Optimal Kullback–Leibler aggregation in mixture density estimation by maximum likelihood

Arnak S. Dalalyan   and   Mehdi Sebbar

Abstract. We study the maximum likelihood estimator of density of n independent observations, under the assumption that it is well approximated by a mixture with a large number of components. The main focus is on statistical properties with respect to the Kullback–Leibler loss. We establish risk bounds taking the form of sharp oracle inequalities both in deviation and in expectation. A simple consequence of these bounds is that the maximum likelihood estimator attains the optimal rate $((\log K)/n)^{1/2}$, up to a possible logarithmic correction, in the problem of convex aggregation when the number K of components is larger than $n^{1/2}$. More importantly, under the additional assumption that the Gram matrix of the components satisfies the compatibility condition, the obtained oracle inequalities yield the optimal rate in the sparsity scenario. That is, if the weight vector is (nearly) D-sparse, we get the rate $(D \log K)/n$. As a natural complement to our oracle inequalities, we introduce the notion of nearly-D-sparse aggregation and establish matching lower bounds for this type of aggregation.

1. Introduction

Assume that we observe n independent random vectors $X_1, \ldots, X_n \in \mathcal{X}$ drawn from a probability distribution P^* that admits a density function f^* with respect to some reference measure ν. The goal is to estimate the unknown density by a mixture density. More precisely, we assume that for a given family of mixture components $f_1, \ldots, f_K$, the unknown density of the observations f^* is well approximated by a convex combination $f_{\boldsymbol{\pi}}$ of these components, where

$$f_{\boldsymbol{\pi}}(\boldsymbol{x}) = \sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x}), \quad \boldsymbol{\pi} \in \mathbb{B}_+^K = \left\{ \boldsymbol{\pi} \in [0,1]^K : \sum_{j=1}^{K} \pi_j = 1 \right\}. \quad (1.1)$$

The assumption that the component densities $\mathcal{F} = \{f_j : j \in [K]\}$ are known essentially means that they are chosen from a dictionary obtained on the basis of previous experiments or expert knowledge.

We focus on the problem of estimation of the density function $f_{\boldsymbol{\pi}}$ and the weight vector $\boldsymbol{\pi}$ from the simplex $\mathbb{B}_+^K$ under the sparsity scenario: the ambient dimension K can be large, possibly larger than the sample size n, but most entries of $\boldsymbol{\pi}$ are either equal to zero or very small.

Our goal is to investigate the statistical properties of the Maximum Likelihood Estimator (MLE), defined by

$$\widehat{\boldsymbol{\pi}} \in \arg\min_{\boldsymbol{\pi} \in \Pi} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log f_{\boldsymbol{\pi}}(X_i) \right\}, \tag{1.2}$$

where the minimum is computed over a suitably chosen subset Π of $\mathbb{B}_+^K$. In the present work, we will consider sets $\Pi = \Pi_n(\mu)$, depending on a parameter $\mu > 0$ and the sample $\{X_1, \ldots, X_n\}$, defined by

$$\Pi_n(\mu) = \left\{ \boldsymbol{\pi} \in \mathbb{B}_+^K : \min_{i \in [n]} \sum_{j=1}^{K} \pi_j f_j(X_i) \geq \mu \right\}. \tag{1.3}$$

Note that the objective function in (1.2) is convex and the same is true for set (1.3). Therefore, the MLE $\widehat{\boldsymbol{\pi}}$ can be efficiently computed even for large K by solving a problem of convex programming. To ease notation, very often, we will omit the dependence of $\Pi_n(\mu)$ on μ and write Π_n instead of $\Pi_n(\mu)$.

The quality of an estimator $\widehat{\boldsymbol{\pi}}$ can be measured in various ways. For instance, one can consider the Kullback–Leibler divergence

$$\text{KL}(f^* || f_{\widehat{\boldsymbol{\pi}}}) = \begin{cases} \int_{\mathcal{X}} f^*(\boldsymbol{x}) \log \frac{f^*(\boldsymbol{x})}{f_{\widehat{\boldsymbol{\pi}}}(\boldsymbol{x})} \nu(d\boldsymbol{x}), & \text{if } P^*\big(f^*(X)/f_{\widehat{\boldsymbol{\pi}}}(X) = 0\big) = 0, \\ +\infty, & \text{otherwise,} \end{cases}$$
$$\tag{1.4}$$

which has the advantage of bypassing identifiability issues. One can also consider the (well-specified) setting where $f^* = f_{\boldsymbol{\pi}^*}$ for some $\boldsymbol{\pi}^* \in \mathbb{B}_+^K$ and measure the quality of estimation through a distance between the vectors $\widehat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}^*$ (such as the ℓ_1-norm $\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_1$ or the Euclidean norm $\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_2$).

The main contributions of the present work are the following:

(a) We demonstrate that in the mixture model there is no need to introduce sparsity favoring penalty in order to get optimal rates of estimation under the Kullback–Leibler loss in the sparsity scenario. In fact, the constraint that the weight vector belongs to the simplex acts as a sparsity inducing penalty. As a consequence, there is no need to tune a parameter accounting for the magnitude of the penalty.

(b) We show that the maximum likelihood estimator of the mixture density simultaneously attains the optimal rate of aggregation for the Kullback–Leibler loss for at least three types of aggregation: model-selection, convex and D-sparse aggregation.

(c) We introduce a new type of aggregation, termed *nearly D-sparse aggregation* that extends and unifies the notions of convex and D-sparse aggregation. We establish strong lower bounds for the nearly D-sparse aggregation and demonstrate that the maximum likelihood estimator attains this lower bound up to logarithmic factors.

1.1. Related work. The results developed in the present work aim to gain a better understanding (a) of the statistical properties of the maximum likelihood estimator over a high-dimensional simplex and (b) of the problem of aggregation of density estimators under the Kullback–Leibler loss. Various procedures of aggregation[1] for density estimation have been studied in the literature with respect to different loss functions. [8, 12, 33] investigated different variants of the progressive mixture rules, also known as mirror averaging [10, 34], with respect to the Kullback–Leibler loss and established model selection type oracle inequalities[2] in expectation. Same type of guarantees, but holding with high probability, were recently obtained in [1, 7] for the procedure termed Q-aggregation, introduced in other contexts by [9, 24].

Aggregation of estimators of a probability density function under the L_2-loss was considered in [26], where it was shown that a suitably chosen unbiased risk estimate minimizer is optimal both for convex and linear aggregation. The goal in the present work is to go beyond the settings of the aforementioned papers in that we want simultaneously to do as well as the best element of the dictionary, the best convex combination of the dictionary elements but also the best sparse convex combination. Note that the latter task was coined D-aggregation in [20] (see also [5]). In the present work, we rename it in D-sparse aggregation, in order to make explicit its relation to sparsity.

Key differences between the latter work and ours are that we do not assume the sparsity index to be known and we are analyzing an aggregation strategy that is computationally tractable even for large K. This is also the case of [3, 6], which are perhaps the most relevant references to the present work. These papers deal with the L_2-loss and investigate the lasso and the Dantzig estimators, respectively, suitably adapted to the problem of density estimation. Their methods handle dictionary elements $\{f_j\}$ which are not necessarily probability density functions, but has the drawback of requiring the choice of a tuning parameter. This choice is a nontrivial problem in practice. Instead, we show here that the optimal rates of sparse aggregation with respect to the Kullback–Leibler loss can be attained by procedure which is tuning parameter free.

Risk bounds for the maximum likelihood and other related estimators in the mixture model have a long history [18, 19, 21]. For the sake of comparison we recall here two elegant results providing non-asymptotic guarantees for the Kullback–Leibler loss.

---

[1] We refer the interested reader to [29] for an up to date introduction into aggregation of statistical procedures.

[2] This means that they prove that the expected loss of the aggregate is almost as small as the loss of the best element of the dictionary $\{f_1, \ldots, f_K\}$.

Theorem 1.1 ([18, Theorem 5.1]). *Let $\mathcal{F}$ be a finite dictionary of cardinality K of density functions such that $\max_{f \in \mathcal{F}} \|f^*/f\|_\infty \leq V$. Then, the maximum likelihood estimator over $\mathcal{F}$, $\widehat{f}_{\mathcal{F}}^{\mathrm{ML}} \in \arg\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f(X_i)$, satisfies the inequality*

$$\mathbf{E}_{f^*}\big[\, \mathrm{KL}\,\big(f^*\|\widehat{f}_{\mathcal{F}}^{\mathrm{ML}}\big)\big] \leq \big(2 + \log V\big)\Big(\min_{f \in \mathcal{F}} \mathrm{KL}(f^*\|f) + \frac{2\log K}{n}\Big). \qquad (1.5)$$

Inequality (1.5) is an inexact oracle inequality in expectation that quantifies the ability of $\widehat{f}_{\mathcal{F}}^{\mathrm{ML}}$ to solve the problem of model-selection aggregation. The adjective inexact refers to the fact that the "bias term" $\min_{f \in \mathcal{F}} \mathrm{KL}(f^*\|f)$ is multiplied by factor strictly larger than one. It is noteworthy that the remainder term $(2\log K)/n$ corresponds to the optimal rate of model-selection aggregation [11, 28]. In relation with Theorem 1.1, it is worth mentioning a result of [33] and [8], see also [14, Theorem 5] and [12, Corollary 5.4], establishing a risk bound similar to (1.5) without the extra factor $2 + \log V$ for the so called mirror averaging aggregate.

Theorem 1.2 ([21, p. 226]). *Let $\mathcal{F}$ be a finite dictionary of cardinality K of density functions and let $\mathcal{C}_k = \{f_{\boldsymbol{\pi}} : \|\boldsymbol{\pi}\|_0 \leq k\}$ be the set of all the mixtures of at most k elements of $\mathcal{F}$ ($k \in [K]$). Assume that f^* and the densities f_k from $\mathcal{F}$ are bounded from below and above by some positive constants m and M, respectively. Then, there is a constant C depending only on m and M such that, for any tolerance level $\delta \in (0,1)$, the maximum likelihood estimator over $\mathcal{C}_k$, $\widehat{f}_{\mathcal{C}_k}^{\mathrm{ML}} \in \arg\max_{f \in \mathcal{C}_k} \sum_{i=1}^{n} \log f(X_i)$, satisfies the inequality*

$$\mathrm{KL}\,\big(f^*\|\widehat{f}_{\mathcal{C}_k}^{\mathrm{ML}}\big) \leq \min_{f \in \mathcal{C}_k} \mathrm{KL}(f^*\|f) + C\Big(\frac{\log(K/\delta)}{n}\Big)^{1/2} \qquad (1.6)$$

with probability at least $1 - \delta$.

This result is remarkably elegant and can be seen as an exact oracle inequality in deviation for D-sparse aggregation (for $D = k$). Furthermore, if we choose $k = K$ in Theorem 1.2, then we get an exact oracle inequality for convex aggregation with a rate-optimal remainder term [28]. However, it fails to provide the optimal rate for D-sparse aggregation.

Closing this section, we would like to mention the recent work [32], where oracle inequalities for estimators of low rank density matrices are obtained. They share a common feature with those obtained in this work: the adaptation to the unknown sparsity or rank is achieved without any additional penalty term. The constraint that the unknown parameter belongs to the simplex acts as a sparsity inducing penalty.

