Online bi-clustering with sparsity priors

Sébastien Loustau

Online bi-clustering with sparsity priors

Sébastien Loustau
loustau@math.univ-angers.fr
Université d'Angers

Abstract

Given a deterministic matrix $M = (m_{ij})$, we want to cluster the row i and the column j before to predict m_{ij}, where the values of the matrix are observed sequentially. The proposed algorithms are PAC-Bayesian procedures with new sparsity priors. Sparsity regret bounds are stated without any assumption on the matrix. These results are based on [13] where online clustering algorithms are suggested. Eventually, we also state minimax lower bounds for these problems, using a classical probabilistic reduction scheme. It shows the minimax optimality of the proposed algorithms in a worst case scenario.

1 Introduction

Bi-clustering or co-clustering is a popular method to analysis data matrices and build recommender systems. In this problem, we mainly observe a random matrix, where rows correspond to a population and columns to variables (or products). This matrix is usually sparse, i.e. with many hidden entries (called ratings). The goal is to reconstruct the matrix by clustering simultaneously the rows and colums of the matrix. This scenario has been applied to many real-world problems such as text mining (see [19]), gene expression ([6]), social networks (see [9]) or collaborative filtering (see [17]). In [18], generalization bounds in terms of KL divergence are proposed for this problem. Assuming the existence of a probabilistic distribution $p(x_1, x_2, y)$ over the triplet of rows, colums and rating, a discriminative predictor $q(y|x_1, x_2)$ is constructed via a PAC-Bayesian approach. The matrix is supposed to have i.i.d. entries and the number of clusters is known in advance. In this paper, we want to investigate a more challenging game in a worst case scenario without the knownledge of the number of clusters.

To tackle this problem, we suggest a high dimensional PAC-Bayesian approach with sparsity priors. Sparsity priors have been introduced in Bayesian estimation by several authors ([11],[15],[16]). The principle is very often to employ heavy-tailed distributions, such as multivariate Laplace, quasi-Cauchy or Pareto priors. In a PAC-Bayesian framework, [7] introduced sparsity priors in mirror averaging procedure to promote sparsity oracle inequalities. More recently, sparsity priors have been used in online learning (see [8], [13]). In [13], we promote sequential clustering algorithms with sparsity priors in the problem of *online clustering* of an individual sequence $(x_t)_{t=1}^T \in \mathbb{R}^{dT}$. On each day t, the forecaster must predict the next instance $x_t \in \mathbb{R}^d$ with at most $p \geq 1$ possible "proposals" or "strategies". On the morning of day t, he has access to the inputs $x_1, \dots, x_{t-1}$ of the previous days. Based on these instances, he must propose a codebook of $p \geq 1$ strategies $\hat{\mathbf{c}}_t = (\hat{c}_{t,1}, \dots, \hat{c}_{t,p}) \in \mathbb{R}^{dp}$. At the end of the day, he receives x_t and incurs a loss - or distortion - $\ell(\hat{\mathbf{c}}_t, x_t)$, where:

$$\ell(\hat{\mathbf{c}}_t, x_t) = \min_{j=1,\dots,p} |\hat{c}_{t,j} - x_t|_2^2,$$

and $|\cdot|_2$ stands for the Euclidean norm in $\mathbb{R}^d$. The goal of the forecaster is to control the cumulative distortion $\sum_{t=1}^T \ell(\hat{\mathbf{c}}_t, x_t)$, with $|\hat{\mathbf{c}}_t|_0$ as small as possible, where $|\hat{\mathbf{c}}_t|_0$ corresponds to the number of non-zero strategies at time t, i.e.:

$$|\mathbf{c}|_0 := \text{card}\{j = 1, \dots, p : c_j \neq (0, \dots, 0)^\top \in \mathbb{R}^d\}, \ \forall \mathbf{c} = (c_1, \dots, c_p) \in \mathbb{R}^{dp}. \tag{1.1}$$

In such a framework, we recommend to reach sparsity regret bounds as in [13] according to:

$$\sum_{t=1}^{T} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{\mathbf{c}\in\mathbb{R}^{dp}} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \lambda |\mathbf{c}|_0 \right\} + r_\lambda(T), \tag{1.2}$$

where $|\cdot|_0$ is defined in (1.1), $r_\lambda(T)$ is a residual term and $\lambda > 0$ is a temperature parameter. In other words, we control the regret of our sequential procedure to have not reached the compromise between fitting the data and compress the information (i.e. the infimum which appears in the right hand side). With a suitable calibration of λ, it gives the following result:

Theorem 1 ([13]) *For any deterministic sequence $(x_t)_{t=1}^T \in \mathbb{R}^{dT}$, any $R > 0$, there exists a sequential algorithm such that:*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{p}_1\ldots,\hat{p}_t)} \ell(\hat{\boldsymbol{c}}_t, x_t) \leq \inf_{\boldsymbol{c}\in\mathcal{B}^p(R)} \left\{ \sum_{t=1}^{T} \ell(\boldsymbol{c}, x_t) + |\boldsymbol{c}|_0 \sqrt{T(3+d)} \log\left(1 + \frac{\sqrt{T}\sum_{j=1}^{p} |c_j|_2}{\sqrt{6}|\boldsymbol{c}|_0}\right) \right\} + C\sqrt{T},$$

where for any $t = 1, \ldots, T$, $\hat{\boldsymbol{c}}_t$ is a randomized codebook with law $\hat{p}_t$.

This result proposes a sparsity regret bound with rates $\sqrt{T}\log T$. If we suppose the existence of a minimizer $\mathbf{c}^\star$ of the RHS of Theorem 1 such that $|\mathbf{c}^\star|_0 = s$ for some sparsity index $s \in \mathbb{N}^*$, we have, for T large enough:

$$\sum_{t=1}^{T} \mathbb{E}\ell(\hat{\mathbf{c}}_t, x_t) - \sum_{t=1}^{T} \ell(\mathbf{c}^\star, x_t) \leq \text{const.} \times s\sqrt{T}\log T.$$

In this paper, we are mainly interested in (1) an online bi-clustering scenario and (2) minimax results for both Theorem 1 and the bi-clustering scenario. Let us consider an individual sequence (x_t, y_t), $t = 1, \ldots, T$ where T is the known horizon whereas for any $t = 1, \ldots, T$:

- the input variable $x_t = (x_{t,1}, \ldots, x_{t,d}) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_d =: \mathcal{X}$,
- the output $y_t \in \mathcal{Y} \subseteq [0, M]$[1].

A seminal example is the construction of recommender systems. In this case, $d = 2$ and $x_t = (x_{t,1}, x_{t,2})$ corresponds to a couple customer×movie whereas y_t is the associated rating (such as $\{\star, \star\star, \star\star\star\}$ for instance). Note also that our analysis is not limited to the bi-clustering problem where $d = 2$ above, since we can consider high dimensional $d > 2$ tensors as well.

Giving the individual sequence (x_t, y_t), $t = 1, \ldots, T$, we want to construct a sequential algorithm as follows. At each time t, an input x_t is observed and we build a prediction $\hat{y}_t$. Then, y_t is given and we pay $(y_t - \hat{y}_t)^2$. This particular loss enjoys the useful property to be λ-exp-concave, which means that $\hat{y} \mapsto e^{-\lambda(\hat{y},y)^2}$ is concave. This property allows to reach fast regret bounds in the deterministic setting (see for instance [5]). The construction of the prediction is based on a mixture of expert's advices constructed thanks to online clustering. More precisely, the algorithm described in Section 2 gives a prediction of the following form:

$$\hat{y}_t = \mathbb{E}_{\vec{\mathbf{c}}\sim\hat{p}_t} g_{\vec{\mathbf{c}}}(x_t). \tag{1.3}$$

In (1.3), the function $g_{\vec{\mathbf{c}}}$ is constructed thanks to a set of d-tensor codebook $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_d) \in \prod_{j=1}^{d} \mathcal{X}_j^{p_j}$. This d-tensor codebook assigns to each component x_j of $x \in \mathcal{X}$ the nearest center of $\mathbf{c}_j$. The associated d-tensor Voronoï cell is denoted as $V_{\vec{\mathbf{c}}}(x)$ and corresponds to the product space

[1] In the sequel, two cases are considered: $\mathcal{Y} = [0, M]$ (online regression) and $\mathcal{Y} = \{1, \ldots, M\}$ (online classification).

of each Voronoï cell $V_{\mathbf{c}_j}(x_j) = \{y_j \in \mathcal{X}_j : \arg\min_{i_j=1,\dots,p_j} |y_j - c_{j,i_j}|_2 = \arg\min_{i_j=1,\dots,p_j} |x_j - c_{j,i_j}|_2\}$. In what follows (see for instance Theorem 2), we consider two different functions $\vec{\mathbf{c}} \mapsto g_{\vec{\mathbf{c}}}(\cdot)$. The first one consists in computing the mean value of the sequence of past outputs y_1, …, y_{t-1} in cell $V_{\vec{\mathbf{c}}}(x_t)$. In this case, $g_{\vec{\mathbf{c}}}(\cdot)$ is written as:

$$g_{\vec{\mathbf{c}}}^{\text{mean}}(x_t) = \frac{\sum_{u=1}^{t-1} y_u \mathbf{1}_{x_u \in V_{\vec{\mathbf{c}}}(x_t)}}{\text{card}\left\{\{x_1, \dots, x_{t-1}\} \cap V_{\vec{\mathbf{c}}}(x_t)\right\}}, \tag{1.4}$$

where in (1.4), we can take $g_{\vec{\mathbf{c}}}(x_1) = M/2$ without loss of generality. Indeed, in (1.4) (and in (1.5) also), $g_{\vec{\mathbf{c}}}(\cdot)$ depends on the past observations $(x_1, y_1), \dots, (x_{t-1}, y_{t-1})$ and then on time t. We omit this dependence for simplicity. Note that when $\mathcal{Y} = \{0, \dots, M\}$, we can also use the majority vote for $g_{\vec{\mathbf{c}}}(\cdot)$, where the majority vote at time t is taken in the Voronoï cell $V_{\vec{\mathbf{c}}}(x_t)$ as follows:

$$g_{\vec{\mathbf{c}}}^{\text{vote}}(x_t) = \arg\max_{k \in \mathcal{Y}} \text{card}\{u = 1, \dots, t-1 : y_u = k \text{ and } x_u \in V_{\vec{\mathbf{c}}}(x_t)\}. \tag{1.5}$$

Equipped with these base forecasters, we want to promote in (1.3) a sparse representation. Here, the sparsity is associated with the set of d-tensor codebooks. Given some vector of integers $\mathbf{m} = (m_1, \dots, m_p) \in \mathbb{N}^p$, we restrict the study to the Euclidean space by considering $\mathcal{X}_j = \mathbb{R}^{m_j}$ for any $j = 1, \dots, p$. Then, $\mathcal{X} = \mathbb{R}^{\sum_{j=1}^d m_j p_j}$ and $\vec{\mathbf{c}} = (\mathbf{c}_1, \dots, \mathbf{c}_d) \in \prod_{j=1}^d \mathbb{R}^{m_j p_j}$. As in [13], we wish that $y_t \approx g_{\vec{\mathbf{c}}^\star}(x_t)$, where $\vec{\mathbf{c}}^\star = (\mathbf{c}_1^\star, \dots, \mathbf{c}_d^\star)$ is such that $\mathbf{c}_j^\star$ has a small ℓ_0-norm for any $j = 1, \dots, d$ where $|\mathbf{c}_j^\star|_0$ is defined in (1.1). Consequently, we are looking for d distincts group-sparsity codebooks $\mathbf{c}_1, \dots, \mathbf{c}_d$. For this purpose, we will use in our algorithm a product of d group-sparsity priors introduced in [13] in online clustering. This prior is defined in Lemma 1 below.

