollary 3*

The proof is similar to the proof of Corollary 1, the only difference lies in the fact that (20) is replaced with

$$
\begin{aligned}
\mathbb{E}_{(\hat{\rho}_1,\ldots,\hat{\rho}_T)}\mathbb{E}_{\mathbf{c}\sim\rho_k}\sum_{t=1}^{T}\frac{\lambda_{t-1}}{2}[\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t)]^2 &\leq \frac{(d+2)C_1^2}{4R^2}\left(1+\sum_{t=2}^{T}\frac{\sqrt{\log(t-1)}}{\sqrt{t-1}}\right)\\
&\leq \frac{(d+2)C_1^2}{4R^2}\left(1+\frac{\sqrt{\log 2}}{\sqrt{2}}+\frac{\sqrt{\log 3}}{\sqrt{3}}+\sum_{t=4}^{T-1}\int_{t-1}^{t}\frac{\sqrt{\log x}}{\sqrt{x}}\mathrm{d}x\right)\\
&\leq \frac{(d+2)C_1^2}{2R^2}\sqrt{T\log T},
\end{aligned}
$$

where the second inequality above is due to the fact that $\frac{\sqrt{\log t}}{\sqrt{t}} \leq \int_{t-1}^{t}\frac{\sqrt{\log x}}{\sqrt{x}}\mathrm{d}x$ when $t\geq 4$ and the last inequality is deduced from the following with change of variable $y=\sqrt{\log x}$, *i.e.*,

$$
\begin{aligned}
\int_3^{T-1}\frac{\sqrt{\log x}}{\sqrt{x}}\mathrm{d}x &= \int_{\sqrt{\log 3}}^{\sqrt{\log(T-1)}} 2y^2 e^{\frac{y^2}{2}}\mathrm{d}y\\
&\leq \sqrt{\log(T-1)}\int_{\sqrt{\log 3}}^{\sqrt{\log(T-1)}} 2y e^{\frac{y^2}{2}}\mathrm{d}y\\
&= 2\sqrt{\log(T-1)}\left(\sqrt{T-1}-\sqrt{3}\right).
\end{aligned}
$$

5.4. *Proof of Corollary 4*

Let us denote by M the index of the last epoch and let $t_M = T$. We assume $M\geq 1$ (otherwise, the corollary follows directly from Corollary 3 applied with an upper bound R_0 of ℓ_2-norm of sequence $(x_t)_{1:T}$). If $R_{t_M}\leq R_{t_{M-1}}$, then we have $R_T = R_{t_M} = R_{t_{M-1}}$, hence one always has $R_{t_M}\geq R_{t_{M-1}}$. In addition, since $M\geq 1$, we also have $R_{t_M}\leq 2\max_{t=1,\ldots,T}|x_t|_2 = 2R$.

Let us introduce for each epoch $r, r=0,1,\ldots,M$ the following notation

$$
E^{(r)} = \sum_{t=t_{r-1}+1}^{t_r-1}\mathbb{E}_{(\hat{\rho}_1,\ldots,\hat{\rho}_t)}\ell(\hat{\mathbf{c}}_t,x_t),
$$

and for $k \in [\![1,p]\!]$, $\mathbf{c} \in \mathscr{C}(k,R)$

$$L^{(r)}(k,\mathbf{c}) = \sum_{t=t_{r-1}+1}^{t_r-1} \ell(\mathbf{c},x_t).$$

Within each epoch $r = 0,1,\dots,M$, since

$$\max_{t=t_{r-1}+1,t_{r-1}+2,\dots,t_r-1} |x_s|_2 \le R_{t_{r-1}}, \tag{25}$$

then applying Corollary 3 to each epoch r can give us that, for each $k \in [\![1,p]\!]$,

$$E^{(r)} - \inf_{\mathbf{c}\in\mathscr{C}\left(k,R_{t_{r-1}}\right)} L^{(r)}(k,\mathbf{c}) \le \left(C(d,\eta)k + C(p,d)\right) R^2_{t_{r-1}} \sqrt{(t_r-1)\log(t_r-1)}, \tag{26}$$

where $C(d,\eta) = \frac{2(d+\eta)}{d+2}$ and $C(p,d) = \frac{2\log p+d}{d+2} + \frac{81(d+2)}{2}$.

In addition, since all observations $x_t, t = t_{r-1}+1,,\dots,t_r-1$ in the epoch r are bounded in a convex ball $B_d\left(R_{t_{r-1}}\right)$, centered in $0 \in \mathbb{R}^d$ with radius $R_{t_{r-1}}$ as indicated by (25), we have for each $\mathbf{c}' \in \mathscr{C}(k,R) \setminus \mathscr{C}\left(k,R_{t_{r-1}}\right)$, $k = 1,2,\dots,p$ that

$$\inf_{\mathbf{c}\in\mathscr{C}\left(k,R_{t_{r-1}}\right)} L^{(r)}(k,\mathbf{c}) \le L^{(r)}(k,\mathbf{c}'). \tag{27}$$

By (26) and (27), we can have that for any $k \in [\![1,p]\!]$ and $\mathbf{c} \in \mathscr{C}(k,R)$, the following inequality holds,

$$E^{(r)} - L^{(r)}(k,\mathbf{c}) \le \left(C(d,\eta)k + C(p,d)\right) R^2_{t_{r-1}} \sqrt{(t_r-1)\log(t_r-1)}.$$

Therefore, for any $\mathbf{c} \in \mathscr{C}(k,R)$, one has

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t,x_t) - \sum_{t=1}^{T} \ell(\mathbf{c},x_t) &= \sum_{r=0}^{M} \left(E^{(r)} - L^{(r)}(k,\mathbf{c})\right) + \sum_{r=0}^{M} \left(\mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_{t_r})} \ell\left(\hat{\mathbf{c}}_{t_r},x_{t_r}\right) - \ell(\mathbf{c},x_{t_r})\right) \\
&\le \sum_{r=0}^{M} \left[C(d,\eta)k + C(p,d)\right] R^2_{t_{r-1}} \sqrt{(t_r-1)\log(t_r-1)} + 4\sum_{r=0}^{M} R^2_{t_r} \\
&\le \sum_{r=0}^{M} \left[C(d,\eta)k + C(p,d)\right] R^2_{t_{r-1}} \sqrt{T\log T} + 4\sum_{r=0}^{M} R^2_{t_r}.
\end{aligned}$$

Since $R_{t_s} \ge 2^{s-r} R_{t_r}$ for $0 \le r \le s \le M-1$, then for $s \le M-1$,

$$\sum_{r=0}^{s} R^2_{t_r} \le \sum_{r=0}^{s} 4^{r-s} R^2_{t_s} \le \frac{4}{3} R^2_{t_s}.$$

Hence,

$$\sum_{r=0}^{M} R^2_{t_{r-1}} \le R^2_{t_{-1}} + \frac{4}{3} R^2_{t_{M-1}} \le \frac{7}{3} R^2_{t_M}$$

$$4\sum_{r=0}^{M} R_{t_r}^2 \leq 4\left(\frac{4}{3}R_{t_{M-1}}^2 + R_{t_M}^2 + \right) \leq \frac{28}{3}R_{t_M}^2$$

Therefore,

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\ldots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) - \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) \leq \frac{7}{3}\left[C(d,\eta)k + C(p,d)\right] R_{t_M}^2 \sqrt{T\log T} + \frac{28}{3}R_{t_M}^2$$
$$\leq \frac{28}{3}\left[C(d,\eta)k + C(p,d)\right] R^2 \sqrt{T\log T} + \frac{112}{3}R^2,$$

where $R = \max_{t=1,2,\ldots,T} |x_t|_2$ and the second inequality is due to the fact that $R_{t_M} \leq 2R$. Taking the infimum of $\sum_{t=1}^{T} \ell(\mathbf{c}, x_t)$ over the set $\mathscr{C}(k,R)$, $k \in [\![1,p]\!]$ leads to

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\ldots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{\mathbf{c}\in\mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \frac{28}{3}\left[C(d,\eta)k + C(p,d)\right] R^2 \sqrt{T\log T} + \frac{112}{3}R^2.$$

Finally, taking the infimum of the right hand side of the above inequality with respect to k terminates the proof.