1.2. Additional notation. In what follows, for any integer $i \in [n]$, we denote by $\boldsymbol{Z}_i$ the K-dimensional vector $[f_1(X_i), \ldots, f_K(X_i)]^\top$ and by $\mathbf{Z}$ the $n \times K$ matrix

$[\mathbf{Z}_1^\top, \ldots, \mathbf{Z}_n^\top]^\top$. We also define $\ell(u) = -\log u$, $u \in (0, +\infty)$, so that the MLE $\hat{\boldsymbol{\pi}}$ is the minimizer of the function

$$L_n(\boldsymbol{\pi}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{Z}_i^\top \boldsymbol{\pi}). \tag{1.7}$$

For any set of indices $J \subseteq [K]$ and any $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top \in \mathbb{R}^K$, we define $\boldsymbol{\pi}_J$ as the K-dimensional vector whose jth coordinate equals π_j if $j \in J$ and 0 otherwise. We denote the cardinality of any $J \subseteq [K]$ by $|J|$. For any set $J \subset \{1, \ldots, K\}$ and any constant $c \geq 0$, we introduce the compatibility constants [30] of a $K \times K$ positive semidefinite matrix $\mathbf{A}$,

$$\kappa_{\mathbf{A}}(J, c) = \inf \left\{ \frac{c^2 |J| \|\mathbf{A}^{1/2} \boldsymbol{v}\|_2^2}{(c \|\boldsymbol{v}_J\|_1 - \|\boldsymbol{v}_{J^c}\|_1)^2} : \boldsymbol{v} \in \mathbb{R}^K, \|\boldsymbol{v}_{J^c}\| < c \|\boldsymbol{v}_J\|_1 \right\}, \tag{1.8}$$

$$\overline{\kappa}_{\mathbf{A}}(J, c) = \inf \left\{ \frac{|J| \|\mathbf{A}^{1/2} \boldsymbol{v}\|_2^2}{\|\boldsymbol{v}_J\|_1^2} : \boldsymbol{v} \in \mathbb{R}^K, \|\boldsymbol{v}_{J^c}\|_1 < c \|\boldsymbol{v}_J\|_1 \right\}. \tag{1.9}$$

The risk bounds established in the present work involve the factors $\kappa_{\mathbf{A}}(J, 3)$ and $\overline{\kappa}_{\mathbf{A}}(J, 1)$. One can easily check that $\overline{\kappa}_{\mathbf{A}}(J, 3) \leq \kappa_{\mathbf{A}}(J, 3) \leq \frac{9}{4} \overline{\kappa}_{\mathbf{A}}(J, 1)$. We also recall that the compatibility constants of a matrix $\mathbf{A}$ are bounded from below by the smallest eigenvalue of $\mathbf{A}$.

Let us fix a function $f_0 \colon \mathcal{X} \to \mathbb{R}$ and denote $\overline{f}_k = f_k - f_0$ and

$$\overline{\mathbf{Z}}_i = [\overline{f}_1(X_i), \ldots, \overline{f}_K(X_i)]^\top, \tag{1.10}$$

for $i \in [n]$. In the results of this work, the compatibility factors are used for the empirical and population Gram matrices of vectors $\overline{\mathbf{Z}}_k$, that is when $\mathbf{A} = \hat{\boldsymbol{\Sigma}}_n$ and $\mathbf{A} = \boldsymbol{\Sigma}$ with

$$\hat{\boldsymbol{\Sigma}}_n = \frac{1}{n} \sum_{i=1}^{n} \overline{\mathbf{Z}}_i \overline{\mathbf{Z}}_i^\top, \quad \boldsymbol{\Sigma} = \mathbf{E}[\overline{\mathbf{Z}}_1 \overline{\mathbf{Z}}_1^\top]. \tag{1.11}$$

The general entries of these matrices are $(\hat{\boldsymbol{\Sigma}}_n)_{k,l} = 1/n \sum_{i=1}^{n} \overline{f}_k(X_i) \overline{f}_l(X_i)$ and $(\boldsymbol{\Sigma})_{k,l} = \mathbf{E}[\overline{f}_k(X_1) \overline{f}_l(X_1)]$, respectively. We assume that there exist positive constants m and M such that for all densities f_k with $k \in [K]$, we have

$$\forall x \in \mathcal{X}, \quad m \leq f_k(x) \leq M. \tag{1.12}$$

We use the notation $V = M/m$. It is worth mentioning that the set of dictionaries satisfying simultaneously this boundedness assumption and the aforementioned compatibility condition is not empty. For instance, one can consider the functions $f_k(x) = 1 + 1/2 \sin(2\pi k x)$ for $k \in [K]$. These functions are probability densities w.r.t. the Lebesgue measure on $\mathcal{X} = [0, 1]$. They are bounded from below and from above by $1/2$ and $3/2$, respectively. Taking $f_0(x) = 1$, the corresponding Gram matrix is $\boldsymbol{\Sigma} = 1/8 \, \mathbf{I}_K$, which has all eigenvalues equal to $1/8$.

1.3. Agenda. The rest of the paper is organized as follows. In Section 2, we state our main theoretical contributions and discuss their consequences. Possible relaxations of the conditions, as well as lower bounds showing the tightness of the established risk bounds, are considered in Section 3. A brief summary of the paper and some future directions of research are presented in Section 4. The proofs of all theoretical results are postponed to Section 5 and Section 6.

2. Oracle inequalities in deviation and in expectation

In this work, we prove several non-asymptotic risk bounds that imply, in particular, that the maximum likelihood estimator is optimal in model-selection aggregation, convex aggregation and D-sparse aggregation (up to log-factors). In all the results of this section we assume the parameter μ in (1.3) to be equal to 0.

Theorem 2.1. *Let F be a set of $K \geq 4$ densities satisfying the boundedness condition (1.12). Denote by $f_{\widehat{\boldsymbol{\pi}}}$ the mixture density corresponding to the maximum likelihood estimator $\widehat{\boldsymbol{\pi}}$ over Π_n defined in (1.7). There are constants $c_1 \leq 32V^3$, $c_2 \leq 288M^2V^6$ and $c_3 \leq 128M^2V^6$ such that, for any $\delta \in (0, 1/2)$, the following inequalities hold*

$$
\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{\substack{J \subset [K] \\ \boldsymbol{\pi} \in \mathbb{B}_+^K}} \left\{ \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) \right.
$$
$$
\left. + c_1 \Big(\frac{\log(K/\delta)}{n} \Big)^{1/2} \|\boldsymbol{\pi}_{J^c}\|_1 + \frac{c_2 |J| \log(K/\delta)}{n \kappa_{\widehat{\boldsymbol{\Sigma}}_n}(J, 3)} \right\},
\tag{2.1}
$$

$$
\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{J \subset [K]} \inf_{\substack{\boldsymbol{\pi} \in \mathbb{B}_+^K \\ \boldsymbol{\pi}_{J^c} = 0}} \left\{ \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + \frac{c_3 |J| \log(K/\delta)}{n \overline{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)} \right\}
\tag{2.2}
$$

with probability at least $1 - \delta$.

The proof of this and the subsequent results stated in this section are postponed to Section 5. Comparing the two inequalities of the above theorem, one can notice two differences. First, the term proportional to $\|\boldsymbol{\pi}_{J^c}\|_1$ is absent in the second risk bound, which means that the risk of the MLE is compared to that of the best mixture with a weight sequences supported by J. Hence, this risk bound is weaker than the first one provided by (2.1). Second, the compatibility factor $\overline{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)$ in (2.2) is larger that its counterpart $\kappa_{\widehat{\boldsymbol{\Sigma}}_n}(J, 3)$ in (2.1). This entails that in the cases where the oracle is expected to be sparse, the remainder term of the bound in (2.1) is slightly looser than that of (2.2).

A first and simple consequence of Theorem 2.1 is obtained by taking $J = \varnothing$ in the right hand side of the first inequality. Then, $\|\boldsymbol{\pi}_{J^c}\|_1 = \|\boldsymbol{\pi}\|_1 = 1$ and we get

$$\mathrm{KL}(f^*\|f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{\boldsymbol{\pi} \in \mathbb{B}_+^K} \mathrm{KL}(f^*\|f_{\boldsymbol{\pi}}) + c_1 \Big(\frac{\log(K/\delta)}{n} \Big)^{1/2}. \tag{2.3}$$

This implies that for every dictionary F, without any assumption on the smallness of the coherence between its elements, the maximum likelihood estimator achieves the optimal rate of convex aggregation, up to a possible[3] logarithmic correction, in the high-dimensional regime $K \geq n^{1/2}$. This essentially recovers the result of [21] recalled in Theorem 1.2 above. In the case of regression with random design, analogous results can be found in [16], [15] and the references therein.

The main compelling feature of our results is that they show that the MLE adaptively achieves the optimal rate of aggregation not only in the case of convex aggregation, but also for the model-selection aggregation and D-(convex) aggregation. For handling these two cases, it is more convenient to get rid of the presence of the compatibility factor of the empirical Gram matrix $\widehat{\boldsymbol{\Sigma}}_n$. The latter can be replaced by the compatibility factor of the population Gram matrix, as stated in the next result.

Theorem 2.2. *Let F be a set of K densities satisfying the boundedness condition (1.12). Denote by $f_{\widehat{\boldsymbol{\pi}}}$ the mixture density corresponding to the maximum likelihood estimator $\widehat{\boldsymbol{\pi}}$ over Π_n defined in (1.7). There are constants $c_4 \leq 32V^3 + 4$, $c_5 \leq 4.5M^2(8V^3 + 1)^2$ and $c_6 \leq 2M^2(8V^3 + 1)^2$ such that, for any $\delta \in (0, 1/2)$, the following inequalities hold*

$$\mathrm{KL}(f^*\|f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{\substack{J \subset [K] \\ \boldsymbol{\pi} \in \mathbb{B}_+^K}} \Bigg\{ \mathrm{KL}(f^*\|f_{\boldsymbol{\pi}}) \\ + c_4 \Big(\frac{\log(K/\delta)}{n} \Big)^{1/2} \|\boldsymbol{\pi}_{J^c}\|_1 + \frac{c_5|J|\log(K/\delta)}{n\kappa_{\boldsymbol{\Sigma}}(J,3)} \Bigg\}, \tag{2.4}$$

$$\mathrm{KL}(f^*\|f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{J \subset [K]} \inf_{\substack{\boldsymbol{\pi} \in \mathbb{B}_+^K \\ \boldsymbol{\pi}_{J^c}=0}} \Bigg\{ \mathrm{KL}(f^*\|f_{\boldsymbol{\pi}}) + \frac{c_6|J|\log(K/\delta)}{n\overline{\kappa}_{\boldsymbol{\Sigma}}(J,1)} \Bigg\} \tag{2.5}$$

with probability at least $1 - 2\delta$.

The main advantage of the upper bounds provided by Theorem 2.2 as compared with those of Theorem 2.1 is that the former is deterministic, whereas the latter involves the compatibility factor of the empirical Gram matrix which is random. The price to pay for getting rid of randomness in the risk bound is the increased values of the constants c_4, c_5 and c_6. Note, however, that this price is not too high, since

---

[3]In fact, the optimal rate of convex aggregation when $K \geq n^{1/2}$ is of order $\big(\log(K/n^{1/2})/n \big)^{1/2}$. Therefore, even the log K term is optimal whenever $K \geq Cn^{1/2+\alpha}$ for some $\alpha > 0$.

obviously $1 \leq M \leq L$ and, therefore, $c_4 \leq 1.25c_1$, $c_5 \leq 1.56c_2$ and $c_6 \leq 1.56c_3$. In addition, the absence of randomness in the risk bound allows us to integrate it and to convert the bound in deviation into a bound in expectation.