In a classical statistical learning context, [18] considers a random generator P with unknown probability distribution on the set $\mathcal{X} \times \mathcal{Y}$ and suggest the following discriminative predictors:

$$h(y|x_1, \dots, x_d) = \sum_{i_1, \dots, i_d} h(y|i_1, \dots, i_d) \Pi_{j=1}^d h(i_j|x_j).$$

In this stochastic setting, the hidden variables $(i_1, \dots, i_d)$ represent the clustering of the input $X = (X_1, \dots, X_d)$. Using a PAC-Bayesian analysis and bounds as in [14], generalization errors in terms of Kullback divergence are proposed. The randomized strategy is based on a density estimation of the law P.

In this paper, the framework is essentially different since we propose to use PAC-Bayesian tools inspired from [13] to get sparsity regret bounds of the following form:

$$\sum_{t=1}^T \ell(y_t, \hat{y}_t) \leq \inf_{\vec{\mathbf{c}} \in \Pi_{j=1}^d \mathbb{R}^{m_j p_j}} \left\{ \sum_{t=1}^T (y_t - g_{\vec{\mathbf{c}}}(x_t))^2 + \text{pen}_0(\vec{\mathbf{c}}) \right\}, \tag{1.6}$$

where $\text{pen}_0(\vec{\mathbf{c}})$ is a penalty function which is proportional to the sum of the ℓ_0-norm of the codebooks $\mathbf{c}_1, \dots, \mathbf{c}_d$. The infimum in the RHS of (1.6) could be seen as a compromise between fitting the data and a sparse representation, where the sparsity is related with the number of clusters in the product space $\mathcal{X}$.

As an example, we can consider recommender systems where we want to predict the value of the rating of a new (customer×movie) couple. In this case, representation $g_{\vec{\mathbf{c}}}$ in (1.4) or (1.5) with a sparse $\vec{\mathbf{c}}$ in terms of ℓ_0-norm (1.1) means that we can propose a simple representation of ratings with a block matrix with a few number of blocks. In this case, it is well-motivated to perform online clustering before predicting the rating.

The rest of the paper is organized as follows. In Section 2, we give sparsity regret bounds for the problem of bi-clustering, reached by a sequential procedure with sparsity priors. Section 3 explores minimax optimality in online clustering and online bi-clustering. Section 4 concludes the paper whereas Section 5-6 are dedicated to the proofs of the main results.

2 General algorithm and sparsity regret bound

Before to describe the algorithm, let us introduce some notations. We denote by $\mathcal{C} := \Pi_{j=1}^{d} \mathbb{R}^{m_j p_j}$ the space of d-tensor codebooks, whereas a decision function at time t is denoted as $g_{\vec{\mathbf{c}}}(\cdot)$ (see (1.4) or (1.5)). We introduce a prior $\pi \in \mathcal{P}(\mathcal{C})$, where $\mathcal{P}(\mathcal{C})$ is the set of probability measure on $\mathcal{C}$, and a temperature parameter $\lambda > 0$. We can now describe the general algorithm and its associated PAC-Bayesian inequality.

2.1 The algorithm of bi-clustering

The principle of the algorithm is to predict y_t according to a mixture of decision functions $g_{\vec{\mathbf{c}}}$, where the mixture is updated by giving the best prediction of y_t at each iteration. At the beginning of the game, $\hat{p}_1 := \pi$. We observe x_1 and predict according to $\hat{y}_1 := \mathbb{E}_{\hat{p}_1} g_{\vec{\mathbf{c}}}(x_1)$, where $g_{\vec{\mathbf{c}}}(x_1)$ is defined above. Then, learning proceeds as the following sequence of trials $t = 1, \ldots, T-1$:

- Get y_t and compute:

$$\hat{p}_{t+1}(d\vec{\mathbf{c}}) = \frac{e^{-\lambda \sum_{u=1}^{t} (y_u - g_{\vec{\mathbf{c}}}(x_u))^2}}{W_t} d\pi(\vec{\mathbf{c}}), \tag{2.1}$$

  where $W_t := \mathbb{E}_{\pi} e^{-\lambda \sum_{u=1}^{t} (y_u - g_{\vec{\mathbf{c}}}(x_u))^2}$ is the normalizing constant.

- Get x_{t+1} and predict $\hat{y}_{t+1} := \mathbb{E}_{\vec{\mathbf{c}} \sim \hat{p}_{t+1}} g_{\vec{\mathbf{c}}}(x_{t+1})$.

Then, we have constructed a sequence of prediction $(\hat{y}_t)_{t=1,\ldots,T}$ which satisfies the following PAC-Bayesian bound.

Proposition 1 *For any deterministic sequence $(x_t, y_t)_{t=1}^{T}$, for any $p \in \mathbb{N}^d$, any $\lambda \leq 1/2M^2$ and any prior $\pi \in \mathcal{P}(\mathcal{C})$, the previous algorithm satisfies:*

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \leq \inf_{\rho \in \mathcal{P}(\mathcal{C})} \left\{ \mathbb{E}_{\vec{\mathbf{c}} \sim \rho} \sum_{t=1}^{T} (y_t - g_{\vec{\mathbf{c}}}(x_t))^2 + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\}, \tag{2.2}$$

where $\mathcal{K}(\rho, \pi)$ denotes the Kullback-Leibler divergence between ρ and the prior π and $g_{\vec{\mathbf{c}}}(\cdot)$ satisfies (1.4) *or* (1.5)*.*

The bound of Proposition 1 gives a control of the cumulative loss of the sequential procedure described above for any choice of prior π. It allows us in the sequel to choose a particular sparsity prior in order to state a sparsity regret bound of the form (1.6).

2.2 Sparsity regret bounds

The main motivation to introduce our prior is to promote sparsity in the following sense. In $g_{\vec{\mathbf{c}}}(\cdot)$, we want a codebook $\vec{\mathbf{c}} \in \mathbb{R}^{\sum_{j=1}^{d} m_j p_j}$ where $\vec{\mathbf{c}} = (\mathbf{c}_1, \ldots, \mathbf{c}_d)$ is such that $|\mathbf{c}_j|_0$ is small for any $j = 1, \ldots, d$, where:

$$|\mathbf{c}_j|_0 = \mathrm{card}\{i_j = 1, \ldots, p_j : c_{j,i_j} = (0, \ldots, 0) \in \mathbb{R}^{m_j}\}.$$

To deal with this issue, we propose a product of d group-sparsity priors introduced as follows:

$$d\pi_{S,d}(\vec{\mathbf{c}}) := \prod_{j=1}^{d} \prod_{i_j=1}^{p_j} \left\{ a_\tau \left(1 + \frac{|c_{j,i_j}|_2^2}{6\tau^2} \right)^{-\frac{3+m_j}{2}} \right\} d\vec{\mathbf{c}}, \tag{2.3}$$

for some constant $a_\tau > 0$. This prior consists of a product of d products of p_j multivariate Student's distribution $\sqrt{2}\tau\mathcal{T}_{m_j}(3)$, where $\tau > 0$ is a scaling parameter and $\mathcal{T}_{m_j}(3)$ is the m_j-multivariate Student with three degrees of freedom. It can be viewed as a generalization of the group-sparsity prior defined in [13] where $d = 1$. Consequently, we use the same multivariate Student's distribution presented in [12], defined as the ratio between a gaussian vector and the square root of an independent χ^2 distribution with 3 degrees of freedom.

It is important to stress that in (2.3), we don't need to threshold the prior at a given radius $R > 0$ such as in [13]. This is due to the presence of the square loss with bounded outputs $y \in \mathcal{Y} \subseteq [0, M]$. A straightforward application of Lemma 1 in [13] to the bi-clustering framework gives the following lemma:

Lemma 1 *Let $p \in \mathbb{N}^d$, $\tau > 0$. Consider the prior $\pi_{S,d}$ defined in (2.3). Let $\vec{\boldsymbol{c}} = (\boldsymbol{c}_1, \dots, \boldsymbol{c}_d)) \in \mathbb{R}^{\sum_{j=1}^d m_j p_j}$. Then, if we denote by $p_{0,d}$ the translated version of $\pi_{S,d}$ with mean $\vec{\boldsymbol{c}}$, we have:*

$$\begin{aligned}\mathcal{K}(p_{0,d}, \pi_{S,d}) &\leq \sum_{j=1}^d \left\{ (3+m_j) \sum_{i_j=1}^{p_j} \log\left(1 + \frac{|c_{j,i_j}|_2}{\sqrt{6}\tau}\right)\right\} \\ &\leq \sum_{j=1}^d \left\{ (3+m_j)|\boldsymbol{c}_j|_0 \log\left(1 + \frac{\sum_{i_j=1}^{p_j}|c_{j,i_j}|_2}{\sqrt{6}\tau|\boldsymbol{c}_j|_0}\right)\right\}.\end{aligned}$$

The first result is a direct consequence of Proposition 1 and the introduction of the sparsity prior (2.3).

Theorem 2 *For any deterministic sequence $(x_t, y_t)_{t=1}^T$, consider prior $\pi_{S,d}$ defined in (2.3) with $\tau = \delta\{\sqrt{24Mp_+}T\}^{-1}$ for some $\delta > 0$, $\lambda = 1/2M^2$ and functions $g_{\vec{\boldsymbol{c}}}(\cdot)$ satisfy (1.4) or (1.5). Then if T is great enough, we have:*

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\vec{\boldsymbol{c}} \in \mathcal{C}} \left\{ \sum_{t=1}^T (y_t - g_{\vec{\boldsymbol{c}}}(x_t))^2 + C \sum_{j=1}^d (3+m_j)|\boldsymbol{c}_j|_0 \log T \right\},$$

where $C > 0$ is a constant independent of T.

This result gives a sparsity regret bound where the penalty term is proportional to the sum of the ℓ_0-norm of the set of codebooks $\vec{\mathbf{c}} = (\mathbf{c}_1, \dots, \mathbf{c}_d)$. The algorithm performs as well as the best compromise between fitting the data and complexity. It is important to highlight that the RHS does not depend on the sequence $(p_1, \dots, p_d)$. Then, as in [13], we can consider large values of p_j and perform model selection clustering to learn the number of clusters.

Unfortunately, this result is essentially asymptotic since it holds for large values of T. This is due to the control of the deviation of the random variable $g_{\vec{\mathbf{c}}'}(x_t)$ to $g_{\vec{\mathbf{c}}}(x_t)$ for any $t = 1, \dots, T$ where $\mathbf{c}' \sim p_{0,d}$ with $p_{0,d}$ defined in Lemma 1. This problem is specific to the context of bi-clustering where the application $\vec{\mathbf{c}} \mapsto g_{\vec{\mathbf{c}}}(x)$ defined in (1.4) or (1.5) is not continuous.

This algorithm in not adaptive since it depends on unknown quantities such as time horizon T. Adaptive choice of $\tau > 0$ could be performed as in [13]. We can also stress that as in [8], we can avoid the boundedness assumption $\mathcal{Y} \subseteq [0, M]$. In this case, the choice of $\lambda > 0$ in the algorithm will depend on the sequence and an adaptive choice could be investigated. We omit these considerations for concision here.