5.5. *Proof of Theorem 2*

The proof for the upper bound is straightforward: by replacing the loss function $\ell(\mathbf{c}, x)$ by the penalized loss $\ell_\alpha(\mathbf{c}, x) = \ell(\mathbf{c}, x) + \alpha|\mathbf{c}|$ with $\alpha = \sqrt{\log T}/\sqrt{T}$ in the proof of Theorem 1, we obtain

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{\rho}_1,\ldots,\hat{\rho}_t} \ell_\alpha(\hat{\mathbf{c}}_t, x_t) \leq \inf_{\rho\in\mathscr{P}_\pi(\mathscr{C})} \Bigg\{ \mathbb{E}_{\mathbf{c}\sim\rho}\left[\sum_{t=1}^{T} \ell_\alpha(\mathbf{c}, x_t)\right] + \frac{\mathscr{K}(\rho,\pi)}{\lambda_T}$$
$$+ \mathbb{E}_{(\hat{\rho}_1,\ldots,\hat{\rho}_T)} \mathbb{E}_{\mathbf{c}\sim\rho}\left[\sum_{t=1}^{T} \frac{\lambda_{t-1}}{2}\left[\ell_\alpha(\mathbf{c}, x_t) - \ell_\alpha(\hat{\mathbf{c}}_t, x_t)\right]^2\right]\Bigg\},$$

and choosing $\lambda = \sqrt{\log T}/\sqrt{T}$ and $p = T^{\frac{1}{4}}$ yields the desired upper bound.

We now proceed to the proof of the lower bound. The trick is to replace the supremum over the (x_t) in $\mathscr{V}_T(s)$ by an expectation.

We first introduce the event $\Omega_{s,R} = \left\{(X_1,\ldots,X_T) \in \mathbb{R}^{dT} :\text{ such that } \left|\mathbf{c}_{T,R}^\star\right| = s\right\}$, where $\mathbf{c}_{T,R}^\star$ is defined as in **Assumption** $\mathscr{H}(s)$. Then, we have

$$\mathscr{V}_T(s) \geq \inf_{(\hat{\rho}_t)} \mathbb{E}_{\mu^T}\left\{\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\ldots,\hat{\rho}_t)}\left(\ell(\hat{\mathbf{c}}_t, X_t) + \frac{\sqrt{\log T}}{\sqrt{T}}|\hat{\mathbf{c}}_t|\right) - \inf_{\mathbf{c}\in\mathscr{C}(s,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, X_t)\right\} \mathbb{1}\left(\Omega_{s,R}\right),$$

where $\mu^T \in \mathscr{P}(\mathbb{R}^{dT})$ is the joint distribution of i.i.d. sample $(X_1,\ldots,X_T)$. Now, we have to choose μ in order to maximize the right-hand side of the above inequality. This is the purpose of the following lemmas.

Lemma 2. *Let $s \in \mathbb{N}^*, s \leq p$. Let $\mu \in \mathscr{P}(\mathbb{R}^d)$ a distribution concentrated on $2s$ fixed points $\mathscr{S}_\mu = \{z_i, z_i + w, i = 1, \dots, s\}$ such that $w = (2\Delta, 0, \dots, 0) \in \mathbb{R}^d$ with $\Delta > 0$ and that $z_1, \dots, z_s \in B_d(R)$. Suppose that for any $i \neq j$, $d(z_i, z_j) \geq 2A\Delta$ for some $A > 0$. Define μ as the uniform distribution over $\mathscr{S}_\mu$. Then, if $A > \sqrt{2} + 1$, we have*

$$\arg \inf_{\mathbf{c} \in \mathscr{C}(s,R)} \mathbb{E}_\mu \ell(\mathbf{c}, X) = \{z_i + w/2, \quad i = 1, \dots, s\} =: \mathbf{c}^\star_{\mu,s}.$$

The proof of Lemma 2 is similar to Bartlett et al. (1998, Section III.A, step 3). The next lemma controls the probability of the event $|\mathbf{c}^\star_{T,R}| \neq s$ with a proper choice of Δ^2 and A in the definition of μ.

Lemma 3. *Let $s \in \mathbb{N}^*$, $2 \leq s \leq p$, and μ is defined in Lemma 2. Then, if we choose $A = \sqrt{2}s + 1$ and*

$$\frac{2(s-1)s\sqrt{\log T}}{(A-1)^2\sqrt{T}} < \Delta^2 < \frac{\sqrt{\log T}}{\sqrt{T}},$$

then for any $\epsilon > 0$ and $T > 8s^2 \log \frac{2s^2}{\epsilon}$, we have

$$\mathbb{P}\left(\left| \mathbf{c}^\star_{T,R} \right| \neq s \right) \leq \epsilon.$$

Proof. For any $k \in [\![1, p]\!]$, let $\mathbf{c}^\star_{T,k}$ firstly denote the optimal partition in $\mathscr{C}(k,R)$ that minimizes the penalized empirical loss on $(X_1, \dots, X_T)$, *i.e.*,

$$\mathbf{c}^\star_{T,k} = \arg \inf_{\mathbf{c} \in \mathscr{C}(k,R)} \left\{ \frac{1}{T} \sum_{t=1}^T \ell(\mathbf{c}, X_t) + |\mathbf{c}| \frac{\sqrt{\log T}}{\sqrt{T}} \right\}.$$

In addition, denote by $\mathbf{c}^\star_{\mu,k}$ the partition minimizing the expected penalized loss, *i.e.*,

$$\mathbf{c}^\star_{\mu,k} = \arg \inf_{\mathbf{c} \in \mathscr{C}(k,R)} \left\{ \mathbb{E}_\mu \ell(\mathbf{c}, X) + |\mathbf{c}| \frac{\sqrt{\log T}}{\sqrt{T}} \right\}.$$

One can notice that in fact $|\mathbf{c}| = k$ in the two above definitions for any $\mathbf{c} \in \mathscr{C}(k,R) \in \mathbb{R}^{dk}$. Next

$$\begin{aligned}
\mathbb{P}\left(\left| \mathbf{c}^\star_{T,R} \right| > s \right) &= \sum_{k=s+1}^{2s} \mathbb{P}\left(\left| \mathbf{c}^\star_{T,R} \right| = k \right) \\
&\leq \sum_{k=s+1}^{2s} \mathbb{P}\left(\frac{1}{T} \sum_{t=1}^T \ell\left(\mathbf{c}^\star_{T,k-1}, X_t\right) - \frac{1}{T} \sum_{t=1}^T \ell\left(\mathbf{c}^\star_{T,k}, X_t\right) > \sqrt{\frac{\log T}{T}} \right) \\
&\leq \sum_{k=s+1}^{2s} \mathbb{P}\left(\frac{1}{T} \sum_{t=1}^T \ell\left(\mathbf{c}^\star_{T,k-1}, X_t\right) > \sqrt{\frac{\log T}{T}} \right) \\
&\leq s\mathbb{P}\left(\frac{1}{T} \sum_{t=1}^T \ell\left(\mathbf{c}^\star_{\mu,s}, X_t\right) > \sqrt{\frac{\log T}{T}} \right) \\
&= s\mathbb{P}\left(\Delta^2 > \sqrt{\frac{\log T}{T}} \right) = 0,
\end{aligned} \tag{28}$$

where the first inequality is induced by the definition of $\mathbf{c}^{\star}_{T,R}$ and the third inequality is due to the fact that we have almost surely

$$\sum_{t=1}^{T} \ell\left(\mathbf{c}^{\star}_{\mu,s}, X_t\right) \geq \sum_{t=1}^{T} \ell\left(\mathbf{c}^{\star}_{T,s}, X_t\right) \geq \sum_{t=1}^{T} \ell\left(\mathbf{c}^{\star}_{T,k-1}, X_t\right), \quad \text{for } k > s.$$