Theorem 2.3 (Bound in expectation). *Let F be a set of K densities satisfying the boundedness condition* (1.12). *Denote by $f_{\hat{\pi}}$ the mixture density corresponding to the maximum likelihood estimator $\hat{\pi}$ over Π_n defined in* (1.7). *There are constants $c_7 \leq 20V^3 + 8$, $c_8 \leq M^2(22V^3 + 3)^2$ and $c_9 \leq M^2(15V^3 + 2)^2$ such that*

$$\mathbf{E}[\mathrm{KL}(f^*||f_{\hat{\pi}})] \leq \inf_{\substack{J \subset [K] \\ \pi \in \mathbb{B}_+^K}} \left\{ \mathrm{KL}(f^*||f_{\pi}) + c_7 \Big(\frac{\log K}{n}\Big)^{1/2} \|\pi_{J^c}\|_1 + \frac{c_8|J|\log K}{n\kappa_{\Sigma}(J,3)} \right\},$$

$$(2.6)$$

$$\mathbf{E}[\mathrm{KL}(f^*||f_{\hat{\pi}})] \leq \inf_{J \subset [K]} \inf_{\substack{\pi \in \mathbb{B}_+^K \\ \pi_{J^c}=0}} \left\{ \mathrm{KL}(f^*||f_{\pi}) + \frac{c_9|J|\log K}{n\overline{\kappa}_{\Sigma}(J,1)} \right\}. \qquad (2.7)$$

In inequality (2.7), upper bounding the infimum over all sets J by the infimum over the singletons, we get

$$\mathbf{E}[\mathrm{KL}(f^*||f_{\hat{\pi}})] \leq \inf_{j \in [K]} \left\{ \mathrm{KL}(f^*||f_j) + \frac{c_9 \log K}{n\overline{\kappa}_{\Sigma}(J,1)} \right\}. \qquad (2.8)$$

This implies that the maximum likelihood estimator $f_{\hat{\pi}}$ achieves the rate $\frac{\log K}{n}$ in model-selection type aggregation. This rate is known to be optimal in the model of regression [24]. If we compare this result with Theorem 1.1 stated in Section 1.1, we see that the remainder terms of these two oracle inequalities are of the same order (provided that the compatibility factor is bounded away from zero), but inequality (2.8) has the advantage of being exact.

We can also apply (2.7) to the problem of convex aggregation with small dictionary, that is for K smaller than $n^{1/2}$. Upper bounding $|J|$ by K, we get

$$\mathbf{E}[\mathrm{KL}(f^*||f_{\hat{\pi}})] \leq \inf_{\pi \in \mathbb{B}_+^K} \mathrm{KL}(f^*||f_{\pi}) + \frac{c_9 K \log K}{n\overline{\kappa}_{\Sigma}([K],1)}. \qquad (2.9)$$

Assuming, for instance, the smallest eigenvalue of Σ bounded away from zero (which is a quite reasonable assumption in the context of low dimensionality), the above upper bound provides a rate of convex aggregation of the order of $\frac{K \log K}{n}$. Up to a logarithmic term, this rate is known to be optimal for convex aggregation in the model of regression.

Finally, considering all the sets J of cardinality smaller than D (with $D \leq K$) and setting $\overline{\kappa}_{\Sigma}(D,1) = \inf_{J:|J| \leq D} \overline{\kappa}_{\Sigma}(J,1)$, we deduce from (2.7) that

$$\mathbf{E}[\mathrm{KL}(f^*||f_{\hat{\pi}})] \leq \inf_{\pi \in \mathbb{B}_+^K : \|\pi\|_0 \leq D} \mathrm{KL}(f^*||f_{\pi}) + \frac{c_9 D \log K}{n\overline{\kappa}_{\Sigma}(D,1)}. \qquad (2.10)$$

According to [25, Theorem 5.3], in the regression model, the optimal rate of D-sparse aggregation is of order $(D/n)\log(K/D)$, whenever $D = o(n^{1/2})$. Inequality (2.10) shows that the maximum likelihood estimator over the simplex achieves this rate up to a logarithmic factor. Furthermore, this logarithmic inflation disappears when the sparsity D is such that, asymptotically, the ratio $\log D / \log K$ is bounded from above by a constant $\alpha < 1$. Indeed, in such a situation the optimal rate

$$\frac{D\log(K/D)}{n} = \frac{D\log K}{n}\left(1 - \frac{\log D}{\log K}\right)$$

is of the same order as the remainder term in (2.10), that is $(D\log K)/n$.

3. Discussion of the conditions and possible extensions

In this section, we start by announcing lower bounds for the Kullback–Leibler aggregation in the problem of density estimation. Then we discuss the implication of the risk bounds of the previous section to the case where the target is the weight vector $\boldsymbol{\pi}$ rather than the mixture density $f_{\boldsymbol{\pi}}$. Finally, we present some extensions to the case where the boundedness assumption is violated.

3.1. Lower bounds for nearly-D-sparse aggregation. As mentioned in previous section, the literature is replete with lower bounds on the minimax risk for various types of aggregation. However most of them concern the regression setting either with random or with deterministic design. Lower bounds of aggregation for density estimation were first established by [23] for the L_2-loss. In the case of Kullback–Leibler aggregation in density estimation, the only lower bounds we are aware are those established by [14] for model-selection type aggregation. It is worth emphasizing here that the results of the aforementioned two papers provide weak lower bounds. Indeed, they establish the existence of a dictionary for which the minimax excess risk is lower bounded by the suitable quantity. In contrast with this, we establish here strong lower bounds that hold for every dictionary satisfying the boundedness and the compatibility conditions.

Let $\mathcal{F} = \{f_1, \ldots, f_K\}$ be a dictionary of density functions on $\mathcal{X} = [0, 1]$. We say that the dictionary $\mathcal{F}$ satisfies the boundedness and the compatibility assumptions if for some positive constants m, M and κ, we have $m \leq f_j(x) \leq M$ for all $j \in [K]$, $x \in \mathcal{X}$. In addition, we assume in this subsection that all the eigenvalues of the Gram matrix $\boldsymbol{\Sigma}$ belong to the interval $[\varkappa_*, \varkappa^*]$, with $\varkappa_* > 0$ and $\varkappa^* < \infty$.

For every $\gamma \in (0, 1)$ and any $D \in [K]$, we define the set of nearly-D-sparse convex combinations of the dictionary elements $f_j \in \mathcal{F}$ by

$$\mathcal{H}_{\mathcal{F}}(\gamma, D) = \left\{ f_{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \mathbb{B}_+^K \text{ such that } \min_{J:|J|\leq D} \|\boldsymbol{\pi}_{J^c}\|_1 \leq \gamma \right\}. \qquad (3.1)$$

In simple words, $f_{\boldsymbol{\pi}}$ belongs to $\mathcal{H}_{\mathcal{F}}(\gamma, D)$ if it admits a γ-approximately D-sparse representation in the dictionary $\mathcal{F}$. We are interested in bounding from below the minimax excess risk

$$\mathcal{R}\big(\mathcal{H}_{\mathcal{F}}(\gamma, D)\big) = \inf_{\hat{f}} \sup_{f^*} \left\{ \mathbf{E}[\mathrm{KL}(f^* \| \hat{f})] - \inf_{f_{\boldsymbol{\pi}} \in \mathcal{H}_{\mathcal{F}}(\gamma, D)} \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}}) \right\}, \quad (3.2)$$

where the inf is over all possible estimators of f^* and the sup is over all density functions over $[0, 1]$. Note that the estimator $\hat{f}$ is not necessarily a convex combination of the dictionary elements. Furthermore, it is allowed to depend on the parameters γ and D characterizing the class $\mathcal{H}_{\mathcal{F}}(\gamma, D)$. It follows from (2.6), that if the dictionary satisfies the boundedness and the compatibility condition, then

$$\mathcal{R}\big(\mathcal{H}_{\mathcal{F}}(\gamma, D)\big) \leq C \left\{ \Big(\frac{\gamma^2 \log K}{n} \Big)^{1/2} + \frac{D \log K}{n} \right\} \bigwedge \Big(\frac{\log K}{n} \Big)^{1/2}, \quad (3.3)$$

for some constant C depending only on m, M and $\varkappa_*$. Note that the last term accounts for the following phenomenon: If the sparsity index D is larger than a multiple of $\sqrt{n}$, then the sparsity bears no advantage as compared to the ℓ_1 constraint. The next result implies that this upper bound is optimal, at least up to logarithmic factors.

Theorem 3.1. *Assume that* $\log(1 + eK) \leq n$. *Let* $\gamma \in (0, 1)$ *and* $D \in [K]$ *be fixed. There exists a constant* A *depending only on* m, M, $\varkappa_*$ *and* $\varkappa^*$ *such that* $\mathcal{R}(\mathcal{H}_{\mathcal{F}}(\gamma, D))$ *is larger than*

$$A \left\{ \Big[\frac{\gamma^2}{n} \log \Big(1 + \frac{K}{\gamma \sqrt{n}} \Big) \Big]^{1/2} + \frac{D \log(1 + K/D)}{n} \right\} \bigwedge \Big[\frac{1}{n} \log \Big(1 + \frac{K}{\sqrt{n}} \Big) \Big]^{1/2}. \quad (3.4)$$

This is the first result providing lower bounds on the minimax risk of aggregation over nearly-D-sparse aggregates. To the best of our knowledge, even in the Gaussian sequence model, such a result has not been established to date. It has the advantage of unifying the results on convex and D-sparse aggregation, as well as extending them to a more general class. Let us also stress that the condition $\log(1 + eK) \leq n$ is natural and unavoidable, since it ensures that the right hand side of (3.3) is smaller than the trivial bound $\log V$.

3.2. Weight vector estimation. The risk bounds carried out in the previous section for the problem of density estimation in the Kullback–Leibler loss imply risk bounds for the problem of weight vector estimation. Indeed, under the boundedness assumption (1.12), the Kullback–Leibler divergence between two mixture densities can be shown to be equivalent to the squared Mahalanobis distance between the weight vectors of these mixtures with respect to the Gram matrix. In order to go

from the Mahalanobis distance to the Euclidean one, we make use of the restricted eigenvalue

$$\kappa_{\boldsymbol{\Sigma}}^{\mathrm{RE}}(s, c) = \inf_{\boldsymbol{v} \in \Delta(s,c)} \|\boldsymbol{\Sigma}^{1/2} \boldsymbol{v}\|_2^2, \qquad (3.5)$$

with $\Delta(s, c) := \{\boldsymbol{v} : \exists\, J \subset [K] \text{ s.t. } |J| \leq s,\ \|\boldsymbol{v}_{J^c}\|_1 \leq c\|\boldsymbol{v}_J\|_1 \text{ and } \|\boldsymbol{v}_J\|_2 = 1\}$. This strategy leads to the next result.