Proof. Let $\vec{\mathbf{c}} \in \mathbb{R}^{\sum_{j=1}^d m_j p_j}$. Let $\rho = p_{0,d}$ defined in Lemma 1. Applying Proposition 1 with ρ and $\lambda \leq 1/2M^2$ leads to:

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \mathbb{E}_{\vec{\mathbf{c}}' \sim p_0} \sum_{t=1}^T (y_t - g_{\vec{\mathbf{c}}'}(x_t))^2 + \frac{\mathcal{K}(p_{0,d}, \pi_{S,d})}{\lambda}. \tag{2.4}$$

To control the first term in the RHS of (2.4), we use the following decomposition:

$$\sum_{t=1}^{T}(y_t-g_{\vec{\mathbf{c}}'}(x_t))^2=\sum_{t=1}^{T}(y_t-g_{\vec{\mathbf{c}}}(x_t))^2+\sum_{t=1}^{T}(g_{\vec{\mathbf{c}}}(x_t)-g_{\vec{\mathbf{c}}'}(x_t))^2+2\sum_{t=1}^{T}(y_t-g_{\vec{\mathbf{c}}}(x_t))(g_{\vec{\mathbf{c}}}(x_t)-g_{\vec{\mathbf{c}}'}(x_t)).$$

At this stage, it is important to note that for any $x\in\mathcal{X}$, we have by construction of $g_{\vec{\mathbf{c}}}$:

$$|g_{\vec{\mathbf{c}}}(x_t)-g_{\vec{\mathbf{c}}'}(x_t)|\leq M\mathbf{1}_{\exists x_u\in\{x_1,\ldots,x_t\}:f_{\vec{\mathbf{c}}}(x_u)\neq f_{\vec{\mathbf{c}}'}(x_u)},\tag{2.5}$$

where $f_{\vec{\mathbf{c}}}:\prod\mathbb{R}^{m_j}\mapsto\prod\{1,\ldots,p_j\}$ is the nearest neighbor quantizer associated with the d-tensor codebook $\vec{\mathbf{c}}$. Then, integrating the previous inequality, we have to control, for any $v\in\{x_1,\ldots,x_t\}$, the probability $\mathbb{P}(f_{\vec{\mathbf{c}}}(v)\neq f_{\vec{\mathbf{c}}'}(v))$. This is done thanks to Lemma 2-3 in Section 6 for a particular choice of $\delta>0$. We lead to:

$$\begin{aligned}\sum_{t=1}^{T}(y_t-\hat{y}_t)^2&\leq\sum_{t=1}^{T}(y_t-g_{\vec{\mathbf{c}}}(x_t))^2+\frac{\mathcal{K}(p_0,\pi_S)}{\lambda}\\&+24M\sum_{j=1}^{d}|\mathbf{c}_j|_0(|\mathbf{c}_j|_0-1)\frac{m_j\tau^2}{\epsilon^2\delta^2}T(T+1)\left(12M\sum_{j=1}^{d}|\mathbf{c}_j|_0(|\mathbf{c}_j|_0-1)\frac{m_j\tau^2\log T}{\delta^2}+1\right).\end{aligned}$$

The choice of $\tau,\delta>0$ in Theorem 2 concludes the proof. ∎

Theorem 2 holds for a family $\{g_{\vec{\mathbf{c}}},\ \vec{\mathbf{c}}\in\mathcal{C}\}$ satisfying (1.4) or (1.5). An inspection of the proof shows that a sufficient condition for the family $\{g_{\vec{\mathbf{c}}},\ \vec{\mathbf{c}}\in\mathcal{C}\}$ is (2.5). Then, next corollary extends the previous regret bound to a richer class of base forecasters defined as:

$$\{g^{\mathrm{k}}_{\vec{\mathbf{c}}},\ \vec{\mathbf{c}}\in\mathcal{C},\ \mathrm{k}\in\{1,\ldots,N\}\},$$

where for any value of $\mathrm{k}=1,\ldots,N$, (2.5) holds for $g^{\mathrm{k}}_{\vec{\mathbf{c}}}$. Functions $g^{\mathrm{k}}_{\vec{\mathbf{c}}}$ includes the previous cases (1.4) and (1.5) but any other labelizer $g_{\vec{\mathbf{c}}}$ constructed thanks to the set of past observations in the cell associated with $\vec{\mathbf{c}}$ could be considered (such as the median for instance). Interestingly, equipped with such a family of N labelizers, we can add to the learning process the choice of $\mathrm{k}=1,\ldots,N$ by considering the following prior in the algorithm described above:

$$\pi_{S,d,N}d(\vec{\mathbf{c}},\mathrm{k})=\prod_{j=1}^{d}\prod_{i_j=1}^{p_j}\left\{a_\tau\left(1+\frac{|c_{j,i_j}|_2^2}{6\tau^2}\right)^{-\frac{3+m_j}{2}}\right\}d\vec{\mathbf{c}}\times\frac{1}{N}\sum_{\mathrm{k}=1}^{N}\delta_{\mathrm{k}}d\mathrm{k}.\tag{2.6}$$

It leads to a sparsity regret bound with an extra $\log N$ term due to the number of base labelizers:

Corollary 1 *For any deterministic sequence $(x_t,y_t)_{t=1}^T$, consider algorithm of Section 2 using prior $\pi_{S,d,N}$ defined in (2.3) with $\tau=\delta\{\sqrt{24Mp_+}T\}^{-1}$ for some $\delta>0$, $\lambda=1/2M^2$. Then:*

$$\sum_{t=1}^{T}(y_t-\hat{y}_t)^2\leq\inf_{(\vec{\mathbf{c}},\mathrm{k})\in\mathcal{C}\times\{1,\ldots,N\}}\left\{\sum_{t=1}^{T}(y_t-g^{\mathrm{k}}_{\vec{\mathbf{c}}}(x_t))^2+C\sum_{j=1}^{d}(3+m_j)|\boldsymbol{c}_j|_0\log T\right\}+2M^2\log N.$$

This result improves Corollary 2 since the infimum is the RHS could involves different indexes k. The prize to pay is an extra $M^2\log N$ term due to the introduction of the parameter k in the algorithm. For instance, consider the case $\mathcal{Y}=[0,M]$. If $N=2$ in Corollary 1 and the family $\{g^{\mathrm{k}}_{\vec{\mathbf{c}}},\ \vec{\mathbf{c}}\in\mathcal{C},\ \mathrm{k}\in\{1,2\}\}$, is made of forecasters (1.4) and a median estimator, the algorithm performs as well as the best strategy between the mean and the median.

3 Minimax regret

In this contribution, we debate several sparsity regret bounds in online (bi-)clustering. These bounds are stated in the worst case scenario and have shown different behaviour with respect to time horizon T. In online clustering, sparsity regret bounds have a residual term of order $\sqrt{T}\log T$ (see Theorem 1) whereas in bi-clustering, we exhibit better rates in $\log T$. These results are not surprising since many online learning problems give rise to similar bounds, depending on the properties of the loss functions. However, in the setting of online clustering, it is natural to ask if better algorithms exist, i.e. if lower regret could be proved for this problem.

In the context of prediction with expert advices, many authors have investigated the minimax value of the game. Given a sequence $(y_t)_{t=1}^T$, and associated experts advices $\mathbf{p}_t := (p_{t,1}, \ldots, p_{t,N})$, [4] have focused on the absolute loss and proved a minimax value of order $O(\sqrt{T\log N})$. In [10], a unified treatment of the problem is suggested with a general class of loss functions. In this context of prediction with a finite - and static - set of experts, the minimax regret is given by:

$$\mathcal{V}_T(N) := \inf_{(\hat{y}_t)} \sup_{(\mathbf{p}_1,\ldots,\mathbf{p}_T)} \sup_{(y_t)} \left\{ \sum_{t=1}^T \ell(\hat{y}_t, y_t) - \min_{k=1,\ldots,N} \sum_{t=1}^T \ell(p_{t,k}, y_t) \right\},$$

where ℓ is a loss function. Asymptotic behaviours for $\mathcal{V}_T(N)$ when $T \to \infty$ have been stated from $\log N$ to $\sqrt{T\log N}$ depending on particular assumptions over the loss function, such as differentiability. Many examples are provided in [10], including the square loss, the logarithmic loss or the absolute loss.

Very often, the proofs of the lower bounds in the deterministic setting use probabilistic arguments. Surprisingly, by considering stochastic i.i.d. generating processes for the sequence of outcomes, we can achieve tight bounds that match - at least asymptotically[2] - to the upper bounds. The starting point is the following inequality:

$$\mathcal{V}_T(N) \geq \inf_{(\hat{y}_t)} \mathbb{E}_{P^{N\times T}} \mathbb{E}_{Q^T} \left\{ \sum_{t=1}^T \ell(\hat{y}_t, Y_t) - \min_{k=1,\ldots,N} \sum_{t=1}^T \ell(p_{t,k}, y_t) \right\},$$

where $p_{t,k}$ are i.i.d. from P and $Y_1, \ldots, Y_T$ are i.i.d. from Q. The rest of the proof consists in finding particular measures P and Q in order to maximize the lower bound. In this section, we want to state the same kind of result in the context of online clustering. Using simple probabilistic tools, we prove minimax results in the context of online clustering and online bi-clustering.

3.1 Minimax regret in online clustering

In this paragraph, we want to investigate the optimality of Theorem 1. For this purpose, we introduce in the sequel the following assumption:

Sparsity assumption $\mathcal{H}(s)$: *Let $R > 0$ and $T \in \mathbb{N}^*$. Then, there exists a sparsity index $s \in \mathbb{N}^*$ such that $|\boldsymbol{c}_{T,R}^\star|_0 = s$, where:*

$$\boldsymbol{c}_{T,R}^\star := \arg\min_{\boldsymbol{c}\in\mathcal{B}(R)^T} \quad \left\{ \sum_{t=1}^T \ell(\boldsymbol{c}, x_t) + |\boldsymbol{c}|_0 \sqrt{T}\log T \right\},$$

where $\mathcal{B}(R)^T = \Pi_{j=1}^T \mathcal{B}(R)$ and $\mathcal{B}(R)$ is the Euclidean ball of $\mathbb{R}^d$.
This sparsity assumption is related with the structure of the individual sequence $x_t, t = 1, \ldots, T$. It means that the sequence could be well-approximated by s codepoints since the infimum is reached for a sparse codebook $\mathbf{c}_{T,R}^\star$. In what follows, we also introduce the set:

$$\omega_{s,R} := \left\{ (x_t)_{t=1}^T \text{ such that } \mathcal{H}(s) \text{ holds} \right\} \subseteq \mathbb{R}^{dT}.$$

[2] More recently, [1] has given non-asymptotic lower bounds in both statistical and online learning by using the same probabilistic reduction scheme.