In order to control the probability $\mathbb{P}(|\mathbf{c}^{\star}_{T,R}| < s)$, let us first consider the Voronoi partition of $\mathbb{R}^d$ induced by the set of points $\{z_i, z_i + w, i = 1, \ldots, s\}$ and for each i define V_i as the union of the Voronoi cells belonging to z_i and $z_i + w$. Let N_i denotes the number of X_t, $t = 1, \ldots, T$ falling in V_i. Hence $(N_1, \ldots, N_s)$ follows a multinomial distribution with parameter $(T, q_1, q_2, \ldots, q_s)$, where $q_1 = q_2 = \cdots = q_s = 1/s$. Then

$$\begin{aligned}
\mathbb{P}\left(\left|\mathbf{c}^{\star}_{T,R}\right| < s\right) &= \sum_{k=1}^{s-1} \mathbb{P}\left(\left|\mathbf{c}^{\star}_{T,R}\right| = k\right) \\
&\leq \sum_{k=1}^{s-1} \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} \ell\left(\mathbf{c}^{\star}_{T,k}, X_t\right) - \frac{1}{T}\sum_{t=1}^{T} \ell\left(\mathbf{c}^{\star}_{T,s}, X_t\right) \leq \frac{(s-k)\sqrt{\log T}}{\sqrt{T}}\right) \\
&\leq \sum_{k=1}^{s-1} \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} \ell\left(\mathbf{c}^{\star}_{T,k}, X_t\right) - \frac{1}{T}\sum_{t=1}^{T} \ell\left(\mathbf{c}^{\star}_{\mu,s}, X_t\right) \leq \frac{(s-k)\sqrt{\log T}}{\sqrt{T}}\right) \\
&\leq (s-1)\mathbb{P}\left(\frac{1}{T}\min_{i=1,\ldots,s} N_i \cdot (A-1)^2\Delta^2 - \Delta^2 \leq \frac{(s-k)\sqrt{\log T}}{\sqrt{T}}\right) \\
&\leq (s-1)s\mathbb{P}\left(N_1 \leq \frac{T\Delta^2 + (s-1)\sqrt{T\log T}}{(A-1)^2\Delta^2}\right).
\end{aligned}$$

The third inequality is due to the fact that $\sum_{t=1}^{T} \ell\left(\mathbf{c}^{\star}_{T,k}, X_t\right) \geq \min_{i=1,\ldots,s} N_i (A-1)^2\Delta^2$ for $k < s$, and the last inequality holds since the marginal distributions of the N_is ($i = 1, \ldots, s$) are the same binomial distribution with parameter $(T, 1/s)$. Finally, we can bound the last term by Hoeffding's inequality, *i.e.,* for any $t > 0$

$$\mathbb{P}(N_1 - \mathbb{E}(N_1) \leq -t) \leq 2\exp\left(-\frac{2t^2}{T}\right).$$

Hoeffding's inequality implies that if $s > 2, A = \sqrt{2}s + 1, T > 8s^2 \log\frac{2s^2}{\epsilon}$ and $\Delta^2 > \frac{2s(s-1)\sqrt{\log T}}{(A-1)^2\sqrt{T}}$, then

$$\mathbb{P}\left(N_1 \leq \frac{T\Delta^2 + (s-1)\sqrt{T\log T}}{(A-1)^2\Delta^2}\right) < \frac{\epsilon}{s^2}.$$

□

Next, we proceed to the proof of Theorem 2. First of all, since $(X_1, \ldots, X_T)$ are i.i.d, following the distribution μ and by the definition of $\Omega_{s,R}$, we can write

$$\inf_{(\hat{\rho}_t)} \mathbb{E}_{\mu^T}\left\{\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\ldots,\hat{\rho}_t)}\left(\ell(\hat{\mathbf{c}}_t, X_t) + \sqrt{\frac{\log T}{T}}|\hat{\mathbf{c}}_t|\right)\right\} \mathbb{1}\left(\Omega_{s,R}\right)$$

$$
\begin{aligned}
&= \inf_{(\hat\rho_t)} \mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_T)} \sum_{t=1}^{T} \mathbb{E}_{\mu^T} \left[\left(\ell(\hat{\mathbf{c}}_t, X_t) + \sqrt{\frac{\log T}{T}} |\hat{\mathbf{c}}_t| \right) \mathbb{1}(\Omega_{s,R}) \right] \\
&\geq \inf_{\hat{\mathbf{c}}} \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^{T} \ell(\hat{\mathbf{c}}, X_t) + \sqrt{T\log T} |\hat{\mathbf{c}}| \right\} \mathbb{1}\left(\Omega_{s,R}\right) \\
&\geq \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c}^\star_{T,R}, X_t) + s\sqrt{T\log T} \right\} \mathbb{1}\left(\Omega_{s,R}\right) \\
&\geq \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c}^\star_{T,R}, X_t) \right\} \left(1 - \mathbb{1}\left(\Omega^C_{s,R}\right)\right) + s\sqrt{T\log T}\,\mathbb{P}\left(\Omega_{s,R}\right) \\
&\geq \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c}^\star_{T,R}, X_t) \right\} - T\Delta^2 \mathbb{P}\left(\Omega^C_{s,R}\right) + s\sqrt{T\log T}\left(\mathbb{P}\left(\Omega_{s,R}\right) - \mathbb{P}\left(\Omega^C_{s,R}\right)\right) \\
&\geq T \inf_{\mathbf{c}\in\mathscr{C}(s,R)} \mathbb{E}_\mu \ell(\mathbf{c}, X) - T\Delta^2 \mathbb{P}\left(\Omega^C_{s,R}\right) + s\sqrt{T\log T}\left(\mathbb{P}\left(\Omega_{s,R}\right) - \mathbb{P}\left(\Omega^C_{s,R}\right)\right),
\end{aligned}
$$

where $\hat{\mathbf{c}}$ in the first inequality is given by

$$
\hat{\mathbf{c}} = \arg\inf_{\mathbf{c}\in\mathscr{C}} \mathbb{E}_{\mu^T} \left[\left(\ell(\mathbf{c}, X_t) + |\mathbf{c}| \sqrt{\log T}/\sqrt{T} \right) \mathbb{1}(\Omega_{s,R}) \right].
$$

Note that $\hat{\mathbf{c}}$ does not depend on t since μ is a symmetric uniform distribution (definition in Lemma 2). The second inequality is due to Jensen's inequality and the fourth inequality relies on the fact that with the definition of $\mathbf{c}^\star_{T,R}$ and μ, we have almost surely that

$$
\sum_{t=1}^{T} \ell\left(\mathbf{c}^\star_{T,R}, X_t\right) \leq \sum_{t=1}^{T} \ell\left(\mathbf{c}^\star_{\mu,s}, X_t\right) + s\sqrt{T\log T} = T\Delta^2 + s\sqrt{T\log T},
$$

where $\Delta > 0$ is related with the choice of μ in Lemma 2 and its value is constrained according to Lemma 3. Then we obtain for any $\epsilon > 0$

$$
\inf_{(\hat\rho_t)} \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^{T} \mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t, X_t) + \frac{\sqrt{\log T}}{\sqrt{T}} |\hat{\mathbf{c}}_t| \right\} \mathbb{1}\left(\Omega_{s,R}\right) \geq T \inf_{\mathbf{c}\in\mathscr{C}(s,R)} \mathbb{E}_\mu \ell(\mathbf{c}, X) - T\epsilon\Delta^2 + s\sqrt{T\log T}(1-2\epsilon). \tag{29}
$$

Moreover, by Jensen's inequality

$$
\mathbb{E}_{\mu^T} \left[\inf_{\mathbf{c}\in\mathscr{C}(s,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, X_t) \mathbb{1}\left(\Omega_{s,R}\right) \right] \leq T \inf_{\mathbf{c}\in\mathscr{C}(s,R)} \mathbb{E}_\mu \ell(\mathbf{c}, X). \tag{30}
$$

Combining (29) and (30), we obtain

$$
\mathscr{V}_T(s) \geq s\sqrt{T\log T}\left(1 - 2\epsilon\left[1 + \frac{\sqrt{T}\Delta^2}{2s\sqrt{\log T}}\right]\right). \tag{31}
$$

Furthermore, by taking $\epsilon = 1/T$ and choosing the minimum value of Δ^2 allowed in Lemma 3, (31) yields

$$
\mathscr{V}_T(s) \geq s\sqrt{T\log T}\left(1 - \frac{2}{T}\left[1 + \frac{s-1}{2s^2}\right]\right).
$$

Finally, we need to ensure that s pairs of points $\{z_i, z_i + w\}$ can be packed in $B_d(R)$ such that the distance between any two of the z_is is at least $2A$. A sufficient condition (Kolmogorov and Tikhomirov, 1961) is

$$s \leq \left(\frac{R - 2\Delta}{2A\Delta}\right)^d.$$

If $\Delta \leq R/6$ (which is satisfied if T is large enough), the above inequality holds if

$$s \leq \left(\frac{R}{3A\Delta}\right)^d$$

As $A = \sqrt{2}s + 1$ and $\Delta^2 < \sqrt{\log T}/\sqrt{T}$, we get the desired result.