Proposition 3.2. *Let F be a set of $K \geq 4$ densities satisfying condition (1.12). Denote by $f_{\widehat{\boldsymbol{\pi}}}$ the mixture density corresponding to the maximum likelihood estimator $\widehat{\boldsymbol{\pi}}$ over Π_n defined in (1.7). Let $\boldsymbol{\pi}^*$ the weight-vector of the best mixture density: $\boldsymbol{\pi}^* \in \arg\min_{\boldsymbol{\pi}} \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}})$, and let J^* be the support of $\boldsymbol{\pi}^*$. There are constants $c_{10} \leq M^2(64V^3 + 8)$ and $c_{11} \leq 4M^2(8V^3 + 1)$ such that, for any $\delta \in (0, 1/2)$, the following inequalities hold*

$$\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_1 \leq \frac{c_{10}|J^*|}{\bar{\kappa}_{\boldsymbol{\Sigma}}(J^*, 1)} \left(\frac{\log(K/\delta)}{n}\right)^{1/2}, \qquad (3.6)$$

$$\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_2 \leq \frac{c_{11}}{\kappa_{\boldsymbol{\Sigma}}^{\mathrm{RE}}(|J^*|, 1)} \left(\frac{2|J^*|\log(K/\delta)}{n}\right)^{1/2}, \qquad (3.7)$$

$$\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_2^2 \leq \frac{c_{11}}{\kappa_{\boldsymbol{\Sigma}}^{\mathrm{RE}}(|J^*|, 1)} \left(\frac{2\log(K/\delta)}{n}\right)^{1/2} \qquad (3.8)$$

with probability at least $1 - 2\delta$.

In simple words, this result tells us that the weight estimator $\widehat{\boldsymbol{\pi}}$ attains the minimax rate of estimation $|J^*|(\log(K)/n)^{1/2}$ over the intersection of the ℓ_1 and ℓ_0 balls, when the error is measured by the ℓ_1-norm, provided that the compatibility factor of the dictionary $\mathscr{F}$ is bounded away from zero. The optimality of this rate — up to logarithmic factors — follows from the fact that the error of estimation of each nonzero coefficients of $\boldsymbol{\pi}^*$ is at least $cn^{-1/2}$ (for some $c > 0$), leading to a sum of the absolute values of the errors at least of the order $|J^*|n^{-1/2}$. The logarithmic inflation of the rate is the price to pay for not knowing the support J^*. It is clear that this reasoning is valid only when the sparsity $|J^*|$ is of smaller order than $n^{1/2}$. Indeed, in the case $|J^*| \geq cn^{1/2}$, the trivial bound $\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_1 \leq 2$ is tighter than the one in (3.6).

Concerning the risk measured by the Euclidean norm, we underline that there are two regimes characterized by the order between upper bounds in (3.7) and (3.8). Roughly speaking, when the signal is highly sparse in the sense that $|J^*|$ is smaller than $(n/\log K)^{1/2}$, then the smallest bound is given by (3.7) and is of the order $(|J^*|\log(K))/n$. This rate is can be compared to the rate $(|J^*|\log(K/|J^*|))/n$, known to be optimal in the Gaussian sequence model. In the second regime corresponding to mild sparsity, $|J^*| > (n/\log K)^{1/2}$, the smallest bound is the one in (3.8). The latter is of order $(\log(K)/n)^{1/2}$, which is known to be optimal in the Gaussian sequence model. For various results providing lower bounds in regression framework we refer the interested reader to [22, 25, 31].

3.3. Extensions to the case of vanishing components. In the previous sections we have deliberately avoided any discussion of the role of the parameter μ, present in the search space $\Pi_n(\mu)$ of the problem (1.2)–(1.3). In fact, when all the dictionary elements are separated from zero by a constant m, a condition assumed throughout previous sections, choosing any value of $\mu \leq m$ is equivalent to choosing $\mu = 0$. Therefore, the choice of this parameter does not impact the quality of estimation. However, this parameter might have strong influence in practice both on statistical and computational complexity of the maximum likelihood estimator. A first step in understanding the influence of μ on the statistical complexity is made in the next paragraphs.

Let us consider the case where the condition $\min_x \min_j f_j(x) \geq m > 0$ fails, but the upper-boundedness condition $\max_x \max_j f_j(x) \leq M$ holds true. In such a situation, we replace the definition $V = M/m$ by $V = M/\mu$. We also define the set $\Pi^*(\mu) = \{\boldsymbol{\pi} \in \mathbb{B}_+^K : P^*(f_{\boldsymbol{\pi}}(X) \geq \mu) = 1\}$. In order to keep mathematical formulae simple, we will only state the equivalent of (2.2) in the case of $m = 0$. All the other results of the previous section can be extended in a similar way.

Proposition 3.3. *Let F be a set of $K \geq 2$ densities satisfying the boundedness condition $\sup_{\boldsymbol{x} \in \mathcal{X}} f_j(\boldsymbol{x}) \leq M$. Denote by $f_{\widehat{\boldsymbol{\pi}}}$ the mixture density corresponding to the maximum likelihood estimator $\widehat{\boldsymbol{\pi}}$ over $\Pi_n(\mu)$ defined in (1.7). There is a constant $\bar{c} \leq 128M^2V^4$ such that, for any $\delta \in (0, 1/2)$,*

$$\mathrm{KL}(f^*\|f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{J \subset [K]} \inf_{\substack{\boldsymbol{\pi} \in \Pi^*(\mu) \\ \boldsymbol{\pi}_{J^c} = 0}} \left\{ \mathrm{KL}(f^*\|f_{\boldsymbol{\pi}}) + \frac{\bar{c}|J|\log(K/\delta)}{n\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)} \right\}$$
$$+ \int_{\mathcal{X}} (\log \mu - \log f_{\widehat{\boldsymbol{\pi}}})_+ f^* d\nu \quad (3.9)$$

on an event of probability at least $1 - \delta$. Furthermore, if $\inf_{\boldsymbol{x} \in \mathcal{X}} f^*(\boldsymbol{x}) \geq \mu$, then, on the same event, we have

$$\|f^* - f_{\widehat{\boldsymbol{\pi}}}\|_{L^2(P^*)}^2 \leq 2M^2 \inf_{J \subset [K]} \inf_{\substack{\boldsymbol{\pi} \in \Pi^*(\mu) \\ \boldsymbol{\pi}_{J^c} = 0}} \left\{ \mathrm{KL}(f^*\|f_{\boldsymbol{\pi}}) + \frac{\bar{c}|J|\log(K/\delta)}{n\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)} \right\}. \quad (3.10)$$

The last term present in the first upper bound,

$$\int_{\mathcal{X}} (\log \mu - \log f_{\widehat{\boldsymbol{\pi}}})_+ f^* d\nu$$

is the price we pay for considering densities that are not lower bounded by a given constant. A simple, non-random upper bound on this term is

$$\int_{\mathcal{X}} \max_{k \in [K]} (\log \mu - \log f_k)_+ f^* d\nu.$$

Providing a tight upper bound on this kind or remainder terms is an important problem which lies beyond the scope of the present work.

4. Conclusion

In this paper, we have established exact oracle inequalities for the maximum likelihood estimator of a mixture density. This oracle inequality clearly highlights the interplay of three sources of error: misspecification of the model of mixture, departure from D-sparsity and stochastic error of estimating D nonzero coefficients. We have also proved a lower bound that show that the remainder terms of our upper bounds are optimal, up to logarithmic terms. This lower bound is valid not only for the maximum likelihood estimator, but for any estimator of the density function. As a consequence, the maximum likelihood estimator has a nearly optimal excess risk in the minimax sense.

In all the results of the present paper, we have assumed that the components of the mixture model are deterministic. From a practical point of view, it might be reasonable to choose these components in a data driven way, using, for instance, a hold-out sample. This question, as well as the problem of tuning the parameter μ, constitute interesting and challenging avenues for future research.

5. Proofs of results stated in previous sections

This section collects the proofs of the theorems and claims stated in previous sections.

5.1. Proof of Theorem 2.1. The main technical ingredients of the proof are a strong convexity argument and a control of the maximum of an empirical process. The corresponding results are stated in Lemma 5.2 and Proposition 5.3, respectively, deferred to Section 5.6. We denote by $\bar{\mathbf{Z}}$ the $n \times K$ matrix $[\bar{\mathbf{Z}}_1, \ldots, \bar{\mathbf{Z}}_K]$.

Since $\hat{\boldsymbol{\pi}}$ is a minimizer of $L_n(\cdot)$, see (1.2) and (1.7), we know that $L_n(\hat{\boldsymbol{\pi}}) \leq L_n(\boldsymbol{\pi})$ for every π. However, this inequality can be made sharper using the (local) strong convexity of the function $\ell(u) = -\log(u)$. Indeed, Lemma 5.2 below shows that

$$\frac{1}{n} \sum_{i=1}^{n} \ell\big(f_{\hat{\boldsymbol{\pi}}}(X_i)\big) \leq \frac{1}{n} \sum_{i=1}^{n} \ell\big(f_{\boldsymbol{\pi}}(X_i)\big) - \frac{1}{2M^2 n} \|\bar{\mathbf{Z}}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2. \qquad (5.1)$$

On the other hand, if we set $\varphi(\pi, \boldsymbol{x}) = \int (\log f_{\boldsymbol{\pi}}) f^* d\nu - \log f_{\boldsymbol{\pi}}(\boldsymbol{x})$, we have $\mathbf{E}_{f^*}[\varphi(\boldsymbol{\pi}, X_i)] = 0$ and

$$\ell\big(f_{\boldsymbol{\pi}}(X_i)\big) = \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}}) - \int_{\mathcal{X}} f^* \log f^* d\nu + \varphi(\pi, X_i). \qquad (5.2)$$

Combining inequalities (5.1) and (5.2), we get

$$\mathrm{KL}(f^* \| f_{\hat{\boldsymbol{\pi}}}) \leq \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}}) - \frac{1}{2M^2 n} \|\bar{\mathbf{Z}}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \big(\varphi(\boldsymbol{\pi}, X_i) - \varphi(\hat{\boldsymbol{\pi}}, X_i)\big).$$

$$(5.3)$$

The next step of the proof consists in establishing a suitable upper bound on the noise term $\Phi_n(\boldsymbol{\pi}) - \Phi_n(\widehat{\boldsymbol{\pi}})$ where

$$\Phi_n(\boldsymbol{\pi}) = \frac{1}{n} \sum_{i=1}^{n} \varphi(\boldsymbol{\pi}, X_i). \tag{5.4}$$

According to the mean value theorem, setting $\zeta_n := \sup_{\overline{\boldsymbol{\pi}} \in \boldsymbol{\Pi}_n} \left\| \nabla \Phi_n(\overline{\boldsymbol{\pi}}) \right\|_\infty$, for every vector $\boldsymbol{\pi} \in \boldsymbol{\Pi}_n$, it holds that

$$|\Phi_n(\widehat{\boldsymbol{\pi}}) - \Phi_n(\boldsymbol{\pi})| \le \sup_{\overline{\boldsymbol{\pi}} \in \boldsymbol{\Pi}_n} \left\| \nabla \Phi_n(\overline{\boldsymbol{\pi}}) \right\|_\infty \|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1 = \zeta_n \|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1. \tag{5.5}$$

This inequality, combined with (5.3), yields

$$\mathrm{KL}(f^* \| f_{\widehat{\boldsymbol{\pi}}}) \le \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}}) - \frac{1}{2M^2 n} \|\overline{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 + \zeta_n \|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1. \tag{5.6}$$

Using the Gram matrix $\widehat{\boldsymbol{\Sigma}}_n = 1/n \overline{\mathbf{Z}}^\top \overline{\mathbf{Z}}$, the quantity $\|\overline{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2$ can be rewritten as

$$\|\overline{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 = n \|\widehat{\boldsymbol{\Sigma}}_n^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2. \tag{5.7}$$

We proceed with applying the following result.