With this notation, we have shown essentially in [13] the existence of an algorithm such that:

$$\sup_{(x_t)\in\omega_{s,R}} \left\{ \sum_{t=1}^{T} \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c}\in\mathcal{B}(R)^T} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) \right\} \leq \text{const.} \times s\sqrt{T}\log T.$$

Then, for any $s \in \mathbb{N}^*$, $R > 0$ we could investigate a lower bound according to:

$$\inf_{(\hat{\mathbf{c}}_t)} \sup_{(x_t)\in\omega_{s,R}} \left\{ \sum_{t=1}^{T} \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c}\in\mathcal{B}(R)^s} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) \right\} \geq \text{const.} \times s\sqrt{T}\log T.$$

Following the guiding thread presented above, we can move to a simple probabilistic setting as follows:

$$\inf_{(\hat{\mathbf{c}}_t)} \sup_{(x_t)\in\omega_{s,R}} \left\{ \sum_{t=1}^{T} \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c}\in\mathcal{B}(R)^s} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) \right\} \geq \inf_{(\hat{\mathbf{c}}_t)} \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^{T} \ell(\hat{\mathbf{c}}_t, X_t) - \mathbb{E}_{\nu^N} \min_{k=1,\dots,N} \sum_{t=1}^{T} \ell(\mathbf{c}_k, X_t) \right\},$$

where X_t, $t = 1,\dots,T$ are i.i.d. with law μ and $\mathbf{c}_k$, $k = 1,\dots,N$ are i.i.d. with law ν. Unfortunately, in the inequality above, the infimum is taken over any $(\hat{\mathbf{c}}_t)_{t=1}^T$, that is with no restriction with respect to the ℓ_0-norm. Then, the RHS could be arbitrarely small and the lower bound does not match with the upper bound of Theorem 1. To impose a sparsity assumption for the sequence $(\hat{\mathbf{c}}_t)$, we need to introduced a penalized loss. Next theorem provides minimax results for an augmented value $\mathcal{V}_T(s)$ defined as:

$$\mathcal{V}_T(s) := \inf_{(\hat{\mathbf{c}}_t)} \sup_{(x_t)\in\omega_{s,R}} \left\{ \sum_{t=1}^{T} \left(\ell(\hat{\mathbf{c}}_t, x_t) + \frac{\log T}{\sqrt{T}} |\hat{\mathbf{c}}_t|_0 \right) - \inf_{\mathbf{c}\in\mathcal{B}(R)^s} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) \right\}. \tag{3.1}$$

In (3.1), we add a penalization term for each $\hat{\mathbf{c}}_t$, in terms of ℓ_0-norm. As a result, to capture the asymptotic behaviour of $\mathcal{V}_T(s)$, we also need to state an upper bound with a penalized loss as in (3.1). This is done in the following theorem that combines an upper and lower bound for the minimax regret.

Theorem 3 *Let $s \in \mathbb{N}^*$, $R > 0$ such that:*

$$s \leq \left\lfloor \frac{3}{2} \left(\frac{R\sqrt{T}}{14\log T} \right)^d \right\rfloor. \tag{3.2}$$

Then:

$$s\sqrt{T}\log T(1 + o_T(1)) \leq \mathcal{V}_T(s) \leq s\sqrt{T}(\log T)^2. \tag{3.3}$$

The proof of the first inequality is based on the probabilistic method described above, where we replace the supremum over the individual sequence in $\mathcal{V}_T(s)$ by an expectation (see Section 5-6 for details).

To prove the second inequality, we can use Proposition 1 in [13] to the penalized loss function $\ell_\alpha(\mathbf{c}, x) = \ell(\mathbf{c}, x) + \alpha|\mathbf{c}|_0$ with $\alpha = \log T/\sqrt{T}$ to get:

$$\sum_{t=1}^{T} \ell_\alpha(\hat{\mathbf{c}}_t, x_t) \leq \inf_{\rho\in\mathcal{P}(\mathbb{R}^{dp})} \left\{ \mathbb{E}_{\vec{\mathbf{c}}\sim\rho} \sum_{t=1}^{T} \ell_\alpha(\mathbf{c}, x_t) + \frac{\mathcal{K}(\rho,\pi)}{\lambda} + \frac{\lambda}{2} \mathbb{E}_{(\hat{p}_1,\dots,\hat{p}_T)} \mathbb{E}_{\vec{\mathbf{c}}\sim\rho} \sum_{t=1}^{T} [\ell_\alpha(\mathbf{c}, x_t) - \ell_\alpha(\hat{\mathbf{c}}_t, x_t)]^2 \right\}.$$

A choice of $\alpha = \log T/\sqrt{T}$, $\lambda = 1/\sqrt{T}$, a sparsity prior with scale parameter $\tau = 1/\sqrt{T}$ and $p = \sqrt{T}$ allows to get the desired upper bound.

3.2 Minimax regret in online bi-clustering

In the context of Section 2, we want to prove the minimax optimality of Theorem 2. For this purpose, we introduce a modified sparsity assumption related to the bi-clustering problem:

Sparsity assumption $\mathcal{H}'(s)$: *There exists a sparsity index* $s \in \mathbb{N}^*$ *such that* $|\boldsymbol{c}_T^\star|_0 = s$ *where:*

$$\boldsymbol{c}_T^\star := \arg\inf_{\vec{\boldsymbol{c}}} \quad \left\{ \sum_{t=1}^T (y_t - g_{\vec{\boldsymbol{c}}}(x_t))^2 + |\vec{\boldsymbol{c}}|_0 \log T \right\}.$$

This sparsity assumption is related with the individual sequence $(x_t, y_t), t = 1, \ldots, T$. Loosely speaking, it means the sequence is made of a small number of clusters of inputs with same labels. In what follows, we also introduce:

$$\omega_s' := \left\{ (x_t, y_t)_{t=1}^T \text{ such that } \mathcal{H}'(s) \text{ holds} \right\} \subseteq \mathbb{R}^{dT}.$$

With this notation, we have shown in Theorem 2 the existence of a sequential algorithm $(\hat{y}_t)_{t=1}^T$ such that for any $s \in \mathbb{N}^*$, for T great enough:

$$\sup_{(x_t,y_t)\in\omega_s'} \left\{ \sum_{t=1}^T (\hat{y}_t - y_t)^2 - \inf_{\vec{\boldsymbol{c}}} \sum_{t=1}^T (y_t - g_{\vec{\boldsymbol{c}}}(x_t))^2 \right\} \leq \text{const.} \times s \log T.$$

Then, for any $s \in \mathbb{N}^*$, we will investigate the order of the minimax value:

$$\mathcal{V}_T'(s) = \inf_{(\hat{y}_t)} \sup_{(x_t,y_t)\in\omega_s'} \left\{ \sum_{t=1}^T (\hat{y}_t - y_t)^2 - \inf_{\vec{\boldsymbol{c}}} \sum_{t=1}^T (y_t - g_{\vec{\boldsymbol{c}}}(x_t))^2 \right\}.$$

Following the guiding thread presented above, in this case we can move to a simple probabilistic setting as follows:

$$\mathcal{V}_T'(s) \geq \inf_{(\hat{y}_t)} \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^T (\hat{y}_t - Y_t) - \mathbb{E}_{\nu^N} \min_{k=1,\ldots,N} \{ \sum_{t=1}^T (Y_t - g_{\vec{\mathbf{c}}_k}(X_t) \} \right\},$$

where (X_t, Y_t), $t = 1, \ldots, T$ are i.i.d. with law μ and $\vec{\mathbf{c}}_k$, $k = 1, \ldots, N$ are i.i.d. with law ν. This inequality is at the origin of the following theorem.

Theorem 4 *Suppose* $\mathcal{Y} = \{0, 1\}$ *and* $s = 2$ *for simplicity. Then, for* T *large enough:*

$$\mathcal{V}_T'(s) \geq C_0 \log T,$$

where $C_0 > 0$ *is an absolute constant.*

4 Conclusion

This paper studies the online bi-clustering scenario where we observe a deterministic matrix sequentially. The goal is to predict the entries of the matrix by clustering the columns and rows, thanks to additional feature variables. We prove sparsity regret bounds by using PAC-Bayesian sequential algorithms with sparsity priors. These results are inspired from [13] where the problem of online clustering is investigated. For completeness, we also propose lower bounds for both online clustering and bi-clustering. In this problem, simple probabilistic tools allow us to show the optimality of our algorithms.

5 Proofs

5.1 Proof of Theorem 3

First, we introduce the event $\Omega_{s,R} = \{(X_1, \ldots, X_T : |\mathbf{c}^*_{T,R}|_0 = s\}$. Then, we have:

$$\mathcal{V}_T(s) \geq \inf_{\hat{\mathbf{c}}_t} \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^{T} \ell(\hat{\mathbf{c}}_t, X_t) + \frac{\log T}{\sqrt{T}} |\hat{\mathbf{c}}_t|_0 - \mathbb{E}_{\nu^N} \min_{i=1,\ldots,N} \sum_{t=1}^{T} \ell(C_i, X_t) \right\} I\left(\Omega_{s,R}\right),$$

where $\nu^N \in \mathcal{P}(\mathbb{R}^{dsM})$ is the law of an i.i.d. sample $(C_1, \ldots, C_N)$ of candidates codebooks such that for any k, $|C_k|_0 = s$ and $|C_k|_2 \leq R$ ν-a.s. whereas $\mu^T \in \mathcal{P}(\mathbb{R}^{dT})$ is the law of the i.i.d. sample $(X_1, \ldots, X_T)$. Now, we have to choose the two measures (ν, μ) in order to maximize the RHS.

First of all, since $(X_1, \ldots, X_T)$ are i.i.d. and by definition of $\Omega_{s,R}$, we can write:

$$\begin{aligned}
\inf_{\hat{\mathbf{c}}_t} \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^{T} \ell(\hat{\mathbf{c}}_t, X_t) + \frac{\log T}{\sqrt{T}} |\hat{\mathbf{c}}_t|_0 \right\} I\left(\Omega_{s,R}\right) &\geq \inf_{\hat{\mathbf{c}}} \mathbb{E}_{\mu^{\otimes T}} \left\{ \sum_{t=1}^{T} \ell(\hat{\mathbf{c}}, X_t) + \sqrt{T} \log T |\mathbf{c}|_0 \right\} I\left(\Omega_{s,R}\right) \\
&\geq \mathbb{E}_{\mu^{\otimes T}} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c}^*_T, X_t) + s\sqrt{T} \log T \right\} I\left(\Omega_{s,R}\right) \\
&\geq \mathbb{E}_{\mu^{\otimes T}} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c}^*_T, X_t) \right\} \left(1 - I\left(\Omega_s^C\right)\right) + s\sqrt{T} \log T \mathbb{P}(\Omega_{s,R}) \\
&\geq \mathbb{E}_{\mu^{\otimes T}} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c}^*_T, X_t) \right\} - T\Delta^2 \mathbb{P}(\Omega_{s,R}^C) + s\sqrt{T} \log T \mathbb{P}(\Omega_{s,R}) \\
&\geq T\mathbb{E}_\mu \ell(c^*_\mu, X) - T| \sup_{\mathbf{c}^\star_T, X} \ell(\mathbf{c}^\star_T, X)|_\infty \mathbb{P}(\Omega_{s,R}^C) + s\sqrt{T} \log T \mathbb{P}(\Omega_{s,R}),
\end{aligned}$$

where $\Delta > 0$ is related with the choice of μ (see Lemma 4 in Section 6). Then, by choosing Δ^2 according to Lemma 5, we arrive at:

$$\inf_{\hat{\mathbf{c}}_t} \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^{T} \ell(\hat{\mathbf{c}}_t, X_t) + \frac{\log T}{\sqrt{T}} |\hat{\mathbf{c}}_t|_0 \right\} I\left(\Omega_{s,R}\right) \geq T\mathbb{E}_\mu \ell(c^*_\mu, X) - T\epsilon_T \Delta^2 + s\sqrt{T} \log T(1 - \epsilon_T),$$

where $\epsilon_T > 0$ appears in Lemma 5. Moreover, with Lemma 7, we have for some constant $\alpha_N > 0$:

$$\mathbb{E}_{\mu^T} \mathbb{E}_{\nu^N} \min_{i=1,\ldots,N} \sum_{t=1}^{T} \ell(C_i, X_t) \leq \alpha_N \sqrt{\log N} \frac{\sqrt{T}\Delta^2}{2} + T\mathbb{E}_\nu \mathbb{E}_\mu \ell(C, X). \tag{5.1}$$