5.6. Proof of Lemma 1

Let D_n denote the event that no "within-model move" is ever accepted in the first n moves. Then $D_1 = D_1^{\text{within}} \cup D_1^{\text{between}}$, where D_1^{within} stands for the event that a "within-model move" is proposed but rejected in one step and D_1^{between} that a "between-model move" is proposed in one step. Then we have

$$\begin{aligned}\mathbb{P}\left[D_1 | (k^{(0)}, \mathbf{c}^{(0)}) = (k, \mathbf{c})\right] =& \mathbb{P}\left[k' \neq k | (k, \mathbf{c})\right] + \mathbb{P}\left[k' = k, \text{but rejected} | (k, \mathbf{c})\right] \\ =& \frac{2}{3} + \frac{1}{3}\left[1 - \int_{\mathbb{R}^{dk}} \alpha\left[(k, \mathbf{c}), (k, \mathbf{c}')\right] \rho_k\left(\mathbf{c}', \mathfrak{c}_k, \tau_k\right) \mathrm{d}\mathbf{c}'\right],\end{aligned}$$

where

$$\begin{aligned}\alpha\left[(k, \mathbf{c}), (k, \mathbf{c}')\right] &= \min\left\{1, \frac{\hat{\rho}_t(\mathbf{c}')\rho_k(\mathbf{c}, \mathfrak{c}_k, \tau_k)}{\hat{\rho}_t(\mathbf{c})\rho_k(\mathbf{c}', \mathfrak{c}_k, \tau_k)}\right\} \\ &= \min\left\{1, h_t\left(\mathbf{c}' | (k, \mathbf{c})\right)\right\}.\end{aligned}$$

Under the assumption of $k' = k$, we have that $\mathbf{c}', \mathbf{c} \in \mathbb{R}^{dk}$, therefore the restriction of $\hat{\rho}_t$ to $\mathbb{R}^{dk}$ is well defined. Moreover, by the definition of π_k in (7), the support of the restriction of $\hat{\rho}_t$ to $\mathbb{R}^{dk}$ is $\mathbb{R}^{dk} \cap \mathscr{E} = (B_d(2R))^k$. Hence the function $(\mathbf{c}', \mathbf{c}) \mapsto h_t\left(\mathbf{c}' | (k, \mathbf{c})\right)$ is strictly positive and continuous on the compact set $(B_d(2R))^k \times (B_d(2R))^k$. As a consequence, the minimum of $h_t\left(\mathbf{c}' | (k, \mathbf{c})\right)$ on $(B_d(2R))^k \times (B_d(2R))^k$ is achieved and we denote it by m_k, *i.e.*,

$$m_k = \inf_{\mathbf{c}', \mathbf{c} \in (B_d(2R))^k} h_t\left(\mathbf{c}' | (k, \mathbf{c})\right) > 0.$$

In addition, due to the continuity and positivity of ρ_k on $\mathbb{R}^{dk}$, it is clear that for any $k \in [\![1, p]\!]$

$$z_k = \int_{(B_d(2R))^k} \rho_k\left(\mathbf{c}', \mathfrak{c}_k, \tau_k\right) \mathrm{d}\mathbf{c}' > 0.$$

Therefore, for any k,

$$\int_{\mathbb{R}^{dk}} \alpha\left[(k, \mathbf{c}), (k, \mathbf{c}')\right] \rho_k\left(\mathbf{c}', \mathfrak{c}_k, \tau_k\right) \mathrm{d}\mathbf{c}' \geq \inf_{k \in [\![1, p]\!]} (m_k z_k)$$

$$=: m^\star > 0.$$

Hence, uniformly on $k \in [\![1,p]\!]$ and $\mathbf{c} \in \mathbb{R}^{dk} \cap \mathscr{E}$, we have,

$$\mathbb{P}[D_1|(k,\mathbf{c})] \leq \left[\frac{2}{3} + \frac{1}{3}(1-m^\star)\right] < 1.$$

To conclude,

$$\mathbb{P}[D|(k,\mathbf{c})] = \lim_{n\longrightarrow\infty} \mathbb{P}[D_n|(k,\mathbf{c})] \leq \lim_{n\longrightarrow\infty} \left[\frac{2}{3} + \frac{1}{3}(1-m^\star)\right]^n = 0.$$

5.7. *Proof of Theorem 3*

For any $\mathbf{c} \in \mathscr{E}$, there exists some $k \in [\![1,p]\!]$ such that $\mathbf{c} \in (B_d(2R))^k \subset \mathscr{E}$. For any $k' \in [\![k-1,k+1]\!]$ and for any $A \in \mathscr{B}\left(\mathbb{R}^{dk'}\right)$ such that $\hat{\rho}_t(A) > 0$, the transition kernel H of the chain is given by

$$H\big(\mathbf{c},\mathbf{c}' \in A\big) = \int \mathbb{1}_{\{v_1 \in A\}} \alpha\left[(k,\mathbf{c}),(k',v_1)\right] q(k,k')\rho_{k'}(v_1,\mathfrak{c}_{k'},\tau_{k'})\mathrm{d}v_1 + r(\mathbf{c})\delta_{\mathbf{c}}(A), \quad (32)$$

where $\rho_{k'}(\cdot,\mathfrak{c}_{k'},\tau_{k'})$ is the multivariate Student distribution in (15) and

$$r(\mathbf{c}) = \sum_{k' \in [\![k-1,k+1]\!]} q(k,k') \int \left(1 - \alpha\left[(k,\mathbf{c}),(k',v_1)\right]\right) \rho_{k'}(v_1,\mathfrak{c}_{k'},\tau_{k'})\mathrm{d}v_1$$

is the probability of rejection when starting at state $\mathbf{c}$, and $\delta_{\mathbf{c}}(\cdot)$ is a Dirac measure in $\mathbf{c}$. One can easily note that $H(\mathbf{c},\mathbf{c}' \in A)$ in (32) is strictly positive, indicating that the chain, when starting from $\mathbf{c}$, has a positive chance to move. Therefore, for any $A \in \mathscr{B}(\mathscr{C})$ such that $\hat{\rho}_t(A) > 0$, we can prove with the Chapman-Kolmogorov equation that there exists some $m \in \mathbb{N}^*$ such that

$$H^m(\mathbf{c},A) > 0,$$

where $H^m(\mathbf{c},A) = \int H^{m-1}(y,A)H(\mathbf{c},\mathrm{d}y)$ is the m-step transition kernel. In other words, the chain is $\hat{\rho}_t$-irreducible. Finally, a sufficient condition for the chain to be aperiodic is that Algorithm 3 allows transitions such as $\left\{\left(k^{(n+1)},\mathbf{c}^{(n+1)}\right) = \left(k^{(n)},\mathbf{c}^{(n)}\right)\right\}$, *i.e.*,

$$\mathbb{P}\left(\alpha\left[(k^{(n)},\mathbf{c}^{(n)}),(k',\mathbf{c}')\right] < 1\right) = \mathbb{P}\left(\frac{\hat{\rho}_t(\mathbf{c}')q(k',k^{(n)})\rho_{k^{(n)}}(\mathbf{c}^{(n)},\mathfrak{c}_{k^{(n)}},\tau_{k^{(n)}})}{\hat{\rho}_t(\mathbf{c}^{(n)})q(k^{(n)},k')\rho_{k'}(\mathbf{c}',\mathfrak{c}_{k'},\tau_{k'})} < 1\right) > 0. \quad (33)$$