Lemma 5.1 ([2, Lemma 2]). *For any pair of vectors $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \mathbb{R}^K$, for any pair of scalars $\theta > 0$ and $\varrho > 1$, for any $K \times K$ symmetric matrix $\mathbf{A}$ and for any set $J \subset [p]$, the following inequality is true*

$$2\theta \varrho^{-1} \left(\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_1 + \varrho \|\boldsymbol{\pi}\|_1 - \varrho \|\boldsymbol{\pi}'\|_1 \right) - \|\mathbf{A}(\boldsymbol{\pi} - \boldsymbol{\pi}')\|_2^2$$
$$\le 4\theta \|\boldsymbol{\pi}_{J^c}\|_1 + \frac{(\varrho + 1)^2 \theta^2 |J|}{\varrho^2 \kappa_{\mathbf{A}^2}(J, c_\varrho)}, \quad (5.8)$$

where $c_\varrho = (\varrho + 1)/(\varrho - 1)$.

Choosing $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_n^{1/2}/(\sqrt{2}\,M)$, $\theta = \zeta_n$ and $\varrho = 2$ (thus $c_\varrho = 3$) we get the inequality

$$\zeta_n \|\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}}\|_1 - \|\mathbf{A}(\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}})\|_2^2 \le 4\zeta_n \|\boldsymbol{\pi}_{J^c}\|_1 + \frac{9\zeta_n^2 |J|}{4\kappa_{\mathbf{A}^2}(J, 3)}, \quad \forall J \in \{1, \dots, p\}. \tag{5.9}$$

One can check that $\kappa_{\mathbf{A}^2}(J, 3) = \kappa_{\widehat{\boldsymbol{\Sigma}}_n}(J, 3)/(2M^2)$. Combining the last inequality with (5.6), we arrive at

$$\mathrm{KL}(f^* \| f_{\widehat{\boldsymbol{\pi}}}) \le \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}}) + 4\zeta_n \|\boldsymbol{\pi}_{J^c}\|_1 + \frac{9M^2 \zeta_n^2 |J|}{2\kappa_{\widehat{\boldsymbol{\Sigma}}_n}(J, 3)}. \tag{5.10}$$

Since the last inequality holds for every $\boldsymbol{\pi}$, we can insert an $\inf_{\boldsymbol{\pi}}$ in the right hand side. Furthermore, in view of Proposition 5.3 below, with probability larger than $1 - \delta$, ζ_n is bounded from above by $8V^3 (\log(K/\delta)/n)^{1/2}$. This completes the proof of (2.1).

To prove (2.2), we follow the same steps as above up to inequality (5.6). Then, we remark that for every $\boldsymbol{\pi}$ in the simplex satisfying $\boldsymbol{\pi}_{J^c} = 0$, it holds

$$\|(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})_{J^c}\|_1 = \|\widehat{\boldsymbol{\pi}}_{J^c}\|_1 = 1 - \|\widehat{\boldsymbol{\pi}}_J\|_1 = \|\boldsymbol{\pi}_J\|_1 - \|\widehat{\boldsymbol{\pi}}_J\|_1 \leq \|(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})_J\|_1. \quad (5.11)$$

Therefore,

$$\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 \geq \frac{\overline{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)\|(\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}})_J\|_1^2}{|J|},$$

we have with probability at least $1 - \delta$

$$\zeta_n\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1 - \frac{1}{2M^2 n}\|\mathbf{Z}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 \leq 2\zeta_n\|(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})_J\|_1 - \frac{1}{2M^2}\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2$$

$$\leq 2\zeta_n\|(\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}})_J\|_1 - \frac{\overline{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)\|(\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}})_J\|_1^2}{2M^2|J|}$$

$$\leq \frac{2\zeta_n^2 M^2 |J|}{\overline{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)}. \quad (5.12)$$

Replacing the right hand term in (5.6) and taking the infimum, we get the claim of the corollary. Since, in view of Proposition 5.3 below, with probability larger than $1 - \delta$, ζ_n is bounded from above by $8V^3(\log(K/\delta)/n)^{1/2}$, we get the claim of (2.2).

5.2. Proof of Theorem 2.2. Let us denote $\boldsymbol{v} = \widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}$. According to (5.6) and (5.7), we have

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + \zeta_n\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1 - \frac{1}{2M^2}\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 \quad (5.13)$$

$$\leq \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + \zeta_n\|\boldsymbol{v}\|_1 \quad (5.14)$$

$$- \frac{1}{2M^2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|_2^2 + \frac{1}{2M^2}\boldsymbol{v}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n)\boldsymbol{v}.$$

As $\boldsymbol{v}$ is the difference of two vectors lying on the simplex, we have $\|\boldsymbol{v}\|_1 \leq 2$. Let $\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_\infty = \max_{j,j'} |(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n)_{j,j'}|$ stand for the largest (in absolute values) element of the matrix $\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n$. We have

$$\boldsymbol{v}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n)\boldsymbol{v} \leq \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_\infty\|\boldsymbol{v}\|_1^2 \leq 2\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_\infty\|\boldsymbol{v}\|_1. \quad (5.15)$$

Setting $\overline{\zeta}_n = \zeta_n + M^{-2}\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_\infty$, we get

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + \overline{\zeta}_n\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1 - \frac{1}{2M^2}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2. \quad (5.16)$$

Following the same steps as those used for obtaining (5.10), we arrive at

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + 4\overline{\zeta}_n\|\boldsymbol{\pi}_{J^c}\|_1 + \frac{9\overline{\zeta}_n^2 M^2 |J|}{2\kappa_{\boldsymbol{\Sigma}}(J, 3)}. \quad (5.17)$$

The last step consists in evaluating the quantiles of the random variable $\bar{\zeta}_n$. To this end, one checks that the Hoeffding inequality combined with the union bound yields

$$\mathbf{P}\{\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_\infty > t\} \leq K(K-1)\exp\left(-2nt^2/M^4\right), \quad \forall t > 0. \qquad (5.18)$$

In other terms, for every $\delta \in (0, 1)$, we have

$$\mathbf{P}\left\{\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_\infty \leq M^2\left(\frac{\log(K^2/\delta)}{2n}\right)^{1/2}\right\} \geq 1 - \delta. \qquad (5.19)$$

Note that for $\delta \leq 1$, we have $\log(K^2/\delta) \leq 2\log(K/\delta)$. Combining with Proposition 5.3, this implies that $\bar{\zeta}_n \leq (8V^3 + 1)(\log(K/\delta)/n)^{1/2}$ with probability larger than $1 - 2\delta$. In view of (5.17), this completes the proof of (2.4). The proof of (2.5) is omitted since it repeats the same arguments as those used for proving (2.2).

5.3. Proof of Theorem 2.3. According to (5.17), for any $\pi \in \Pi$ and any $J \subset \{1, \dots, K\}$, we have

$$\mathbf{E}[\mathrm{KL}(f^*\|f_{\hat{\pi}})] \leq \mathrm{KL}(f^*\|f_\pi) + 4\|\pi_{J^c}\|_1\mathbf{E}[\bar{\zeta}_n] + \frac{9M^2|J|}{2\kappa_\mathbf{\Sigma}(J,3)}\mathbf{E}[\bar{\zeta}_n^2]. \qquad (5.20)$$

Recall now that $\bar{\zeta}_n = \zeta_n + M^{-2}\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_\infty$ and, according to Proposition 5.3, we have

$$\mathbf{E}[\zeta_n] \leq 4V^3\left(\frac{2\log(2K^2)}{n}\right)^{1/2} \quad \text{and} \quad \mathbf{Var}[\zeta_n] \leq \frac{V^2}{2n}. \qquad (5.21)$$

Using Theorem A.1, one easily checks that

$$\mathbf{E}[\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_\infty] \leq M^2\left(\frac{\log(2K^2)}{2n}\right)^{1/2}. \qquad (5.22)$$

This implies that

$$\mathbf{E}[\bar{\zeta}_n] \leq (8V^3 + 1)\left(\frac{\log(2K^2)}{2n}\right)^{1/2}. \qquad (5.23)$$

Similarly, in view of the Efron–Stein inequality, we have $\mathbf{Var}[\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_\infty] \leq \frac{M^4}{2n}$. This implies that

$$\mathbf{E}[\bar{\zeta}_n^2] \leq (\mathbf{E}[\bar{\zeta}_n])^2 + \left\{\left(\mathbf{Var}[\zeta_n]\right)^{1/2} + M^{-2}\left(\mathbf{Var}[\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_\infty]\right)^{1/2}\right\}^2 \qquad (5.24)$$

$$\leq (8V^3 + 1)^2\frac{\log(2K^2)}{2n} + \frac{(V+1)^2}{2n} \qquad (5.25)$$

$$\leq 1.615(8V^3 + 1)^2\frac{\log K}{n}. \qquad (5.26)$$

Combining (5.23), (5.26) and (5.20), we get the desired result.

5.4. Proof of Proposition 3.2. Using the strong convexity of the function $u \mapsto -\log u$ over the interval $[m, M]$ and the fact that $\boldsymbol{\pi}^*$ minimizes the convex function $\boldsymbol{\pi} \mapsto \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}})$, we get

$$\mathrm{KL}(f^* \| f_{\hat{\boldsymbol{\pi}}}) \geq \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}^*}) + \frac{1}{2M^2} \| \hat{\boldsymbol{\Sigma}}_n^{1/2} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*) \|_2^2. \tag{5.27}$$

Combining with (5.16), in which we replace $\boldsymbol{\pi}$ by $\boldsymbol{\pi}^*$, we get

$$\| \boldsymbol{\Sigma}^{1/2} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*) \|_2^2 \leq 2M^2 \overline{\zeta}_n \| \hat{\boldsymbol{\pi}} - \boldsymbol{\pi}^* \|_1. \tag{5.28}$$

Let us set $\boldsymbol{v} = \hat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*$. If $\boldsymbol{v} = 0$, then the claims are trivial. In the rest of this proof, we assume $\|\boldsymbol{v}\|_1 > 0$. In view of (5.11), we have $\|\boldsymbol{v}\|_1 \leq 2\|\boldsymbol{v}_{J^*}\|_1$. Therefore, using the definition of the compatibility factor, we get

$$\|\boldsymbol{v}\|_1^2 \leq 4 \|\boldsymbol{v}_{J^*}\|_1^2 \leq \frac{4|J^*| \, \| \boldsymbol{\Sigma}^{1/2} \boldsymbol{v} \|_2^2}{\overline{\kappa}(J^*, 1)} \leq \frac{8|J^*| \, M^2 \overline{\zeta}_n \|\boldsymbol{v}\|_1}{\overline{\kappa}(J^*, 1)}. \tag{5.29}$$

We have already checked that $\overline{\zeta}_n \leq (8V^3 + 1)(\log(K/\delta)/n)^{1/2}$ with probability larger than $1 - 2\delta$. Dividing both sides of inequality (5.29) by $\|\boldsymbol{v}\|_1$ and using the aforementioned upper bound on $\overline{\zeta}_n$, we get the desired bound on $\|\boldsymbol{v}\|_1 = \|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_1$.