The choice of ν in Lemma 7 leads to $\mathbb{E}_\mu \ell(c^*_\mu, X) = \mathbb{E}_\nu \mathbb{E}_\mu \ell(C, X)$ and by choosing N large enough, one has eventually for a constant $a' > 0$:

$$\mathcal{V}_T(s) \geq s\sqrt{T} \log T \left(1 - \epsilon_T \left[1 - \frac{T}{s\Phi_T^{-1}(1 - \epsilon_T)} \right] + a' \frac{\log T \sqrt{\log N}}{s\Phi_T^{-1}(1 - \epsilon_T)} \right), \tag{5.2}$$

where $\Phi_T^{-1}(\cdot)$ is defined in Lemma 7. Furthermore, the choice of $\epsilon_T = 1/T$ gathering with a simple normal approximation of the binomial distribution $B(T, 1/2)$, for T large enough, leads to:

$$\Phi_T^{-1}(1 - \epsilon_T) \leq \frac{T}{2} + \frac{\sqrt{T}}{4} \sqrt{2 \log T} \leq T.$$

Then, we have in (5.2):

$$\mathcal{V}_T(s) \geq s\sqrt{T} \log T \left(1 - \frac{1}{T} \left[1 - \frac{1}{s} + a' \frac{\log T \sqrt{\log N}}{s} \right] \right),$$

which gives the desired result. ∎

5.2 Proof of Theorem 4

We restrict ourselve the $d = 1$ for simplicity. By written $\Omega'_s = \{(X_1, Y_1, \ldots, X_T, Y_T) : |\mathbf{c}^\star_T|_0 = s\}$, where $\mathbf{c}^\star_T$ is defined in $\mathcal{H}'(s)$, we have:

$$\mathcal{V}'_T(s) \geq \inf_{\hat{y}_t} \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^T (\hat{y}_t - Y_t)^2 - \mathbb{E}_{\nu^N} \min_{k=1,\ldots,N} \sum_{t=1}^T (Y_t - g_{\vec{\mathbf{c}}_k}(X_t))^2 \right\} I\left(\Omega'_s\right),$$

where ν is the law of an i.i.d. sample $(\vec{\mathbf{c}}_1, \ldots, \vec{\mathbf{c}}_N)$ of candidates d-tensor codebooks such that for any k, $|\vec{\mathbf{c}}_k|_0 = s$ ν-a.s. whereas μ is the law of the i.i.d. sample $(X_1, Y_1), \ldots, (X_T, Y_T)$. Now, we have to choose the two measures (ν, μ) in order to maximize the RHS.

First of all, since (X_t, Y_t) are i.i.d.:

$$\inf_{\hat{y}_t} \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^T (\hat{y}_t - Y_t)^2 \right\} I\left(\Omega'_s\right) \geq T \left(\min_y \mathbb{E}_Y (Y - y)^2 - \mathbb{P}(|\mathbf{c}^\star_T|_0 > s) \right).$$

A good choice of μ, ν will ensure $\mathbb{P}(|\mathbf{c}^\star_T|_0 > s) \leq c(\log T)/T$ in Lemma 10 below. To control the second term, we use a martingale version of the central limit theorem (see Lemma 9) to get, for some contsant ${}_a lpha_N > 0$:

$$\mathbb{E}_{\mu^T}\mathbb{E}_{\nu^N} \min_{k=1,\ldots,N} \sum_{t=1}^T (Y_t - g_{\vec{\mathbf{c}}_k}(X_t))^2 \leq \alpha_N \sqrt{T} s_T \sqrt{\log N} + \mathbb{E}_{\mu^T}\mathbb{E}_{\nu^N} \sum_{t=1}^T \mathbb{E}\left[(Y_t - g_{\vec{\mathbf{c}}}(X_t))^2 | \mathcal{F}_{t-1} \right],$$

where $s_T^2 = \sum_{t=1}^T \mathrm{Var}(Y_t - g_{\vec{\mathbf{c}}_k}(X_t))^2$ and $\mathcal{F}_{t-1} = \sigma(\{(X_u, Y_u), u = 1, \ldots, t-1\})$. Then, we obtain:

$$\mathcal{V}'_T(s) \geq T \min_y \mathbb{E}_Y (Y - y)^2 - \mathbb{E}_{\mu^T}\mathbb{E}_{\nu^N} \sum_{t=1}^T \mathbb{E}\left[(Y_t - g_{\vec{\mathbf{c}}}(X_t))^2 | \mathcal{F}_{t-1} \right] + \alpha_N s_T \sqrt{\log N} - \log T.$$

Now, the choice of μ, ν in Lemma 8 gives:

$$\mathbb{E}_{\mu^T}\mathbb{E}_{\nu^N} \sum_{t=1}^T \mathbb{E}\left[(Y_t - g_{\vec{\mathbf{c}}}(X_t))^2 | \mathcal{F}_{t-1} \right] = T \min_y \mathbb{E}_Y (Y - y)^2 + \sum_{t=1}^T \mathbb{E}(\epsilon - g_{\vec{\mathbf{c}}_k}(X_t))^2,$$

where $\epsilon > 0$ is such that Y_t is a Bernoulli variable with parameter $\epsilon > 0$. The precise value of ϵ will be chosen in the sequel. We hence get:

$$\mathcal{V}'_T(s) \geq -\sum_{t=1}^T \mathbb{E}(\epsilon - g_{\vec{\mathbf{c}}}(X_t))^2 + \alpha_N s_T \sqrt{\log N} - \log T.$$

Last step is to compute the order of s_T and $\sum_{t=1}^T \mathbb{E}(\epsilon - g_{\vec{\mathbf{c}}}(X_t))^2$. This is done is Lemma 8 and a choice of $\epsilon = \log T^2/T$ leads to:

$$\mathcal{V}'_T(s) \geq -\sum_{t=1}^T a_t + \alpha_N s_T \sqrt{\log N} - \log T \geq \log T,$$

where:

$$a_t \sim \frac{\epsilon}{t} + \frac{1}{t} \text{ and } s_T = \sqrt{\sum_{t=1}^T \epsilon + \frac{\epsilon}{t}}.$$

∎

6 Appendix

6.1 Auxiliary lemmas for Theorem 2

The proof of Theorem 2 is based on the control of the following probability:

Lemma 2 *For any $\vec{\boldsymbol{c}} \in \mathcal{C}$, let $\vec{\boldsymbol{c}}' \sim p_{0,d}$ where $p_{0,d}$ is defined in Lemma 1. Then for any $v \in \mathcal{X}$, for any $\delta, \epsilon > 0$, we have:*

$$\mathbb{P}_{\vec{\boldsymbol{c}}'}\left(f_{\vec{\boldsymbol{c}}}(v) \neq f_{\vec{\boldsymbol{c}}'}(v)\right) \leq 24 \sum_{j=1}^{d} |\boldsymbol{c}_j|_0(|\boldsymbol{c}_j|_0 - 1)\frac{m_j \tau^2}{\epsilon^2 \delta^2},$$

provided that the following condition holds:

$$\forall j = 1, \ldots, d,\ \psi_{v_j, \boldsymbol{c}_{j,i_j}}(\epsilon) \geq \epsilon \text{ and } |v_j - \delta V(\boldsymbol{c}_j)| \geq \epsilon, \tag{6.1}$$

where:

$$\psi_{v_j, \boldsymbol{c}_{j,i_j}}(\epsilon) = \sqrt{d(v_j, [c_{j,i_j}, c_{j,i'_j}])^2 + \left(d(c_{j,i_j}, \partial V(\boldsymbol{c}_j) + \epsilon\right)^2} - \sqrt{d(v_j, [c_{j,i_j}, c_{j,i'_j}])^2 + \left(d(c_{j,i_j}, \partial V(\boldsymbol{c}_j) - \epsilon\right)^2}.$$

Proof: Notice that, for any $v \in \mathcal{X}$:

$$\begin{aligned}\mathbb{P}_{\vec{\boldsymbol{c}}'}(\{f_{\vec{\boldsymbol{c}}}(v) \neq f_{\vec{\boldsymbol{c}}'}(v)\}) &= \mathbb{P}\left((\exists j \in \{1, \ldots, d\},\ \exists (i_j, i'_j) \in \{1, \ldots, |\mathbf{c}_j|_0\}^2 : \pi_j(f_{\vec{\boldsymbol{c}}}(v)) = i_j \neq i'_j = \pi_j(f_{\vec{\boldsymbol{c}}'}(v))\right) \\ &\leq \sum_{j=1}^{d} \sum_{(i_j, i'_j)} \mathbb{E}_{c'_{j,i'_j}} \int_{\mathcal{B}(v_j, |v_j - c'_{j,i'_j}|)^C} dp_0^{j,i_j}(c'_{j,i_j}) d(c'_{j,i_j}),\end{aligned}$$

where $p_0(d\mathbf{c}') = \Pi_{j=1}^{2} \Pi_{i_j=1}^{p_j} p_0^{j,i_j}(dc'_{j,i_j})$ and $\mathbb{E}_{c'_{j,i'_j}}$ is the expectation with respect to p_0^{j,i'_j} whereas $\mathcal{B}(v_j, |v_j - c'_{j,i'_j}|)$ is the Euclidean ball in $\mathbb{R}^{m_j}$ with center v_j and radius $|v_j - c'_{j,i'_j}|$. Then, we get for any $\epsilon, \delta > 0$:

$$\begin{aligned}&\mathbb{P}_{\vec{\boldsymbol{c}}'}(\{f_{\vec{\boldsymbol{c}}}(v) \neq f_{\vec{\boldsymbol{c}}'}(v)\}) \leq \\ &\sum_{j=1}^{d} \sum_{(i_j, i'_j)} \left(\mathbb{E}_{c'_{j,i'_j}} \int_{\mathcal{B}(v_j, |v_j - c'_{j,i'_j}|)^C} dp_0^{j,i_j}(c'_{j,i_j}) d(c'_{j,i_j}) \mathbf{1}_{|c'_{j,i'_j} - c_{j,i'_j}| \leq \epsilon\delta/2} + \mathbb{P}_{c'_{j,i'_j}}\left(|c'_{j,i'_j} - c_{j,i'_j}| > \frac{\epsilon\delta}{2}\right) \right) \\ &\leq \sum_{j=1}^{d} \sum_{(i_j, i'_j)} \left(\mathbb{E}_{c'_{j,i'_j}} \int_{\mathcal{B}(v_j, |v_j - c'_{j,i'_j}|)^C} dp_0^{j,i_j}(c'_{j,i_j}) d(c'_{j,i_j}) \mathbf{1}_{|c'_{j,i'_j} - c_{j,i'_j}| \leq \epsilon\delta/2} + \mathbb{P}\left(|\mathcal{T}_{m_j}(3)| > \frac{\epsilon\delta}{2\tau}\right) \right).\end{aligned}$$

Last step is to control the first term is the previous decomposition. Using simple geometry, we can notice that for any $\epsilon, \delta > 0$, if $|v_j - \partial V(\mathbf{c}_j)|_2 > \epsilon$, the following assertion holds:

$$\psi_{v_j, \mathbf{c}_{j,i_j}}(\epsilon)/\epsilon \geq \delta \Longrightarrow \mathcal{B}(c_{j,i_j}, \epsilon\delta/2) \subseteq \mathcal{B}(v_j, d(v_j, \mathcal{B}(c_{j,i'_j}, \epsilon\delta/2))),$$

where $\psi_{v_j, \mathbf{c}_{j,i_j}}(\epsilon)$ is defined in the lemma. We then obtain, provided that $\psi_{v_j, \mathbf{c}_{j,i_j}}(\epsilon) \geq \epsilon$ and $|v_j - \delta V(\mathbf{c}_j)| \geq \epsilon$:

$$\begin{aligned}\mathbb{P}_{\mathbf{c}'}(\{f_{\vec{\boldsymbol{c}}}(v) \neq f_{\vec{\boldsymbol{c}}'}(v)\})) &\leq \sum_{j=1}^{d} \sum_{(i_j, i'_j)} \left(\int_{\mathcal{B}(c_{j,i_j}, \epsilon\delta/2)^C} dp_0^{j,i_j}(c'_{j,i_j}) d(c'_{j,i_j}) + \mathbb{P}(|\mathcal{T}_{m_j}(3)| > \frac{\epsilon\delta}{2\tau}) \right) \\ &= 2 \sum_{j=1}^{d} |\mathbf{c}_j|_0(|\mathbf{c}_j|_0 - 1)\mathbb{P}(|\mathcal{T}_{m_j}(3)| > \frac{\epsilon\delta}{2\tau}) \\ &\leq 2 \sum_{j=1}^{d} |\mathbf{c}_j|_0(|\mathbf{c}_j|_0 - 1)\frac{12 m_j \tau^2}{\epsilon^2 \delta^2},\end{aligned}$$

where we use in last line a simple Markov's inequality to $|\mathcal{T}_{m_j}(3)|^2/d_j \sim \mathcal{F}(m_j, 3)$.