Since for any $\mathbf{c}' \in A \subset \mathscr{B}\left(\mathbb{R}^{dk'}\right) \cap \mathscr{E}^c$ such that $\mathbb{P}\left(\mathbf{c}' \in A\right) = \int_A \rho_{k'}(\mathbf{c}',\mathfrak{c}_{k'},\tau_{k'})\mathrm{d}\mathbf{c}' > 0$, we have $\hat{\rho}_t(\mathbf{c}') = 0$, (33) holds. Therefore,

$$\mathbb{P}\left(\frac{\hat{\rho}_t(\mathbf{c}')q(k',k^{(n)})\rho_{k^{(n)}}(\mathbf{c}^{(n)},\mathfrak{c}_{k^{(n)}},\tau_{k^{(n)}})}{\hat{\rho}_t(\mathbf{c}^{(n)})q(k^{(n)},k')\rho_{k'}(\mathbf{c}',\mathfrak{c}_{k'},\tau_{k'})} < 1\right) \geq \mathbb{P}\left(\mathbf{c}' \in A\right) > 0.$$

The chain is therefore aperiodic. Finally, the Harris recurrence of the chain is a consequence of Lemma 1 (based on Roberts and Rosenthal, 2006, Theorem 20). As a conclusion, the chain converges to the target distribution $\hat{\rho}_t$.

Acknowledgements

The authors gratefully acknowledge financial support from iAdvize and ANRT (CIFRE grant 2014-00757).

References

P. Alquier. *Transductive and Inductive Adaptive Inference for Regression and Density Estimation*. PhD thesis, Université Paris 6, 2006.

P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14:243–280, 2013.

P. Alquier and B. Guedj. An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization. *Mathematical Methods of Statistics*, 2017.

P. Alquier and K. Lounici. PAC-Bayesian theorems for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

J.-Y. Audibert. *Une approche PAC-bayésienne de la théorie statistique de l'apprentissage*. PhD thesis, Université Paris 6, 2004.

J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3): 211–246, 2001.

W. Barbakh and C. Fyfe. Online clustering algorithms. *International Journal of Neural Systems*, 18(3):185–194, 2008.

P. L. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44(5): 1802–1813, 1998.

J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

R. B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. École d'Été de Probabilités de Saint-Flour 2001. Springer, 2004.

O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture notes – Monograph Series*. Institute of Mathematical Statistics, 2007.

N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *Journal of Computer and System Sciences*, 59(3):392–411, 1999.

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

N. Cesa-Bianchi, P. M. Long, and M. K. Warmuth. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7(3):604–619, 1996.

N. Cesa-Bianchi, D. Helmbold, N. Freund, Y. Haussler, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007. ISSN 1573-0565. . URL https://doi.org/10.1007/s10994-006-5001-7.

A. Choromanska and C. Monteleoni. Online clustering with experts. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 227–235, 2012.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory (COLT 2007), Lecture Notes in Computer Science*, pages 97–111, 2007.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78(5): 1423–1443, 2012.

P. Dellaportas, J. J. Forster, and I. Ntzoufras. On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36, 2002.

A. Fischer. On the number of groups in clustering. *Statistics and Probability Letters*, 81:1771–1781, 2011.

S. Gerchinovitz. *Prédiction de suites individuelles et cadre statistique classique : étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation*. PhD thesis, Université Paris-Sud, 2011.

A. D. Gordon. *Classification*, volume 82 of *Monographs on Statistics and Applied Probability*. Chapman Hall/CRC, Boca Raton, 1999.

P. J. Green. Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7:264–291, 2013.

B. Guedj and S. Robbiano. PAC-Bayesian high dimensional bipartite ranking. *Journal of Statistical Planning and Inference*, 2017.

S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: theory and practice. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):511–528, 2003.

J. A. Hartigan. *Clustering Algorithms*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, 1975.

L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience, Hoboken, 1990.

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Computational Learning Theory: 4th European Conference (EuroCOLT '99)*, pages 153–167. Springer, 1999.

A. N. Kolmogorov and V. M. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *American Mathematical Society Translations*, 17:277–364, 1961.

W. J. Krzanowski and Y. T. Lai. A criterion for determination the number of clusters in a data set. *Biometrics*, 44:23–34, 1988.

L. Li. *PACBO: PAC-Bayesian Online Clustering*, 2016. URL https://CRAN.R-project.org/package=PACBO. R package version 0.1.0.

E. Liberty, R. Sriharsha, and M. Sviridenko. An algorithm for online k-means clustering. In *Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 81–89. SIAM, 2016.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–216, 1994.

D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999a.

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999b.

G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179, 1985.

A. Petralias and P. Dellaportas. An MCMC model search algorithm for regression problems. *Journal of Statistical Computation and Simulation*, 83(9):1722–1740, 2013.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2004.

G. O. Roberts and J. S. Rosenthal. Harris Recurrence of Metropolis-Within-Gibbs and Trans-Dimensional Markov Chains. *Annals of Applied Probability*, 16(4):2123–2139, 2006.

M. Seeger. PAC-Bayesian generalization bounds for gaussian processes. *Journal of Machine Learning Research*, 3:233–269, 2002.

M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2003.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9. ACM, 1997. .

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society*, 63:411–423, 2001.

V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

O. Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017.

Appendix A: Extension to a different prior

For the sake of completion, this appendix presents additional regret bounds for a different heavy-tailed prior. Doing so, we stress that the quasi-Bayesian approach is flexible in the sense that it allows for regret bounds for a large variety of priors.

Let us consider π_k as a product of k independent truncated multivariate Student distributions with 3 degrees of freedom in $\mathbb{R}^d$, namely, for any $\mathbf{c} \in \mathbb{R}^{dk} \subset \mathscr{C}$,

$$\mathrm{d}\pi_k(\mathbf{c}, \tau_0, 2R) = \prod_{j=1}^{k} \left\{ C_{2R,\tau_0}^{-1} \left(1 + \frac{|c_j|_2^2}{6\tau_0^2}\right)^{-\frac{3+d}{2}} \mathbb{1}_{\{|c_j|_2 \leq 2R\}} \right\} \mathrm{d}\mathbf{c}, \tag{34}$$

where $\tau_0 > 0$ and $R > 0$ are respectively the scale and truncation parameters, and C_{2R,τ_0} is the normalizing constant accounting for the truncation. When $R = +\infty$, $\pi_k(\mathbf{c}, \tau_0, 2R)$ amounts to a distribution without truncation. In the following, we shorten $\pi_k(\mathbf{c}, \tau_0, 2R)$ to π_k whenever no confusion is possible.

Denote by ν the multivariate Student distribution in $\mathbb{R}^d$, with mean vector $0 \in \mathbb{R}^d$, scale parameter 1, and 3 degrees of freedom. Fix $k \in [\![1, p]\!]$, $R > 0$ and $\mathfrak{c} \in \mathscr{C}(k, R)$, and recall that $\Xi(k, R)$ denotes the hypercube in $\mathbb{R}^k$ defined by

$$\Xi(k, R) := \left\{ \xi = (\xi_j)_{j=1,\ldots,k} \in \mathbb{R}^k : 0 < \xi_j \leq R, \forall j \right\}.$$

For any $k \in [\![1, p]\!]$, $\mathbf{c} \in \mathbb{R}^{dk} \subset \mathscr{C}$, $\mathfrak{c} \in \mathscr{C}(k, R)$, $\xi \in \Xi(k, R)$, $0 < \tau^2 \leq \sqrt{3}R^2/(6\sqrt{d})$ and $R > 0$, we define the probability distribution ρ_k on $\mathbb{R}^{dk}$ by

$$\rho_k(\mathbf{c}, \mathfrak{c}, \tau, \xi) = \prod_{j=1}^{k} \left\{ C_{\xi_j,\tau}^{-1} \left(1 + \frac{|c_j - \mathfrak{c}_j|_2^2}{6\tau^2}\right)^{-\frac{3+d}{2}} \mathbb{1}_{\{|c_j - \mathfrak{c}_j|_2 \leq \xi_j\}} \right\}, \tag{35}$$

where $C_{\xi_j,\tau}$ are normalizing constants defined as $C_{\xi_j,\tau} = \mathbb{P}\left(|\nu|_2 \leq \xi_j/\sqrt{2}\tau\right)/A_{d,\tau}$, where $A_{d,\tau}$ is the constant in the density of ν. Moreover, when $(\xi_j)_{j=1,\ldots,k} = +\infty$, we let $\rho_k(\mathbf{c}, \mathfrak{c}, \tau, \xi)$ denote the multivariate Student distribution without truncation. In the sequel, we will shorten $\rho_k(\mathbf{c}, \mathfrak{c}, \tau, \xi)$ to ρ_k whenever no confusion is possible.