In order to bound the error $\boldsymbol{v} = \hat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*$ in the Euclidean norm, we denote by $\hat{J}$ the set of $D = |J^*|$ indices corresponding to D largest entries of the vector $(|v_1|, \ldots, |v_K|)$. Since $\|\boldsymbol{v}\|_1 \leq 2\|\boldsymbol{v}_{J^*}\|_1$, we clearly have $\|\boldsymbol{v}\|_1 \leq 2\|\boldsymbol{v}_{\hat{J}}\|_1$. Therefore,

$$\|\boldsymbol{v}\|_2^2 = \|\boldsymbol{v}_{\hat{J}}\|_2^2 + \|\boldsymbol{v}_{\hat{J}^c}\|_2^2 \tag{5.30}$$

$$\leq \|\boldsymbol{v}_{\hat{J}}\|_2^2 + \|\boldsymbol{v}_{\hat{J}^c}\|_\infty \|\boldsymbol{v}_{\hat{J}^c}\|_1 \tag{5.31}$$

$$\leq \|\boldsymbol{v}_{\hat{J}}\|_2^2 + \frac{\|\boldsymbol{v}_{\hat{J}}\|_1}{D} \|\boldsymbol{v}_{\hat{J}^c}\|_1 \tag{5.32}$$

$$\leq \|\boldsymbol{v}_{\hat{J}}\|_2^2 + \frac{1}{D} \|\boldsymbol{v}_{\hat{J}}\|_1^2 \leq 2\|\boldsymbol{v}_{\hat{J}}\|_2^2. \tag{5.33}$$

Combining this inequality with the definition of the restricted eigenvalue and inequality (5.28) above, we arrive at

$$\|\boldsymbol{v}_{\hat{J}}\|_2^2 \leq \frac{\| \boldsymbol{\Sigma}^{1/2} \boldsymbol{v} \|_2^2}{\kappa^{\mathrm{RE}}(D, 1)} \leq \frac{2M^2 \overline{\zeta}_n \|\boldsymbol{v}\|_1}{\kappa^{\mathrm{RE}}(D, 1)} \tag{5.34}$$

$$\leq \frac{4M^2 \overline{\zeta}_n (\|\boldsymbol{v}_{\hat{J}}\|_1 \wedge 1)}{\kappa^{\mathrm{RE}}(D, 1)} \leq \frac{4M^2 \overline{\zeta}_n (\sqrt{D} \|\boldsymbol{v}_{\hat{J}}\|_2 \wedge 1)}{\kappa^{\mathrm{RE}}(D, 1)}. \tag{5.35}$$

Dividing both sides by $\|\boldsymbol{v}_{\hat{J}}\|_2$, taking the square and using (5.33), we get

$$\|\boldsymbol{v}\|_2 \leq \sqrt{2} \|\boldsymbol{v}_{\hat{J}}\|_2 \leq \frac{4\sqrt{2}M^2 |J^*|^{1/2} \overline{\zeta}_n}{\kappa^{\mathrm{RE}}(|J^*|, 1)} \bigwedge \frac{2\sqrt{2}M \overline{\zeta}_n^{1/2}}{\kappa^{\mathrm{RE}}(|J^*|, 1)^{1/2}}. \tag{5.36}$$

This inequality, in conjunction with the upper bound on $\overline{\zeta}_n$ used above, completes the proof of the second claim.
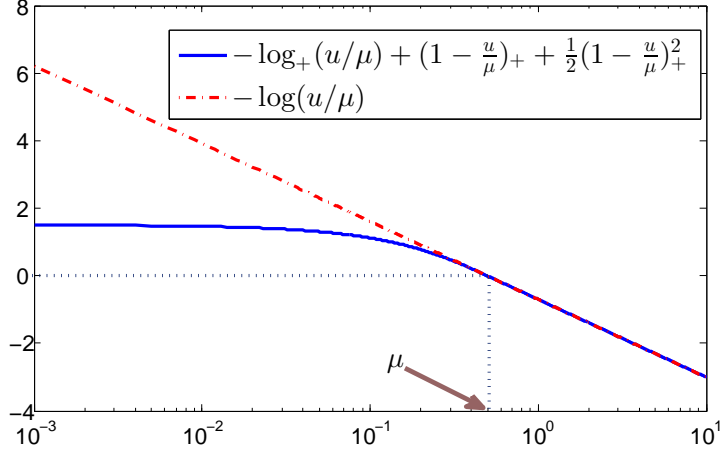
Figure 1. The plot of the function $u \mapsto \bar{\ell}(u)$, used in the proof of Proposition 3.3, superposed on the plot of the function $u \mapsto \ell(u) = -\log u$. We see that the former is a strongly convex surrogate of the latter.

5.5. Proof of Proposition 3.3. We repeat the proof of Theorem 2.1 with some small modifications. First of all, we replace the function $\ell(u) = -\log(u)$ by the function

$$\bar{\ell}(u) = \begin{cases} -\log(u/\mu), & \text{if } u \geq \mu, \\ (1 - \frac{u}{\mu}) + \frac{1}{2}(1 - \frac{u}{\mu})^2, & \text{if } u \in (0, \mu). \end{cases} \tag{5.37}$$

One easily checks that this function is twice continuously differentiable with a second derivative satisfying $M^{-2} \leq \bar{\ell}''(u) \leq \mu^{-2}$ for every $u \in (0, M)$. Furthermore, since $\bar{\ell}(u) = \ell(u/\mu)$ for every $u \geq \mu$, we have $\bar{L}_n(\hat{\pi}) = L_n(\hat{\pi})$, where we have used the notation $\bar{L}_n(\pi) = \frac{1}{n} \sum_{i=1}^{n} \bar{\ell}(f_\pi(X_i))$. Therefore, similarly to (5.1), we get

$$\frac{1}{n} \sum_{i=1}^{n} \bar{\ell}(f_{\hat{\pi}}(X_i)) \leq \frac{1}{n} \sum_{i=1}^{n} \bar{\ell}(f_\pi(X_i)) - \frac{1}{2M^2 n} \|\bar{\mathbf{Z}}(\hat{\pi} - \pi)\|_2^2, \tag{5.38}$$

for every $\pi \in \Pi^*(\mu)$. Let us define $\bar{\varphi}(\pi, x) = \bar{\ell}(f_\pi(x)) - \int \bar{\ell}(f_\pi) f^* d\nu$ and $\bar{\Phi}_n(\pi) = \frac{1}{n} \sum_{i=1}^{n} \bar{\varphi}(\pi, X_i)$. We have

$$\int \bar{\ell}(f_{\hat{\pi}}) f^* d\nu$$

$$\leq \int \bar{\ell}(f_\pi) f^* d\nu - \frac{1}{2M^2 n} \|\bar{\mathbf{Z}}(\hat{\pi} - \pi)\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} (\varphi(\pi, X_i) - \varphi(\hat{\pi}, X_i))$$

$$\tag{5.39}$$

$$\leq \int \bar{\ell}(f_{\pi}) \, f^* d\nu - \frac{1}{2M^2 n} \|\bar{\mathbf{Z}}(\hat{\pi} - \pi)\|_2^2 + \underbrace{\sup_{\pi \in \Pi_n(0)} \|\nabla \bar{\Phi}_n(\pi)\|_{\infty} \, \|\hat{\pi} - \pi\|_1}_{:= \xi_n}.$$

$$(5.40)$$

Notice that $\pi \in \Pi^*(\mu)$ implies that $\bar{\ell}(f_{\pi}) = \log \mu - \log f_{\pi}$ and that $\bar{\ell}(f_{\hat{\pi}}) \geq \log \mu - \log f_{\hat{\pi}} - (\log \mu - \log f_{\hat{\pi}})_+$. Therefore, along the lines of the proof of (2.2) (see, namely, (5.12)), we get

$$\mathrm{KL}(f^* \| f_{\hat{\pi}}) \leq \mathrm{KL}(f^* \| f_{\pi}) + \frac{2\xi_n^2 M^2 |J|}{\bar{\kappa}_{\hat{\Sigma}_n}(J, 1)} + \int_{\mathcal{X}} (\log \mu - \log f_{\hat{\pi}})_+ f^* d\nu. \quad (5.41)$$

We can repeat now the arguments of Proposition 5.3 with some minor modifications. First of all, we rewrite ξ_n as $\xi_n = \max_{l=1,\dots,K} \xi_{l,n}$ with $\xi_{l,n} = \sup_{\pi \in \Pi_n(0)} |\partial_l \bar{\Phi}_n(\pi)|$. One checks that the bounded difference inequality and the Efron–Stein inequality can be applied with an additional factor 2, since for $F_l(\mathbf{X}) = \sup_{\pi \in \Pi_n(0)} |\partial_l \bar{\Phi}_n(\pi)|$, we have

$$|F_l(\mathbf{X}) - F_l(\mathbf{X}')| \leq \frac{2M}{n\mu} = \frac{2V}{n}. \quad (5.42)$$

Therefore, for every $l \in [K]$, with probability larger than $1 - (\delta/K)$, we have

$$\xi_{l,n} \leq \mathbf{E}[\xi_{l,n}] + V \left(\frac{2 \log(K/\delta)}{n} \right)^{1/2} \quad \text{and} \quad \mathbf{Var}[\xi_n] \leq \frac{(2V)^2}{n}.$$

By the union bound, we obtain that with probability larger than $1 - \delta$,

$$\xi_n \leq \max_l \mathbf{E}[\xi_{l,n}] + V \left(\frac{2 \log(K/\delta)}{n} \right)^{1/2}.$$

Thus, to upper bound $\mathbf{E}[\xi_{l,n}]$, we use the symmetrization argument:

$$\mathbf{E}[\xi_{l,n}] \leq 2\mathbf{E}\left[\sup_{\pi \in \Pi_n(0)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \bar{\ell}'(f_{\pi}(X_i)) f_l(X_i) \right| \right] \quad (5.43)$$

$$\leq 2M \mathbf{E}\left[\sup_{\pi \in \Pi_n(0)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \bar{\ell}'(f_{\pi}(X_i)) \right| \right] \quad (5.44)$$

$$\leq \frac{2M}{\mu} \mathbf{E}\left[\left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \right| \right] + 2M \mathbf{E}\left[\sup_{\pi \in \Pi_n(0)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i [\bar{\ell}'(f_{\pi}(X_i)) - \bar{\ell}'(0)] \right| \right],$$

$$(5.45)$$

where the second inequality comes from [4, Th. 11.5]. Note that the function $\bar{\ell}'$, the derivative of $\bar{\ell}$ defined in (5.37), is by construction Lipschitz with constant $1/\mu^2$.

Therefore, in view of the contraction principle,

$$\mathbf{E}[\xi_{l,n}] \leq \frac{2M}{\mu}\mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)^2\right]^{1/2} + \frac{4M}{\mu^2}\mathbf{E}\left[\sup_{\boldsymbol{\pi}\in\Pi_n(0)}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_{\boldsymbol{\pi}}(X_i)\right] \quad (5.46)$$

$$\leq \frac{2M}{\mu\sqrt{n}} + \frac{4M}{\mu^2}\mathbf{E}\left[\sup_{k\in[K]}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_k(X_i)\right] \quad (5.47)$$

$$\leq \frac{2M}{\mu\sqrt{n}} + \frac{8M^2}{\mu^2}\left(\frac{\log K}{2n}\right)^{1/2} \leq \frac{2V^2(1+2\sqrt{2\log K})}{\sqrt{n}}. \quad (5.48)$$

As a consequence, we proved that with probability larger than $1-\delta$, we have

$$\xi_n \leq 8V^2\left(\frac{\log K}{n}\right)^{1/2}.$$

This completes the proof of the first inequality. In order to prove the second one, we simply change the way we have evaluated the term $\int \bar{\ell}(f_{\hat{\boldsymbol{\pi}}})f^*$ in the left hand side of (5.39). Since $\bar{\ell}$ is strongly convex with a second order derivative bounded from below by $1/M^2$, we have

$$\bar{\ell}(f_{\hat{\boldsymbol{\pi}}}) \geq \bar{\ell}(f^*) + \bar{\ell}'(f^*)(f_{\hat{\boldsymbol{\pi}}} - f^*) + \frac{1}{2M^2}(f_{\hat{\boldsymbol{\pi}}} - f^*)^2.$$

Since f^* is always larger than μ, the derivative $\bar{\ell}'(f^*)$ equals $1/f^*$. Integrating over $\mathcal{X}$, we get the second inequality of the proposition.