Lemma 3 *For any $\vec{\mathbf{c}} \in \mathcal{C}$, for T great enough, for any $x_1, \dots, x_t \in \mathcal{X}$, let us consider $\epsilon = 1/\sqrt{\log T}$ and:*

$$\delta = \min_{v \in \{x_1,\dots,x_t\}} \min_{j=1,2} \frac{d(c_{j,i_j}, \partial V(\boldsymbol{c}_j))}{\sqrt{d(v_j, [c_{j,i_j}, c_{j,i'_j}])^2 + \left(d(c_{j,i_j}, \partial V(\boldsymbol{c}_j)\right)^2}}.$$

Then, (6.1) *holds for any $v \in \{x_1, \dots, x_t\}$.*

Proof of the lemma : First notice that we can suppose the set of codebooks $\vec{\mathbf{c}}$ is such that $(x_t)_{t=1}^T \cap \cup_j \partial V(\mathbf{c}_j) = \emptyset$. Indeed, if this condition does not hold, we can slightly modify $\vec{\mathbf{c}}$ to have the desired property without changing the value of $g_{\vec{\mathbf{c}}}(x_t)$, for any $t = 1, \dots, T$. Moreover, for T great enough, for any $v \in \mathcal{X}$:

$$\frac{\psi_{v_j, \mathbf{c}_{j,i_j}}(\epsilon)}{2\epsilon} := \frac{\phi(a+\epsilon) - \phi(a-\epsilon)}{2\epsilon},$$

where $\phi(u) = \sqrt{b^2 + u^2}$, $a = d(c_{j,i_j}, \partial V(\mathbf{c}_j)$ and $b = d(v_j, [c_{j,i_j}, c_{j,i'_j}])$. Then, if $\epsilon := \epsilon(T) \to 0$, for any $\epsilon' > 0$, $\exists T_0(\epsilon') \in \mathbb{N}$ such that for any $T \geq T_0(\epsilon')$:

$$\phi'(a) = \frac{d(c_{j,i_j}, \partial V(\mathbf{c}_j)}{\sqrt{d(v_j, [c_{j,i_j}, c_{j,i'_j}])^2 + \left(d(c_{j,i_j}, \partial V(\mathbf{c}_j)\right)^2}} \geq \frac{\delta}{2} + \epsilon' \Rightarrow \psi_{v_j, \mathbf{c}_{j,i_j}}(\epsilon)/\epsilon \geq \delta.$$

Hence, since $d(c_{j,i_j}, \partial V(\mathbf{c}_j)) > 0$, we can choose

$$\delta \leq \min_{v \in \{x_1,\dots,x_t\}} \min_{j=1,2} \frac{d(c_{j,i_j}, \partial V(\mathbf{c}_j))}{\sqrt{d(v_j, [c_{j,i_j}, c_{j,i'_j}])^2 + \left(d(c_{j,i_j}, \partial V(\mathbf{c}_j)\right)^2}}$$

provided that T large enough to have

$$\epsilon' < \min_{v \in \{x_1,\dots,x_t\}} \min_{j=1,2} \frac{d(c_{j,i_j}, \partial V(\mathbf{c}_j)}{\sqrt{d(v_j, [c_{j,i_j}, c_{j,i'_j}])^2 + \left(d(c_{j,i_j}, \partial V(\mathbf{c}_j)\right)^2}}.$$

Then, by choosing $\epsilon = 1/\sqrt{\log T}$, (6.1) holds for any $v \in \{x_1, \dots, x_t\}$.

6.2 Auxiliary lemmas for Theorem 3-4

6.2.1 Online clustering

The proof of Theorem 3 is based on the following intermediate lemmas.

Lemma 4 *Let $s \in \mathbb{N}^*$ is divisible by 3. Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ a distribution concentrated on $2m = 4s/3$ points $\mathcal{S}_\mu := \{z_i, z_i + w,\ i = 1, \dots, m\}$ such that $w = (2\Delta, 0, \dots, 0) \in \mathbb{R}^d$ with $\Delta > 0$. Suppose for any $i \neq j$, $d(z_i, z_j) \geq A\Delta$, for some $A > 0$. Define μ as the uniform distribution over $\mathcal{S}_\mu$. Then, if $A \geq \sqrt{2} + 1$, we have:*

$$\arg\min_{\boldsymbol{c} \in \mathbb{R}^{ds}} \mathbb{E}_\mu \ell(\boldsymbol{c}, X) = \{\boldsymbol{c}^\star_{\mu,1}, \dots, \boldsymbol{c}^\star_{\mu,k}\} =: \mathcal{M}_\mu,$$

where $k = \binom{m}{m/2}$ and $\boldsymbol{c}^\star_{\mu,j}$ is such that for $m/2$ values of $i \in \{1, \dots, m\}$, $\boldsymbol{c}^\star_{\mu,j}$ has codepoints at both z_i and $z_i + w$ and for the remaining $m/2$ values of i, $\boldsymbol{c}^\star_{\mu,j}$ has a single codepoint at $z_i + w/2$.

Proof : The proof of this claim can be found in the Appendix of [2].

Next lemma controls the probability that $|\mathbf{c}^\star_T|_0 > s$ with a proper choice of Δ^2 in the definition of μ.

Lemma 5 *Let $s \in \mathbb{N}^*$ and assume s is divisible by 3. Let μ defined in Lemma 4. Then, if we choose:*

$$\Delta^2 \leq \frac{\sqrt{T}\log T}{\Phi_T^{-1}(1-\epsilon_T)},$$

where $\Phi_T^{-1}(\cdot)$ is the generalized inverse of $\Phi_T : x \mapsto \mathbb{P}(B(T,1/2) \leq x)$, we have:

$$\mathbb{P}(|\boldsymbol{c}_T^\star|_0 > s) \leq \epsilon_T,$$

where $(X_1, \ldots, X_T)$ are i.i.d. with law μ and:

$$\boldsymbol{c}_T^\star := \arg\min_{\boldsymbol{c}\in\mathbb{R}^{dT}} \left\{ \sum_{t=1}^T \ell(\boldsymbol{c}, X_t) + \sqrt{T}\log T |\boldsymbol{c}|_0 \right\}.$$

Proof of the Lemma : We consider the case $s = 3$ for simplicity. In this case, by construction of μ, we have $|\mathcal{S}_\mu| = 4$ and then $|\mathbf{c}_T^\star|_0 \leq 4$ $\mu-$almost surely. We hence have the following assertion:

$$\sum_{t=1}^T \ell(c_3^\star, X_t) - \ell(c_4^\star, X_t) \leq \sqrt{T}\log T \Rightarrow |\mathbf{c}_T^\star|_0 \leq 3,$$

where $\mathbf{c}_k^\star := \arg\min_{|\mathbf{c}|_0=k} \sum_{t=1}^T \ell(\mathbf{c}, X_t)$ for any $k \in \mathbb{N}^*$. To end up the proof we have to show the following intermediate lemma:

Lemma 6 *Let $\boldsymbol{c}_k^\star$ defined above for some $k \in \mathbb{N}^\star$. Then, if $A \geq 7$ in Lemma 4, almost surely, for $m/2$ values of $i \in \{1, \ldots, m\}$, $\boldsymbol{c}_k^\star$ has codepoints at both z_i and $z_i + w$, and for the remaining $m/2$ values of i, $\boldsymbol{c}_k^\star$ has a single codepoint at $z_i + w/2$.*

Proof of the intermediate lemma: For any codebook $\vec{\mathbf{c}} = \{c_1, \ldots, c_k\}$, we consider $V_i = V(z_i) \cup V(z_i + w)$ and denote by m_i the number of points in $\mathbf{c} \cap V_i$. Then we consider the following codebook $\mathbf{c}'$. For any $i = 1, \ldots, m$:

- if $m_i \geq 2$, $z_i, z_i + w$ are codepoints of $\mathbf{c}'$,
- if $m_i = 1 \cup 0$, $z_i + w/2$ is a codepoints of $\mathbf{c}'$.

Note that $\mathbf{c}'$ could have more than k codepoints but suppose for a moment that $\mathbf{c}'$ has excactly k codepoints. We want to show that for T large enough, we have a.s.:

$$1/T \sum_{t=1}^T \ell(\mathbf{c}', X_t) \leq 1/T \sum_{t=1}^T \ell(\mathbf{c}, X_t). \tag{6.2}$$

Deote $r_i(\mathbf{c}) = |z_i - \mathbf{c}(z_i)|_2^2 \mu_T(z_i) + |z_i + w - \mathbf{c}(z_i + w)|_2^2 \mu_T(z_i + w)$, where $\mu_T = 1/T \sum_{t=1}^T \delta_{X_t}$ is the empirical measure and $\mathbf{c}(z)$ is the codepoint in $\vec{\mathbf{c}}$ associated with z. Then we have $\sum_{t=1}^T \ell(\mathbf{c}', x_t) = T \sum_{i=1}^m r_i(\mathbf{c})$. We will show that for any $i = 1, \ldots, m$, $r_i\mathbf{c}') \leq r_i(\mathbf{c})$. If $m_i \geq 2$, we have clearly $r_i(\mathbf{c}') = 0 \leq r_i(\mathbf{c})$. If $m_i = 1$ and $\mathbf{c}(z_i) = \mathbf{c}(z_i + w)$, then $r_i(\mathbf{c}') \leq r_i(\mathbf{c})$ since $\mathbf{c}'$ has a codepoint at $z_i + w/2$. If $m_i = 1$ and $\mathbf{c}(z') \notin V_i$ for $z' \in \{z_i, z_i + w\}$, we have:

$$r_i(\mathbf{c}) \geq \mu_T(z_i)|z_i - \mathbf{c}(z_i)|_2^2 \geq \mu_T(z_i) \left(\frac{A\Delta}{2} - \Delta \right)^2,$$

by construction of measure μ in Lemma 4. Then, since in this case $r_i(\mathbf{c}') = \mu_T(z_i, z_i + w)\Delta^2$, we have $r_i(\mathbf{c}') \leq r_i(\mathbf{c})$ if:

$$\mu_T(z_i, z_i + w)\Delta^2 \leq \mu_T(z_i) \left(\frac{A\Delta}{2} - \Delta \right)^2. \tag{6.3}$$

Then, for any $\epsilon > 0$, for any $T > T_\epsilon$, if $A \geq 2(3+\epsilon)$, (6.3) holds true.