Lemma 4. *Assume that q and π_k in (4) are defined respectively as in (6) and (34), and that ρ_k is defined as (35) for each $k \in [\![1, p]\!]$. For the probability distribution $\rho(\mathbf{c}, \mathfrak{c}, \tau, \xi) = \mathbb{1}_{\{\mathbf{c} \in \mathbb{R}^{dk}\}} \rho_k(\mathbf{c}, \mathfrak{c}, \tau, \xi)$ defined on $\mathscr{C}$, if $R \geq \max_{t=1,\ldots,T} |x_t|_2$, then*

$$\begin{aligned}\mathscr{K}(\rho, \pi) \leq \sum_{j=1}^{k} \left[\frac{3+d}{2} \log\left(1 + \frac{\xi_j^2}{6\tau^2}\right) - \frac{d}{2} \log \xi_j^2 \right] - k \log c_d \\ + (3+d)k \log\left(1 + \frac{\tau}{\tau_0} + \frac{\sum_{j=1}^{k} |\mathfrak{c}_j|_2}{\sqrt{6}k\tau_0}\right) + kd \log \tau_0 + \log p + \eta(k-1).\end{aligned}$$

Proof. By the definition of the Kullback-Leibler divergence, we have

$$\mathscr{K}(\rho, \pi) = \mathscr{K}(\rho_k, \pi_k) + \log \frac{1}{q(k)} =: A + B, \tag{36}$$

where

$$
\begin{aligned}
A &= \int_{\mathbb{R}^{dk}} \log\left[\prod_{j=1}^{k} \frac{C_{2R,\tau_0}}{C_{\xi_j,\tau}} \left(\frac{\tau_0^2}{\tau^2}\frac{6\tau^2+|c_j-\mathfrak{c}_j|_2^2}{6\tau_0^2+|c_j|_2^2}\right)^{-\frac{3+d}{2}}\right]\rho_k(\mathbf{c})\mathrm{d}\mathbf{c} \\
&= \sum_{j=1}^{k}\log\frac{C_{2R,\tau_0}}{C_{\xi_j,\tau}} + \frac{3+d}{2}\int_{\mathbb{R}^{dk}}\sum_{j=1}^{k}\log\left(\frac{\tau^2}{\tau_0^2}\frac{6\tau_0^2+|c_j|_2^2}{6\tau^2+|c_j-\mathfrak{c}_j|_2^2}\right)\rho_k(\mathbf{c})\mathrm{d}\mathbf{c} \\
&= \sum_{j=1}^{k}\log\frac{\mathbb{P}\left(|\nu|_2 \le \frac{2R}{\sqrt{2}\tau_0}\right)}{\mathbb{P}\left(|\nu)|_2 \le \frac{\xi_j}{\sqrt{2}\tau}\right)} + kd\log\frac{\tau_0}{\tau} + \frac{3+d}{2}\int_{\mathbb{R}^{dk}}\sum_{j=1}^{k}\log\left(\frac{\tau^2}{\tau_0^2}\frac{6\tau_0^2+|c_j|_2^2}{6\tau^2+|c_j-\mathfrak{c}_j|_2^2}\right)\rho_k(\mathbf{c})\mathrm{d}\mathbf{c} \\
&=: A_1 + A_2 + A_3. \qquad (37)
\end{aligned}
$$

By the definition of the multivariate Student distribution ν,

$$
\begin{aligned}
\mathbb{P}\left(|\nu|_2 \le \frac{\xi_j}{\sqrt{2}\tau}\right) &= \int_{|\nu|_2 \le \frac{\xi_j}{\sqrt{2}\tau}} \frac{\Gamma(\frac{3+d}{2})}{\Gamma(\frac{3}{2})(3\pi)^{\frac{d}{2}}}\left(1+\frac{|\nu|_2^2}{3}\right)^{-\frac{3+d}{2}} \mathrm{d}\nu \\
&\ge \left(1+\frac{\xi_j^2}{6\tau^2}\right)^{-\frac{3+d}{2}} \frac{\Gamma(\frac{3+d}{2})}{\Gamma(\frac{3}{2})(3\pi)^{\frac{d}{2}}}\int_{|\nu|_2 \le \frac{\xi_j}{\sqrt{2}\tau}} \mathrm{d}\nu \\
&= c_d\tau^{-d}\left(1+\frac{\xi_j^2}{6\tau^2}\right)^{-\frac{3+d}{2}} \xi_j^d,
\end{aligned}
$$

where $\Gamma(\cdot)$ is the Gamma function and $c_d = \frac{\Gamma\left(\frac{3+d}{2}\right)}{\Gamma(\frac{3}{2})\Gamma\left(\frac{d}{2}+1\right)6^{\frac{d}{2}}}$. Hence, the term A_1 in (37) verifies

$$
\begin{aligned}
A_1 &= k\log\mathbb{P}\left(|\nu|_2 \le \frac{2R}{\sqrt{2}\tau_0}\right) - \sum_{j=1}^{k}\log\mathbb{P}\left(|\nu)|_2 \le \frac{\xi_j}{\sqrt{2}\tau}\right) \\
&\le -\sum_{j=1}^{k}\log\mathbb{P}\left(|\nu|_2 \le \frac{\xi_j}{\sqrt{2}\tau}\right) \\
&\le \sum_{j=1}^{k}\left[\frac{3+d}{2}\log\left(1+\frac{\xi_j^2}{6\tau^2}\right) - \frac{d}{2}\log\xi_j^2\right] + kd\log\tau - k\log c_d. \qquad (38)
\end{aligned}
$$

In addition, we have

$$
\begin{aligned}
\frac{6\tau_0^2+|c_j|_2^2}{6\tau^2+|c_j-\mathfrak{c}_j|_2^2} &\le 1 + \frac{2|\mathfrak{c}_j|_2}{2\sqrt{6}\tau}\frac{2\sqrt{6}\tau|c_j-\mathfrak{c}_j|_2}{6\tau^2+|c_j-\mathfrak{c}_j|_2^2} + \frac{|\mathfrak{c}_j|_2^2}{6\tau^2+|c_j-\mathfrak{c}_j|_2^2} + \frac{\tau_0^2}{\tau^2} \\
&= 1 + \frac{|\mathfrak{c}_j|_2}{\sqrt{6}\tau} + \frac{|\mathfrak{c}_j|_2^2}{6\tau^2} + \frac{\tau_0^2}{\tau^2} \le \left(1+\frac{|\mathfrak{c}_j|_2}{\sqrt{6}\tau}+\frac{\tau_0}{\tau}\right)^2,
\end{aligned}
$$

where we used the Cauchy–Schwarz inequality. Due to the above inequality, the term A_3 in (37) satisfies

$$\begin{aligned} A_3 &\le (3+d) \int \sum_{j=1}^{k} \log\left(1+\frac{\tau}{\tau_0}+\frac{|\mathfrak{c}_j|_2}{\sqrt{6}\tau_0}\right) \rho_k(\mathbf{c}) \mathrm{d}\mathbf{c} \\ &\le (3+d)k \int \log\left(1+\frac{\tau}{\tau_0}+\frac{\sum_{j=1}^{k}|\mathfrak{c}_j|_2}{\sqrt{6}k\tau_0}\right) \rho_k(\mathbf{c}) \mathrm{d}\mathbf{c} \\ &= (3+d)k \log\left(1+\frac{\tau}{\tau_0}+\frac{\sum_{j=1}^{k}|\mathfrak{c}_j|_2}{\sqrt{6}k\tau_0}\right). \end{aligned} \tag{39}$$