5.6. Auxiliary results. We start by a general convex result based on the strong convexity of the $-\log$ function to derive a bound on the estimated log-likelihood.

Lemma 5.2. *Let us assume that* $M = \max_{j\in[K]}\|f_j\|_\infty < \infty$. *Then, for any* $\boldsymbol{\pi}\in\mathbb{B}_+^K$, *it holds that*

$$L_n(\hat{\boldsymbol{\pi}}) \leq L_n(\boldsymbol{\pi}) - \frac{1}{2M^2 n}\|\bar{\mathbf{Z}}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2. \quad (5.49)$$

Proof. Recall that $\hat{\boldsymbol{\pi}}$ minimizes the function L_n defined in (1.7) over Π_n. Furthermore, the function $u \mapsto \ell(u)$ is clearly strongly convex with a second order derivative bounded from below by $1/M^2$ over the set $u \in (0, M]$. Therefore, for every $\hat{u} \in (0, M]$, the function $\tilde{\ell}$ given by:

$$\tilde{\ell}(u) = \ell(u) - \frac{1}{2M^2}(\hat{u} - u)^2, \quad u \in (0, M], \quad (5.50)$$

is convex. This implies that the mapping

$$\boldsymbol{\pi} \mapsto \tilde{L}_n(\boldsymbol{\pi}) = L_n(\boldsymbol{\pi}) - \frac{1}{2M^2 n}\|\mathbf{Z}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 \quad (5.51)$$

is convex over the set $\boldsymbol{\pi} \in \mathbb{B}_+^K$. This yields[4]

$$\widetilde{L}_n(\boldsymbol{\pi}) - \widetilde{L}_n(\widehat{\boldsymbol{\pi}}) \geq \sup_{\boldsymbol{v} \in \partial \widetilde{L}_n(\widehat{\boldsymbol{\pi}})} \boldsymbol{v}^\top(\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}}), \quad \forall \boldsymbol{\pi} \in \mathbb{B}_+^K. \tag{5.52}$$

Using the Karush–Kuhn–Tucker conditions and the fact that $\widehat{\boldsymbol{\pi}}$ minimizes L_n, we get $\boldsymbol{0}_K \in \partial L_n(\widehat{\boldsymbol{\pi}}) = \partial \widetilde{L}_n(\widehat{\boldsymbol{\pi}})$. This readily gives $\widetilde{L}_n(\boldsymbol{\pi}) - \widetilde{L}_n(\widehat{\boldsymbol{\pi}}) \geq 0$, for any $\boldsymbol{\pi} \in \mathbb{B}_+^K$. The last step is to remark that $\mathbf{Z}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}) = \overline{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})$, since both $\widehat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$ have entries summing to one. $\square$

The core of our results lies in the following proposition which bound the deviations of the empirical process part.

Proposition 5.3 (Supremum of empirical process). *For any $\boldsymbol{\pi} \in \mathbb{B}_+^K$ and $x \in \mathcal{X}$, define $\varphi(\pi, x) = \int (\log f_{\boldsymbol{\pi}}) f^* - \log f_{\boldsymbol{\pi}}(x)$ and consider $\Phi_n(\pi) = \frac{1}{n} \sum_{i=1}^n \varphi(\boldsymbol{\pi}, X_i)$. If $K \geq 2$, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\zeta_n = \sup_{\boldsymbol{\pi} \in \Pi_n} \big\| \nabla \Phi_n(\boldsymbol{\pi}) \big\|_\infty \leq 8V^3 \Big(\frac{\log(K/\delta)}{n} \Big)^{1/2}. \tag{5.53}$$

Furthermore, we have $\mathbf{E}[\zeta_n] \leq 4V^3 (2 \log(2K^2)/n)^{1/2}$ and $\mathbf{Var}[\zeta_n] \leq V^2/(2n)$.

Proof. To ease notation, let us denote

$$g_{\boldsymbol{\pi}, l}(x) = \frac{f_l(x)}{f_{\boldsymbol{\pi}}(x)} - \mathbf{E}\bigg[\frac{f_l(X)}{f_{\boldsymbol{\pi}}(X)} \bigg]$$

and

$$F(\mathbf{X}) = \sup_{\boldsymbol{\pi} \in \Pi_n} \big\| \nabla \Phi_n(\boldsymbol{\pi}) \big\|_\infty = \sup_{(\boldsymbol{\pi}, l) \in \Pi_n \times [K]} \bigg| \frac{1}{n} \sum_{i=1}^n g_{\boldsymbol{\pi}, l}(X_i) \bigg|, \tag{5.54}$$

where $\mathbf{X} = (X_1, \ldots, X_n)$. To derive a bound on F, we will use the McDiarmid concentration inequality that requires the bounded difference condition to hold for F. For some $i_0 \in [n]$, let $\mathbf{X}' = (X_1, \ldots, X_{i_0}', \ldots, X_n)$ be a new sample obtained from $\mathbf{X}$ by modifying the i_0th element X_i and by leaving all the others unchanged. Then, we have

$$F(\mathbf{X}) - F(\mathbf{X}') = \sup_{(\boldsymbol{\pi}, l) \in \Pi_n \times [K]} \bigg| \frac{1}{n} \sum_{i=1}^n g_{\boldsymbol{\pi}, l}(X_i) \bigg| - \sup_{(\boldsymbol{\pi}, l) \in \Pi \times [K]} \bigg| \frac{1}{n} \sum_{i=1}^n g_{\boldsymbol{\pi}, l}(X_i') \bigg| \tag{5.55}$$

$$\leq \sup_{(\boldsymbol{\pi}, l) \in \Pi_n \times [K]} \bigg| \frac{1}{n} \sum_{i=1}^n g_{\boldsymbol{\pi}, l}(X_i) - \frac{1}{n} \sum_{i=1}^n g_{\boldsymbol{\pi}, l}(X_i') \bigg| \tag{5.56}$$

---

[4]We denote by ∂g the sub-differential of a convex function g.

$$= \sup_{(\boldsymbol{\pi},l) \in \Pi_n \times [K]} \left| \frac{1}{n} \big(g_{\boldsymbol{\pi},l}(X_{i_0}) - g_{\boldsymbol{\pi},l}(X'_{i_0}) \big) \right| \leq \frac{V}{n}, \qquad (5.57)$$

where the last inequality is a direct consequence of assumption (1.12). Therefore, using the McDiarmid concentration inequality recalled in Theorem A.2 below, we check that the inequality

$$F(\mathbf{X}) \leq \mathbf{E}\big(F(\mathbf{X})\big) + V \sqrt{\frac{\log(1/\delta)}{2n}} \qquad (5.58)$$

holds with probability at least $1 - \delta$. Furthermore, in view of the Efron–Stein inequality, we have

$$\mathbf{Var}[\zeta_n] = \mathbf{Var}[F(\mathbf{X})] \leq \frac{V^2}{2n}. \qquad (5.59)$$

Let us denote

$$\mathcal{G} := \big\{ (f_l/f_{\boldsymbol{\pi}}) - 1, (\boldsymbol{\pi}, l) \in \Pi_n \times [K] \big\}$$

and $\mathfrak{R}_{n,q}(\mathcal{G})$ the Rademacher complexity of $\mathcal{G}$ given by

$$\mathfrak{R}_n(\mathcal{G}) = \mathbf{E}_\epsilon \left[\sup_{(\boldsymbol{\pi},l) \in \Pi_n \times [K]} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \Big(\frac{f_l(X_i)}{f_{\boldsymbol{\pi}}(X_i)} - 1 \Big) \right| \right], \qquad (5.60)$$

with $\epsilon_1, \ldots, \epsilon_n$ independent and identically distributed Rademacher random variables independent of $X_1, \ldots, X_n$. Using the symmetrization inequality (see, for instance, [13, Theorem 2.1]) we have

$$\mathbf{E}[F(\mathbf{X})] = \mathbf{E}[\zeta_n] \leq 2\mathbf{E}[\mathfrak{R}_n(\mathcal{G})]. \qquad (5.61)$$

Lemma 5.4. *The Rademacher complexity defined in* (5.60) *satisfies*

$$\mathfrak{R}_n(\mathcal{G}) \leq 4V^3 \sqrt{\frac{\log K}{n}}. \qquad (5.62)$$

Proof. The proof relies on the contraction principle of [17] that we recall in Appendix A for the convenience. We apply this principle to the random variables

$$X_{i,(\boldsymbol{\pi},l)} = f_{\boldsymbol{\pi}}(X_i)/f_l(X_i) - 1$$

and to the function $\psi(x) = (1 + x)^{-1} - 1$. Clearly ψ is Lipschitz on $[\frac{1}{V} - 1, V - 1]$ with the Lipschitz constant equal to V^2 and $\psi(0) = 0$. Therefore

$$\mathfrak{R}_n(\mathcal{G}) \leq \mathbf{E}_\epsilon \left[\sup_{(\boldsymbol{\pi},l)} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \psi\big(X_{i,(\boldsymbol{\pi},l)}\big) \right] + \mathbf{E}_\epsilon \left[\sup_{(\boldsymbol{\pi},l)} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (-\psi)\big(X_{i,(\boldsymbol{\pi},l)}\big) \right]$$

$$\leq 2V^2 \mathbf{E}_\epsilon \left[\sup_{(\boldsymbol{\pi},l) \in \Pi_n \times [K]} \frac{1}{n} \sum_{i=1}^n \epsilon_i X_{i,(\boldsymbol{\pi},l)} \right]$$

$$= 2V^2 \mathbf{E}_\epsilon \left[\sup_{(\boldsymbol{\pi},l) \in \Pi_n \times [K]} \frac{1}{n} \sum_{i=1}^n \epsilon_i \left(\frac{f_{\boldsymbol{\pi}}(X_i)}{f_l(X_i)} - 1 \right) \right]. \tag{5.63}$$

Expanding $f_{\boldsymbol{\pi}}(X_i)$ we obtain

$$\mathbf{E}_\epsilon \left[\sup_{(\boldsymbol{\pi},l)} \frac{1}{n} \sum_{i=1}^n \epsilon_i \left(\frac{f_{\boldsymbol{\pi}}(X_i)}{f_l(X_i)} - 1 \right) \right] = \mathbf{E}_\epsilon \left[\sup_{(\boldsymbol{\pi},l)} \sum_{k=1}^K \frac{\pi_k}{n} \sum_{i=1}^n \epsilon_i \left(\frac{f_k(X_i)}{f_l(X_i)} - 1 \right) \right]$$

$$= \mathbf{E}_\epsilon \left[\max_{k,l \in [K]} \frac{1}{n} \sum_{i=1}^n \epsilon_i \left(\frac{f_k(X_i)}{f_l(X_i)} - 1 \right) \right]. \tag{5.64}$$

We apply now Theorem A.1 with $s = (k,l)$, $N = K^2$, $a = -V$, $b = V$ and $Y_{i,s} = \epsilon_i \left(\frac{f_k(X_i)}{f_l(X_i)} - 1 \right)$. This yields

$$\mathbf{E}_\epsilon \left[\max_{k,l \in [K]} \frac{1}{n} \sum_{i=1}^n \epsilon_i \left(\frac{f_k(X_i)}{f_l(X_i)} - 1 \right) \right] \leq 2V \left(\frac{\log K^2}{2n} \right)^{1/2}. \tag{5.65}$$

This completes the proof of the lemma. $\qquad\square$

Combining inequalities (5.58, 5.61) and Lemma 5.4, we get that the inequality

$$F(\mathbf{X}) \leq 8V^3 \left(\frac{\log K}{n} \right)^{1/2} + V \left(\frac{\log(1/\delta)}{2n} \right)^{1/2} \tag{5.66}$$

holds with probability at least $1 - \delta$. Noticing that $V \geq 1$ and, for $K \geq 2$, $\delta \in (0, K^{-1/31})$ we have

$$8\sqrt{\log K} + \sqrt{(1/2)\log(1/\delta)} \leq 8\sqrt{\log(K/\delta)},$$

we get the first claim of the proposition. The second claim is a direct consequence of Lemma 5.4 and (5.61). $\qquad\square$

6. Proof of the lower bound for nearly-D-sparse aggregation

We prove the minimax lower bound for estimation in Kullback–Leibler risk using the following slightly adapted version of Theorem 2.5 from [27]. Throughout this section, we denote by $\lambda_{\min,\boldsymbol{\Sigma}}(k)$ and $\lambda_{\max,\boldsymbol{\Sigma}}(k)$, respectively, the smallest and the largest eigenvalue of all $k \times k$ principal minors of the matrix $\boldsymbol{\Sigma}$.