Now if $m_i = 0$, $\mathbf{c}(z_i)$ and $\mathbf{c}(z_i + w)$ are not in V_i. Then $r_i(\mathbf{c}) \geq \mu_T(z_i, z_i + w)\left(A\Delta/2 - \Delta\right)^2$ and since $r_i(\mathbf{c}') = \mu_T(z_i, z_i + w)\Delta^2$, we have $r_i(\mathbf{c}') \leq r_i(\mathbf{c})$ for $A \geq 4$. Then, we have (6.2) for instance for $A \geq 7$ by choosing $\epsilon = 1/2$ above.

Now suppose $|\mathbf{c}'|_0 > k'$. Then we can choose $k' - k$ couples $(z_i, z_i + w)$ which are codepoints of $\mathbf{c}'$ and replace them by $z_i + w/2$. Then if we denote by $\mathbf{c}_k^\star$ this quantizer, we have clearly $|\mathbf{c}_k^\star|_0 = k$ and:

$$1/T \sum_{t=1}^{T} \ell(\mathbf{c}^\star, X_t) \leq 1/T \sum_{t=1}^{T} \ell(\mathbf{c}', X_t) + \sum_{i=1}^{k'-k} \mu_T(z_i, z_i + w)\Delta^2.$$

Moreover, there is at least $k' - k$ indices where $m_i = 0$, then we have also:

$$1/T \sum_{t=1}^{T} \ell(\mathbf{c}', X_t) \leq 1/T \sum_{t=1}^{T} \ell(\mathbf{c}, X_t) - \sum_{i=1}^{k'-k} \mu_T(z_i', z_i' + w)\left(\left[\frac{A\Delta}{2} - \Delta\right]^2 - \Delta^2\right).$$

Then, gathering with the previous inequality, we arrive at:

$$1/T \sum_{t=1}^{T} \ell(\mathbf{c}^\star, X_t) \leq 1/T \sum_{t=1}^{T} \ell(\mathbf{c}, X_t) + \sum_{i=1}^{k'-k} \mu_T(z_i, z_i + w)\left(\left[\frac{A\Delta}{2} - \Delta\right]^2 - \Delta^2\right) - \sum_{i=1}^{k'-k} \mu_T(z_i', z_i' + w)\Delta^2.$$

To control the RHS, note that for any $\epsilon > 0$, there exists T_ϵ such that for $T \geq T_\epsilon$, we have $\mu_T(z_i, z_i + w) \leq 1/m + \epsilon$ and $\mu_T(z_i', z_i' + w) \geq 1/m - \epsilon$. Then, for T large enough, we have the result since $A \geq 7$.

Using this intermediate lemma, we hence have:

$$\sum_{t=1}^{T} \ell(c_3^\star, X_t) - \ell(c_4^\star, X_t) = \sum_{t=1}^{T} \ell(c_3^\star, X_t) \leq \Delta^2 \mathrm{card}\{t = 1, \ldots, T : X_t \in \{z_i, z_i + w\}\},$$

where $i = 1$ or $i = 2$. Noting that the random variable $\mathrm{card}\{t = 1, \ldots, T : X_t \in \{z_i, z_i + w\}\}$ has a Binomial distribution $\mathcal{B}(T, 1/2)$, We arrive at:

$$\mathbb{P}(\mathbf{c}_T^\star|_0 > 3) \leq \mathbb{P}\left(\mathcal{B}(T, 1/2) \geq \frac{\sqrt{T}\log T}{\Delta^2}\right).$$

The choice of Δ in Lemma 5 concludes the proof.

Eventually, the following lemma controls the expectation of the minimum of the cumulative loss in our probabilistic setting.

Lemma 7 *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ defined in Lemma 4, $\nu \in \mathcal{P}(\mathbb{R}^{ds})$ the uniform law over $\mathcal{M}_\mu = \{\boldsymbol{c}_{\mu,1}^\star, \ldots, \boldsymbol{c}_{\mu,k}^\star\}$ is defined in Lemma 4 . Then, for any $\epsilon > 0$, there exists $T_\epsilon \in \mathbb{N}^*$ such that for any $T \geq T_\epsilon$:*

$$\mathbb{E}_{\mu^T}\mathbb{E}_{\nu^N} \min_{i=1,\ldots,N} \sum_{t=1}^{T} \ell(C_i, X_t) \leq (\epsilon - a_N)\sqrt{\log N}\frac{\sqrt{T}\Delta^2}{2} + T\mathbb{E}_\nu\mathbb{E}_\mu\ell(C, X),$$

where $\lim_{N\to\infty} a_N = \sqrt{2}$.
Moreover, if:

$$s \leq \left\lfloor \frac{3}{2}\left(\frac{R - 2\Delta}{A\Delta}\right)^d \right\rfloor, \tag{6.4}$$

then $|C_k|_2 \leq R$ almost surely, when C_k is the uniform law over $\mathcal{M}_\mu$.

Proof of the lemma: Let us fix a sequence $\mathbf{c}_i \in \{\mathbf{c}^\star_{\mu,1}, \mathbf{c}^\star_{\mu,2}\}, i = 1, \dots, N$ and consider the random variable:

$$\xi_{i,T} := \frac{\sum_{t=1}^T \ell(\mathbf{c}_i, X_t) - \sum_{t=1}^T \mathbb{E}_\mu \ell(\mathbf{c}_i, X_t)}{\sqrt{\sum_{t=1}^T \mathrm{Var}_\mu \ell(\mathbf{c}_i, X_t)}}.$$

For a fixed sequence $(\mathbf{c}_i)$, for any $t \neq t'$, the random variables $\ell(\mathbf{c}_i, X_t)$ and $\ell(\mathbf{c}_i, X_{t'})$ are independent with bounded moment of order 2. Then, with the central limit theorem, we have that for any $i = 1, \dots, N$, each $\xi_{i,T}$ converges in law to a standard normal distribution when $T \to \infty$. Then, applying dominated convergence theorem leads to:

$$\lim_{T\to\infty} \mathbb{E}_{\nu^{\otimes N}} \mathbb{E}_{\mu^{\otimes T}} \min_{i=1,\dots,N} \xi_{i,T} = -a_N \sqrt{\log N}, \tag{6.5}$$

since we know that $\mathbb{E} \min_{i=1,\dots,N} \zeta_i = -a_N \sqrt{\log N}$ when ζ_i are i.i.d. standard normal distribution. Now, note that we want a same kind of result for $\sum_{t=1}^T \ell(\mathbf{c}_i, X_t)$ instead of the normalized random variables $\xi_{i,T}$. By definition of $\mu_{\Delta,m}$ and $\mathbf{c}_i$, $i = 1, \dots, N$, we have coarsely that:

$$\forall i,\ \forall t,\ \mathrm{Var}_\mu \ell(\mathbf{c}_i, X_t) = 1/2\left(\Delta^2 - \Delta^2/2\right)^2 + 1/2\left(0 - \Delta^2/2\right)^2 = \Delta^4/4.$$

Then, $\xi_{i,T} = 2(\sum_{t=1}^T \ell(\mathbf{c}_i, X_t) - \sum_{t=1}^T \mathbb{E}_\mu \ell(\mathbf{c}_i, X_t))/(\sqrt{T}\Delta^2)$ and with (6.5), we have the desired result.

To show that $|C_k|_2 \leq R$, we need to place m pairs of points $(z_i, z_i + 2\Delta)$ in $\mathcal{B}(R)$ such that $|z_i - z - j|_2 \geq A$. It is well-known (see [2]) that m points could be packed in $\mathcal{B}(R - 2\Delta)$ as long as:

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

Then, condition (6.4) gives the result since $2s = 3m$.

6.2.2 Online bi-clustering

The proof of Theorem 4 is based on the following lemmas.

Lemma 8 *Let $\mu \in \mathcal{P}(\mathbb{R}^d \times \{0,1\})$ such that for a generic couple $(X, Y) \sim \mu$, Y is a Bernoulli with parameter $\epsilon > 0$, $X \in \{a, b, c, d\}$ such that $Y|X = x$ is a Bernoulli with parameter p_x where $(p_a, p_b, p_c, p_d) = (1-\epsilon/4, 1-3\epsilon/4, \epsilon/2, \epsilon/2)$. Moreover, let $\nu \in \mathcal{P}(\mathcal{C})$ such that $\nu = 1/2(\delta_{\vec{\mathbf{c}}_1} + \delta_{\vec{\mathbf{c}}_2})$, where $V_{\vec{\mathbf{c}}_1}(a) = V_{\vec{\mathbf{c}}_1}(c)$, $V_{\vec{\mathbf{c}}_1}(b) = V_{\vec{\mathbf{c}}_1}(b)$ whereas $V_{\vec{\mathbf{c}}_2}(a) = V_{\vec{\mathbf{c}}_2}(b)$, $V_{\vec{\mathbf{c}}_2}(c) = V_{\vec{\mathbf{c}}_2}(d)$. Then, if $\epsilon := o_T(1)$, we have:*

$$\mathbb{E}\sum_{t=1}^T (\epsilon - g_{\vec{\mathbf{c}}}(X_t))^2 = (\log T + T\epsilon)(1 + o_T(1)) \text{ and } s_T = \sqrt{\log T + T\epsilon}(1 + o_T(1)).$$

Proof. By construction, denoting $x_u = \mathbb{P}(X_u)$, for $u \in \{a, b, c, d\}$, we have $x_a = x_b = \epsilon/(2(1-\epsilon))$ and $x_c = x_d = (2 - 3\epsilon)/(2(1 - \epsilon))$. Then, one obtains for any $t = 1, \dots, T$:

$$\mathbb{E}(\epsilon - g_{\vec{\mathbf{c}}}(X_t))^2 = \mathbb{E}_{(X,Y)_1^{t-1}} \sum_{(u,v)\in\Lambda} \frac{x_u + x_v}{2}(\epsilon - \bar{Y}^u_{t-1} - \bar{Y}^v_{t-1})^2,$$

where $\Lambda = \{(a, b), (a, c), (b, d), (c, d)\}$, $\bar{Y}^u_{t-1}$ is the empirical mean at time $t - 1$ when $X_k = u$. Moreover, if $x_{u,v} = x_u + x_v$, $\epsilon_{u,v} = \mathbb{P}(Y = 1|X \in \{u, v\})$ and $\bar{Y}^{u,v}_{t-1}$ is the empirical mean when $X_k \in \{u, v\}$ for $k = 1, \dots, t - 1$, by simple algebra, for a given $(u, v) \in \Lambda$:

$$\frac{x_u + x_v}{2}\mathbb{E}_{(X,Y)_1^{t-1}}(\epsilon - \bar{Y}^u_{t-1} - \bar{Y}^v_{t-1})^2 = \frac{x_u + x_v}{2}\mathbb{E}_{X_1^{t-1}}\mathbb{E}\left[\epsilon^2 + (\bar{Y}^{u,v}_{t-1})^2 - 2\epsilon\bar{Y}^{u,v}_{t-1}|X_1, \dots, X_{t-1}\right]$$

$$\frac{x_u + x_v}{2}\mathbb{E}_{X_1^{t-1}}\mathbb{E}\left[\epsilon^2 + \frac{\epsilon_{u,v}(1 - \epsilon_{u,v})}{\mathrm{card}\{u = 1, \dots, t-1 : X_u \in \{u, v\}\}} + \epsilon^2_{u,v} - 2\epsilon\epsilon_{u,v}|X_1, \dots, X_{t-1}\right] := \varphi(u, v).$$