Combining (36), (37), (38), (39) with (18) completes the proof. □

Corollary 5. *For any sequence* $(x_t)_{1:T} \in \mathbb{R}^{dT}$*, for any* $\lambda > 0$*, if* q *and* π_k *in* (4) *are taken respectively as in* (6) *and* (34) *with parameter* $\eta \ge 0$*,* $\tau_0 > 0$ *and* $R \ge \max_{t=1,\dots,T} |x_t|_2$*, Algorithm 1 satisfies, for any* $0 < \tau^2 \le (\sqrt{3}R^2)/(6\sqrt{d})$*,*

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t, x_t) \le \inf_{k\in[\![1,p]\!]} \inf_{\mathbf{c}\in\mathscr{C}(k,R)} &\left\{ \sum_{t=1}^{T} \ell(\mathbf{c},x_t) + \frac{kd}{\lambda}\log\frac{\tau_0}{c_d\tau} + \frac{\eta}{\lambda}k \right. \\ &\left. + \frac{(3+d)k}{\lambda} \log\left(1+\frac{\tau}{\tau_0}+\frac{\sum_{j=1}^{k}|c_j|_2}{\sqrt{6}k\tau_0}\right) + \frac{1}{\lambda}\sqrt{kd(12\tau^2T\lambda+3k)} \right\} + \frac{\lambda T}{2}C_1^2 + \frac{\log p}{\lambda}, \end{aligned}$$

where $C_1 = (2R + \max_{t=1,\dots,T}|x_t|_2)^2$ *and* $c_d = \left(\frac{\Gamma(\frac{3+d}{2})}{\Gamma(\frac{3}{2})\Gamma(\frac{d}{2}+1)}\right)^{1/d}$.

Proof. By Proposition 1,

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t, x_t) \le \inf_{k\in[\![1,p]\!]} \inf_{\substack{\rho\in\mathscr{P}_\pi(\mathscr{C}) \\ \rho=\rho_k \mathbb{1}_{\{\mathbf{c}\in\mathbb{R}^{dk}\}}}} &\left\{ \mathbb{E}_{\mathbf{c}\sim\rho} \sum_{t=1}^{T} [\ell(\mathbf{c},x_t)] + \frac{\mathscr{K}(\rho,\pi)}{\lambda} \right. \\ &\left. + \frac{\lambda}{2} \mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_T)} \mathbb{E}_{\mathbf{c}\sim\rho} \sum_{t=1}^{T} [\ell(\mathbf{c},x_t) - \ell(\hat{\mathbf{c}}_t,x_t)]^2 \right\} \end{aligned} \tag{40}$$

As in (17), the first term on the right-hand side of (40) may be upper bounded.

$$\sum_{t=1}^{T} \mathbb{E}_{\mathbf{c}\sim\rho}[\ell(\mathbf{c},x_t)] \le \sum_{t=1}^{T} \ell(m,x_t) + T \max_{j=1,\dots,k} \xi_j^2. \tag{41}$$

For the second term in the right-hand side of (40), by Lemma 4,

$$\begin{aligned} \frac{\mathscr{K}(\rho,\pi)}{\lambda} \le \frac{(3+d)k}{\lambda} \log\left(1+\frac{\tau}{\tau_0}+\frac{\sum_{j=1}^{k}|\mathfrak{c}_j|_2}{\sqrt{6}k\tau_0}\right) + \frac{1}{\lambda}\sum_{j=1}^{k}\left[\frac{3+d}{2}\log\left(1+\frac{\xi_j^2}{6\tau^2}\right) - \frac{d}{2}\log\xi_j^2\right] \\ + \frac{kd}{\lambda}\log\tau_0 - \frac{k}{\lambda}\log c_d + \frac{\eta}{\lambda}(k-1) + \frac{\log p}{\lambda}. \end{aligned} \tag{42}$$

Likewise to (20), the third term on the right-hand side of (40) is upper bounded by

$$\frac{\lambda}{2}\mathbb{E}_{(\hat\rho_1,\ldots,\hat\rho_T)}\mathbb{E}_{\mathbf{c}\sim\rho_k}\sum_{t=1}^{T}[\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t)]^2 \le \frac{\lambda T}{2}C_1^2. \tag{43}$$

Combining inequalities (41), (42) and (43) yields for $\xi\in\Xi(k,R)$ and $0<\tau^2\le\sqrt{3}R^2/(6\sqrt{d})$ that

$$\sum_{t=1}^{T}\mathbb{E}_{(\hat\rho_1,\hat\rho_2,\ldots,\hat\rho_t)}\ell(\hat{\mathbf{c}}_t,x_t)\le\inf_{k\in[\![1,p]\!]}\inf_{\mathfrak{c}\in\mathscr{C}(k,R)}\left\{\sum_{t=1}^{T}\ell(\mathfrak{c},x_t)+\xi_j^2+\frac{(3+d)k}{\lambda}\log\left(1+\frac{\tau}{\tau_0}+\frac{\sum_{j=1}^k|\mathfrak{c}_j|_2}{\sqrt{6}k\tau_0}\right)\right.$$
$$\left.+T\max_{j=1,\ldots,k}\xi_j^2+\frac{3+d}{2\lambda}\sum_{j=1}^{k}\log\left(1+\frac{\xi_j^2}{6\tau^2}\right)-\frac{d}{2\lambda}\sum_{j=1}^{k}\log\xi_j^2+\frac{kd}{\lambda}\log\tau_0-\frac{k}{\lambda}\log c_d+(k-1)\right\}$$
$$+\frac{\lambda T}{2}C_1^2+\frac{\log p}{\lambda}.$$

Let $\hat\xi_j=\xi_j^2/6\tau^2$ for any $j=1,\ldots,k$, then $0<\hat\xi_j\le R^2/6\tau^2$ since $\xi=(\xi_j)_{j=1,\ldots,k}\in\Xi(k,R)$. This yields

$$\begin{aligned}
&T\max_{j=1,\ldots,k}\xi_j^2+\frac{3+d}{2\lambda}\sum_{j=1}^{k}\log\left(1+\frac{\xi_j^2}{6\tau^2}\right)-\frac{d}{2\lambda}\sum_{j=1}^{k}\log\xi_j^2\\
&=6\tau^2T\max_{j=1,\ldots,k}\hat\xi_j+\frac{3}{2\lambda}\sum_{j=1}^{k}\log\left(1+\hat\xi_j\right)+\frac{d}{2\lambda}\sum_{j=1}^{k}\log\left(1+\frac{1}{\hat\xi_j}\right)-\frac{kd}{2\lambda}\log(6\tau^2)\\
&\le 6\tau^2T\max_{j=1,\ldots,k}\hat\xi_j+\frac{3}{2\lambda}\sum_{j}^{k}\hat\xi_j+\frac{d}{2\lambda}\sum_{j=1}^{k}\frac{1}{\hat\xi_j}-\frac{kd}{2\lambda}\log(6\tau^2)\\
&\le\left(6\tau^2T+\frac{3k}{2\lambda}\right)\max_{j=1,\ldots,k}\hat\xi_j+\frac{d}{2\lambda}\sum_{j=1}^{k}\frac{1}{\hat\xi_j}-\frac{kd}{2\lambda}\log(6\tau^2).
\end{aligned}\tag{44}$$

The minimum of the right-hand side of (44) is reached for

$$\hat\xi_1=\cdots=\hat\xi_k=\sqrt{\frac{kd}{12\tau^2T\lambda+3k}}\le\frac{R^2}{6\tau^2},\quad\text{if }0<\tau^2\le\frac{\sqrt{3}R^2}{6\sqrt{d}}.$$

Therefore for a fixed k, $\mathfrak{c}\in\mathscr{C}(k,R)$ and $0<\tau^2\le\frac{\sqrt{3}R^2}{6\sqrt{d}}$,

$$\inf_{\xi\in\Xi(k,R)}\left\{T\max_{j=1,\ldots,k}\xi_j^2+\frac{3+d}{2\lambda}\sum_{j=1}^{k}\log\left(1+\frac{\xi_j^2}{6\tau^2}\right)-\frac{d}{2\lambda}\sum_{j=1}^{k}\log\xi_j^2\right\}\le\frac{1}{\lambda}\sqrt{kd(12\tau^2T\lambda+3k)}$$
$$-\frac{kd}{2\lambda}\log6\tau^2.$$