Now, we use the asymptotic expansion of inverse moment of a binomial distribution stated in [20] as follows. If Y is a Bernoulli with parameter (n,p), we have:

$$\begin{aligned}\mathbb{E}Y^{-1} &= \frac{1}{np}\left(1+\frac{1-p}{np}+\frac{(1-p)(2-p)}{n^2p^2}+\cdots\right)\\ &= \frac{1}{np}\left(1+\mathcal{O}_{np}(1)\right).\end{aligned}$$

Then, applying the previous expansion for $Y=\bar{Y}_{t-1}^{u,v}$, and summing over $(u,v)\in\Lambda$, if $\epsilon=o_T(1)$, we obtain:

$$\sum_{t=1}^{T}\mathbb{E}(\epsilon-g_{\vec{\mathbf{c}}}(X_t))^2 \sim \sum_{t=1}^{T}\epsilon+\frac{1}{t}\sim \epsilon T+\log T. \tag{6.6}$$

By denoting $\mathcal{F}_{t-1}=\sigma(X_1,\ldots,X_{t-1},Y_1,\ldots,Y_{t-1})$, the computation of s_T is decomposed as follows:

$$\begin{aligned}s_T &= \sum_{t=1}^{T}\mathbb{E}_{(X,Y)_1^{t-1}}\mathbb{E}\left[(Y_t-g_{\vec{\mathbf{c}}}(X_t))^2-\mathbb{E}[(Y_t-g_{\vec{\mathbf{c}}}(X_t))^2|\mathcal{F}_{t-1}]\right]^2\\ &= \sum_{t=1}^{T}\Big(\mathbb{E}\left[(Y_t-\epsilon)^2-\mathbb{E}[(Y_t-\epsilon)^2|\mathcal{F}_{t-1}]\right]^2+\mathbb{E}\left[(\epsilon-g_{\vec{\mathbf{c}}}(X_t))^2-\mathbb{E}\left[(\epsilon-g_{\vec{\mathbf{c}}}(X_t))^2|\mathcal{F}_{t-1}\right]\right]^2\\ &+4\mathbb{E}(Y_t-\epsilon)^2(\epsilon-g_{\vec{\mathbf{c}}}(X_t))^2\Big).\end{aligned}$$

Then, with (6.6) and simple algebra, one gets the result.

Lemma 9 *Let $\xi_t=(Y_t-g_{\vec{\mathbf{c}}}(X_t))^2$ for any t. Then, we have:*

$$\frac{1}{s_T}\sum_{t=1}^{T}\left(\xi_t-\mathbb{E}(\xi_t|\mathcal{F}_{t-1})\right)\longrightarrow\mathcal{N}(0,1).$$

The proof is based on a martingale version of central limit theorem due to [3]. We hence have to check the Lindenberg condition for martingale:

$$\forall\delta>0,\ \frac{1}{s_T^2}\sum_{t=1}^{T}\mathbb{E}\left[(\xi_t-\mathbb{E}(\xi_t|\mathcal{F}_{t-1}))^2\mathbf{1}_{|\xi_t|>\delta s_T}|\mathcal{F}_{t-1}\right]\to 0 \text{ in proba. as } T\to\infty.$$

By using Lemma 8, the boundness of ξ_t and a simple Markov inequality, we have:

$$\begin{aligned}&\mathbb{E}\left[(\xi_t-\mathbb{E}(\xi_t|\mathcal{F}_{t-1}))^2\mathbf{1}_{|\xi_t|>\delta s_T}|\mathcal{F}_{t-1}\right]\le\mathbb{P}\left(|\xi_t|>\delta s_T|\mathcal{F}_{t-1}\right)\\ &\le\frac{\mathrm{Var}\left((Y_t-\epsilon)^2|\mathcal{F}_{t-1}\right)+\mathrm{Var}\left(\epsilon-g_{\vec{\mathbf{c}}}(X_t))^2|\mathcal{F}_{t-1}\right)+\epsilon(1-\epsilon)\mathbb{E}\left((\epsilon-g_{\vec{\mathbf{c}}}(X_t))^2|\mathcal{F}_{t-1}\right)}{s_T^2\delta^2}\\ &\le\frac{\epsilon(1+(1/t)+\epsilon/(t-1)^2)}{s_T^2\delta^2}.\end{aligned}$$

Then, summing over t, provided that $\epsilon\sim(\log T)^2/T$, we have for any $\delta'>0$ with Lemma 8:

$$\mathbb{P}\left(\frac{1}{s_T^2}\sum_{t=1}^{T}\mathbb{E}\left[(\xi_t-\mathbb{E}(\xi_t|\mathcal{F}_{t-1}))^2\mathbf{1}_{|\xi_t|>\delta s_T}|\mathcal{F}_{t-1}\right]>\delta'\right)\le\frac{\epsilon(T+\log T+\epsilon)}{s_T^4\delta^2\delta'}\to 0 \text{ as } T\to\infty.$$

Lemma 10 *Let (X_t,Y_t), $t=1,\ldots,T$ i.i.d. with law μ defined in Lemma 8. Then, we have:*

$$\mathbb{P}(|\boldsymbol{c}_T^\star|_0>2)\le C\frac{\log T}{T}.$$

In the sequel, we introduce $\mathbf{c}_k^\star = \arg\min_{|\mathbf{c}|_0=k}(\sum_{t=1}^T (Y_t - g_{\mathbf{c}}(X_t))^2 + |\mathbf{c}|_0 \log T$ for $k=2$ and $k=3$. Then by construction of the i.i.d. sequence (X_t, Y_t), $t=1,\dots,T$ in Lemma 8, we have:

$$\sum_{t=1}^T (Y_t - g_{\mathbf{c}_2^\star}(X_t))^2 = \sum_{t\in S_{a,b}} (Y_t - \bar{Y}_{t-1}^{a,b})^2 + \sum_{t\in S_{c,d}} (Y_t - \bar{Y}_{t-1}^{c,d})^2,$$

where $S_{a,b}$ and $S_{c,d}$ denotes each $t=1,\dots,T$ such that $X_t \in \{a,b\}$ (and respectively $X_t \in \{c,d\}$) whereas $\bar{Y}_{t-1}^{a,b}$ and $\bar{Y}_{t-1}^{c,d}$ are defined in the proof of Lemma 8. Moreover, we have:

$$\sum_{t=1}^T (Y_t - g_{\mathbf{c}_3^\star}(X_t))^2 = \sum_{t\in S_{c,d}} (Y_t - \bar{Y}_{t-1}^{c,d})^2 + \sum_{t\in S_a} (Y_t - \bar{Y}_{t-1}^{a})^2 + \sum_{t\in S_b} (Y_t - \bar{Y}_{t-1}^{b})^2,$$

where S_a and S_b denotes each $t=1,\dots,T$ such that $X_t \in \{a\}$ (and respectively $X_t \in \{b\}$) whereas $\bar{Y}_t^a$ is defined in Lemma 8. Furthermore, a necessary condition to have $|\mathbf{c}_T^\star|_0 = 3$ is:

$$\sum_{t=1}^T (Y_t - g_{\mathbf{c}_2^\star}(X_t))^2 - \sum_{t=1}^T (Y_t - g_{\mathbf{c}_3^\star}(X_t))^2 \geq \log T.$$

Then, with the previous computations, gathering with a Markov inequality:

$$\begin{aligned}
\mathbb{P}\left(|\mathbf{c}_T^\star| = 3\right) &\leq \mathbb{P}\left(\sum_{x\in\{a,b\}} \sum_{t\in S_x} (2Y_t - \bar{Y}_{t-1}^{a,b} - \bar{Y}_{t-1}^{x})(\bar{Y}_{t-1}^{x} - \bar{Y}_{t-1}^{a,b}) \geq \log T\right) \\
&\leq \mathbb{P}\left(\sum_{x\in\{a,b\}} \sum_{t\in S_x} \mathbf{1}_{Y_t=1}(\bar{Y}_{t}^{x} - \bar{Y}_{t-1}^{a,b}) \geq \log T\right) \\
&\leq \mathbb{P}\left(\sum_{x\in\{a,b\}} \sum_{t\in S_x} \mathbf{1}_{Y_t=1}(\bar{Y}_{t}^{x} - \bar{Y}_{t-1}^{a,b}) \geq \log T\right) \\
&\leq \frac{\mathbb{E}\left[\sum_{x\in\{a,b\}} \sum_{t\in S_x} \mathbf{1}_{Y_t=1}(\bar{Y}_{t}^{x} - \bar{Y}_{t-1}^{a,b})\right]}{\log T} \\
&\leq \frac{\mathbb{E}\left[\sum_{x\in\{a,b\}} \sum_{t\in S_x} \mathbf{1}_{Y_t=1}(\bar{Y}_{t}^{x} - \bar{Y}_{t-1}^{a,b})\right]}{\log T} \sim \frac{\log T}{T},
\end{aligned}$$

where last line comes from the choice of $\epsilon \sim (\log T)^2/T$. A same argument shows that $\mathbb{P}(|\mathbf{c}_T^\star|_0) = 4$ is small, and then since by construction $|\mathbf{c}_T^\star|_0 \leq 4$ a.s., the proof is completed. ∎

References

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

[2] P. L. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inform. Theory*, 44(5):1802–1813, 1998.

[3] Brown B.M. Martingale Central Limit Theorems. *Ann. Math. Statist.*, 42(1):59–66, 1971.

[4] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R.E. Schapire, and M. Warmuth. How to use expert advice. *Journal for the Association of Computing Machinery*, 44 (3):427–485, 1997.

[5] N. Cesa-Bianchi and G. Lugosi. *Learning, Prediction and Games*. Cambridge University Press, 2006.

[6] Y. Cheng and G.M. Church. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 2000.

[7] A. S. Dalalyan and A.B. Tsybakov. Miror averaging with sparsity priors. *Bernoulli*, 18 (3):914–944, 2012.

[8] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14:729–769, 2013.

[9] D. Gnatyshak, D. Ignatov, A. Semenov, and J. Poelmans. Analysing online social network data with biclustering and triclustering. In *Proceedings of the Conference on Concept Discovery in Unstructured Data*, pages 30–39. Dmitry I. Ignatov, Sergei O. Kuznetsov, Jonas Poelmans (Eds.), 2012.

[10] D. Haussler, J. Kivinen, and M. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44 (5):1906–1925, 1998.

[11] I.M. Johnstone and B.W. Silverman. Empirical bayes selection of wavelet thresholds. *The Annals of Statistics*, 33:1700–1752, 2005.

[12] S. Kotz and S. Nadarajah. *Multivariate t distribution and their applications*. Cambridge University Press, 2004.

[13] S. Loustau. Online clustering of individual sequences. Submitted, 2014.

[14] D.A. Mac Allester. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234. ACM, 1998.

[15] V. Rivoirard. Nonlinear estimation over weak besov spaces and minimax bayes method. *Bernoulli*, 12:609–632, 2006.

[16] M.W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

[17] Y. Seldin. A pac-bayesian approach to structure learning. Phd Thesis, The Hebrew University of Jerusalem, 2009.

[18] Y. Seldin and N. Tishby. Pac-bayesian analysis of co-clustering and beyong. *Journal of Machine Learning Research*, 11:3595–3646, 2010.

[19] N. Slonim and N. Tishby. Document clustering using word clusters via the information botleneck method. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrievial*, 2000.

[20] M. Znidaric. Asymptotic expansion for inverse moments of binomial and poisson distributions. *The Open Statistics & Probability Journal*, 1:7–10, 2009.