Hence

$$\sum_{t=1}^{T}\mathbb{E}_{(\hat\rho_1,\hat\rho_2,\ldots,\hat\rho_t)}\ell(\hat{\mathbf{c}}_t,x_t)\le\inf_{k\in[\![1,p]\!]}\inf_{\mathfrak{c}\in\mathscr{C}(k,R)}\left\{\sum_{t=1}^{T}\ell(\mathfrak{c},x_t)+\frac{(3+d)k}{\lambda}\log\left(1+\frac{\tau}{\tau_0}+\frac{\sum_{j=1}^k|\mathfrak{c}_j|_2}{\sqrt{6}k\tau_0}\right)\right.$$

$$+\frac{1}{\lambda}\sqrt{kd(12\tau^2T\lambda+3k)}+\frac{kd}{\lambda}\log\frac{\tau_0}{\sqrt{6}\tau c_d^{1/d}}+\frac{\eta}{\lambda}(k-1)\Bigg\}+\frac{\lambda T}{2}C_1^2+\frac{\log p}{\lambda}.$$

which concludes the proof. □

Tuning parameters λ, τ and η can be chosen to obtain a sublinear regret bound for the cumulative loss of Algorithm 1.

Corollary 6. *For any sequence $(x_t)_{1:T}\in\mathbb{R}^{dT}$, under the assumptions of Corollary 5, if $T\geq 12d\tau_0^4/c_d^2R^4$, $\lambda=\sqrt{\log T}/\sqrt{T}$, $\tau^2=\tau_0^2T^{-1/2}(c_d)^{-2}$ and $\eta\geq 0$, Algorithm 1 satisfies*

$$\sum_{t=1}^T\mathbb{E}_{(\hat\rho_1,\hat\rho_2,\ldots,\hat\rho_t)}\ell(\hat{\mathbf{c}}_t,x_t)\leq\inf_{k\in[\![1,p]\!]}\inf_{\mathbf{c}\in\mathscr{C}(k,R)}\Bigg\{\sum_{t=1}^T\ell(\mathbf{c},x_t)+(3+d)k\sqrt{T}\log\left(1+\frac{1}{c_dT^{\frac14}}+\frac{\sum_{j=1}^k|c_j|_2}{\sqrt{6}k\tau_0}\right)$$
$$+\frac{kd}{4}\sqrt{T\log T}+\left(\sqrt{3k^2d+12\tau_0^2(c_d)^{-2}}+\eta k\right)\sqrt{T}\Bigg\}+\left(\log p+\frac{C_1^2}{2}\right)\sqrt{T},$$

where $C_1=(2R+\max_{t=1,\ldots,T}|x_t|_2)^2$ and $c_d=\left(\frac{\Gamma(\frac{3+d}{2})}{\Gamma(\frac32)\Gamma(\frac d2+1)}\right)^{1/d}$.

In the adaptive setting (Algorithm 2), applying Theorem 1 to the specific q and π_k in (6) and (34) leads to the following result.

Corollary 7. *For any deterministic sequence $(x_t)_{1:T}\in\mathbb{R}^{dT}$, under the assumptions of Corollary 5, set $T\geq 12d\tau_0^4/c_d^2R^4$, $\eta\geq 0$, $R\geq\max_{t=1,\ldots,T}|x_t|_2$ and $\lambda_t=\sqrt{\log t}/\sqrt{t}$ for any $t\in[\![1,T]\!]$ and $\lambda_0=1$. Then Algorithm 2 satisfies*

$$\sum_{t=1}^T\mathbb{E}_{(\hat\rho_1,\hat\rho_2,\ldots,\hat\rho_t)}\ell(\hat{\mathbf{c}}_t,x_t)\leq\inf_{k\in[\![1,p]\!]}\inf_{\mathbf{c}\in\mathscr{C}(k,R)}\Bigg\{\sum_{t=1}^T\ell(\mathbf{c},x_t)+(3+d)k\sqrt{T}\log\left(1+\frac{1}{c_dT^{\frac14}}+\frac{\sum_{j=1}^k|c_j|_2}{\sqrt{6}k\tau_0}\right)$$
$$+\frac{kd}{4}\sqrt{T\log T}+\left(\sqrt{3k^2d+12\tau_0^2(c_d)^{-2}}+\eta k\right)\sqrt{T}\Bigg\}+\left(\log p+C_1^2\right)\sqrt{T},$$

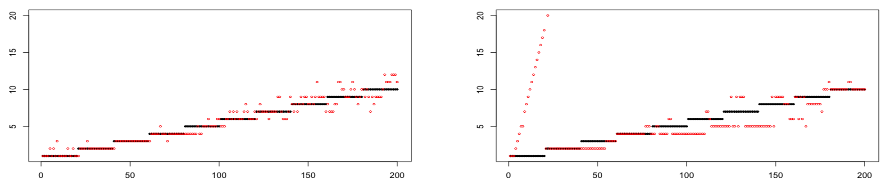
where $C_1=(2R+\max_{t=1,\ldots,T}|x_t|_2)^2$ and $c_d=\left(\frac{\Gamma(\frac{3+d}{2})}{\Gamma(\frac32)\Gamma(\frac d2+1)}\right)^{1/d}$.

Proof. The proof is similar to the proof of Corollary 5, the only difference lies in the fact that (43) is replaced by

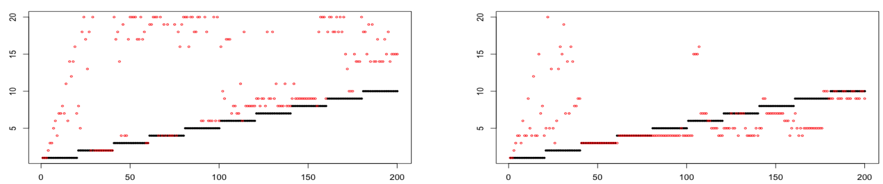
$$\mathbb{E}_{(\hat\rho_1,\ldots,\hat\rho_T)}\mathbb{E}_{\mathbf{c}\sim\rho_k}\sum_{t=1}^T\frac{\lambda_{t-1}}{2}[\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t)]^2\leq C_1^2\sqrt{T\log T}.$$

□

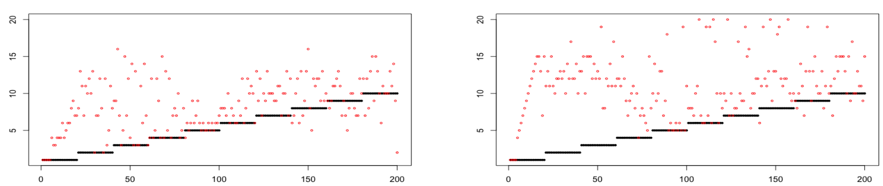
For the sake of completion, we present in Figure 6 the performance of PACBO and its seven competitors for estimating the true number $k_t^\star$ of clusters along time. We acknowledge that no theoretical guarantee is derived for the estimation of $k_t^\star$ yet the practical behavior is remarkable.



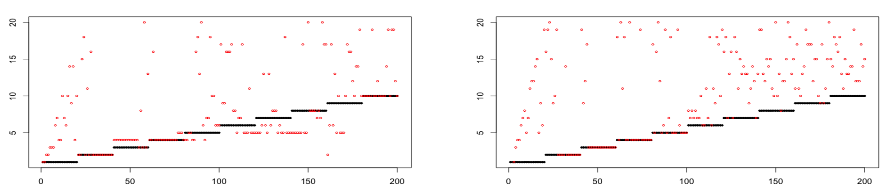
(a) PACBO (left) and Silhouette (right)



(b) Calinski (left) and Hartigan (right)



(c) Djump (left) and DDSE (right)



(d) Lai (left) and Gap (right)

Figure 6: True (black) and estimated (red) number of clusters as functions of t.