Theorem 6.1. *For some integer $L \geq 4$ assume that $\mathcal{H}_{\mathcal{F}}(\gamma, D)$ contains L elements $f_{\boldsymbol{\pi}^{(1)}}, \ldots, f_{\boldsymbol{\pi}^{(L)}}$ satisfying the following two conditions.*

(i) $\mathrm{KL}(f_{\boldsymbol{\pi}^{(j)}} \| f_{\boldsymbol{\pi}^{(k)}}) \geq 2s > 0$, *for all pairs (j, k) such that $1 \leq j < k \leq L$.*

(ii) *For product densities f_ℓ^n defined on $\mathcal{X}^n$ by*

$$f_\ell^n(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = f_{\boldsymbol{\pi}^{(\ell)}}(\boldsymbol{x}_1) \times \cdots \times f_{\boldsymbol{\pi}^{(\ell)}}(\boldsymbol{x}_n)$$

it holds

$$\max_{\ell \in [L]} \mathrm{KL}(f_\ell^n \| f_1^n) \leq \frac{\log L}{16}. \tag{6.1}$$

Then

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma, D)} \mathbf{P}_f\big(\mathrm{KL}(f \| \hat{f}) \geq s\big) \geq 0.17. \tag{6.2}$$

To establish the bound claimed in Theorem 3.1, we will split the problem into two parts, corresponding to the following two subsets of $\mathcal{H}_{\mathcal{F}}(\gamma, D)$

$$\mathcal{H}_{\mathcal{F}}(0, D) = \big\{ f_{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \mathbb{B}_+^K \text{ s.t. } \exists J \subset [K] \text{ with } \|\boldsymbol{\pi}_{J^c}\|_1 = 0 \text{ and } |J| \leq D \big\},$$

$$\mathcal{H}_{\mathcal{F}}(\gamma, 1) = \Big\{ f_{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \mathbb{B}_+^K \text{ s.t. } \pi_1 = 1 - \gamma \text{ and } \sum_{j=2}^K \pi_j = \gamma \Big\}. \tag{6.3}$$

We will show that over $\mathcal{H}_{\mathcal{F}}(0, D)$, we have a lower bound of order $\log(1 + K/D)/n$ while over $\mathcal{H}_{\mathcal{F}}(\gamma, 1)$, a lower bound of order $\big[\frac{\gamma^2}{n} \log\big(1 + K/(\gamma\sqrt{n})\big)\big]^{1/2}$ holds true. Therefore, the lower bound over $\mathcal{H}_{\mathcal{F}}(\gamma, D)$ is larger than the average of these bounds.

For any $M \geq 1$ and $k \in [M - 1]$, let Ω_k^M be the subset of $\{0, 1\}^M$ defined by

$$\Omega_k^M := \big\{ \boldsymbol{\omega} \in \{0, 1\}^M : \|\boldsymbol{\omega}\|_1 = k \big\}. \tag{6.4}$$

Before starting, we remind here a version of the Varshamov–Gilbert lemma (see, for instance, [25, Lemma 8.3]) which will be helpful for deriving our lower bounds.

Lemma 6.2. *Let $M \geq 4$ and $k \in [M/2]$ be two integers. Then there exist a subset $\Omega \subset \Omega_k^M$ and an absolute constant C_1 such that*

$$\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1 \geq \frac{k+1}{4} \quad \forall \boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega \text{ s.t. } \boldsymbol{\omega} \neq \boldsymbol{\omega}' \tag{6.5}$$

and $L = |\Omega|$ satisfies $L \geq 4$ and

$$\log L \geq C_1 k \log\Big(1 + \frac{eM}{k}\Big). \tag{6.6}$$

We will also use the following lemma that allows us to relate the KL-divergence $\mathrm{KL}(f_{\boldsymbol{\pi}} \| f_{\boldsymbol{\pi}'})$ to the Euclidean distance between the weight vectors $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$.

Lemma 6.3. *If the dictionary $\mathcal{F}$ satisfies the boundedness assumption* (1.12)*, then for any $f_{\boldsymbol{\pi}}, f_{\boldsymbol{\pi}'} \in \mathcal{H}_{\mathcal{F}}(\gamma, D)$ we have*

$$\frac{1}{2V^2 M} \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\pi}' - \boldsymbol{\pi})\|_2^2 \leq \mathrm{KL}(f_{\boldsymbol{\pi}} \| f_{\boldsymbol{\pi}'}) \leq \frac{V^2}{2m} \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\pi}' - \boldsymbol{\pi})\|_2^2. \qquad (6.7)$$

Proof. Using the Taylor expansion, one can check that for any $u \in [1/V, V]$, we have

$$(1 - u) + \frac{1}{2V^2}(u - 1)^2 \leq -\log u \leq (1 - u) + \frac{V^2}{2}(u - 1)^2.$$

Therefore,

$$\frac{1}{2V^2} \int_{\mathcal{X}} \left(\frac{f_{\boldsymbol{\pi}'}}{f_{\boldsymbol{\pi}}} - 1 \right)^2 f_{\boldsymbol{\pi}} \, d\nu \leq \mathrm{KL}(f_{\boldsymbol{\pi}} \| f_{\boldsymbol{\pi}'}) \leq \frac{V^2}{2} \int_{\mathcal{X}} \left(\frac{f_{\boldsymbol{\pi}'}}{f_{\boldsymbol{\pi}}} - 1 \right)^2 f_{\boldsymbol{\pi}} \, d\nu. \quad (6.8)$$

Since $\mathcal{F}$ satisfies the boundedness assumption, we get

$$\frac{1}{2MV^2} \int_{\mathcal{X}} \left(f_{\boldsymbol{\pi}'} - f_{\boldsymbol{\pi}} \right)^2 d\nu \leq \mathrm{KL}(f_{\boldsymbol{\pi}} \| f_{\boldsymbol{\pi}'}) \leq \frac{V^2}{2m} \int_{\mathcal{X}} \left(f_{\boldsymbol{\pi}'} - f_{\boldsymbol{\pi}} \right)^2 d\nu. \quad (6.9)$$

The claim of the lemma follows from these inequalities and the fact that

$$\int_{\mathcal{X}} \left(f_{\boldsymbol{\pi}'} - f_{\boldsymbol{\pi}} \right)^2 d\nu = \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\pi}' - \boldsymbol{\pi})\|_2^2. \qquad \square$$

6.1. Lower bound on $\mathcal{H}_{\mathcal{F}}(0, D)$. We show that the lower bound

$$\frac{D}{n} \log \left(1 + \frac{eK}{D} \right) \bigwedge \left(\frac{1}{n} \log \left(1 + \frac{K}{\sqrt{n}} \right) \right)^{1/2}$$

holds when we consider the worst case error for f^* belonging to the set $\mathcal{H}_{\mathcal{F}}(0, D)$.

Proposition 6.4. *If $\log(1 + eK) \leq n$ then, for the constant*

$$C_2 = \frac{C_1 m \bar{\kappa}_{\boldsymbol{\Sigma}}(2D, 0)}{2^9 V^2 M (C_1 m \vee 4V^2 \lambda_{\max, \boldsymbol{\Sigma}}(2D))} \geq \frac{C_1 m \varkappa_*}{2^9 V^2 M (C_1 m \vee 4V^2 \varkappa^*)}, \quad (6.10)$$

we have

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(0,D)} \mathbf{P}_f \left(\mathrm{KL}(f \| \hat{f}) \geq C_2 \frac{D \log(1 + K/D)}{n} \bigwedge \left(\frac{\log \left(1 + K/\sqrt{n} \right)}{n} \right)^{1/2} \right)$$

$$\geq 0.17. \quad (6.11)$$

Proof. We assume that $D \leq K/2$. The case $D > K/2$ can be reduced to the case $D = K/2$ by using the inclusion $\mathcal{H}_{\mathcal{F}}(0, K/2) \subset \mathcal{H}_{\mathcal{F}}(0, D)$. Let us set

$$A_1 = 4 \vee 16V^2 \lambda_{\max, \boldsymbol{\Sigma}}(2D)/(C_1 m)$$

and denote by d the largest integer such that

$$d \le D \quad \text{and} \quad d^2 \log\left(1 + \frac{eK}{d}\right) \le A_1 n. \tag{6.12}$$

According to Lemma 6.2, there exists a subset $\Omega = \{\boldsymbol{\omega}^{(\ell)} : \ell \in [L]\}$ of Ω_d^K of cardinality $L \ge 4$ satisfying

$$\log L \ge C_1 d \log(1 + eK/d)$$

such that for any pair of distinct elements $\boldsymbol{\omega}^{(\ell)}, \boldsymbol{\omega}^{(\ell')} \in \Omega$ we have

$$\|\boldsymbol{\omega}^{(\ell)} - \boldsymbol{\omega}^{(\ell')}\|_1 \ge d/4.$$

Using these binary vectors $\boldsymbol{\omega}^{(\ell)}$, we define the set $\mathcal{D} = \{\boldsymbol{\pi}^{(1)}, \dots, \boldsymbol{\pi}^{(L)}\} \subset \mathbb{B}_+^K$ as follows:

$$\boldsymbol{\pi}^{(1)} = \boldsymbol{\omega}^{(1)}/d, \quad \boldsymbol{\pi}^{(\ell)} = (1-\varepsilon)\boldsymbol{\pi}^{(1)} + \varepsilon\boldsymbol{\omega}^{(\ell)}/d, \quad \ell = 2, \dots, L. \tag{6.13}$$

Clearly, for every $\varepsilon \in [0, 1]$, the vectors $\boldsymbol{\pi}^{(\ell)}$ belong to $\mathbb{B}_+^K$. Furthermore, for any pair of distinct values $\ell, \ell' \in [L]$, we have

$$\|\boldsymbol{\pi}^{(\ell)} - \boldsymbol{\pi}^{(\ell')}\|_q^q = (\varepsilon/d)^q \|\boldsymbol{\omega}^{(\ell)} - \boldsymbol{\omega}^{(\ell')}\|_1 \ge (\varepsilon/d)^q d/4.$$

In view of Lemma 6.3, this yields

$$\mathrm{KL}(f_{\boldsymbol{\pi}^{(\ell)}} \| f_{\boldsymbol{\pi}^{(\ell')}}) \ge \frac{\overline{\kappa}_{\boldsymbol{\Sigma}}(2d, 0)}{4V^2 M d} \|\boldsymbol{\pi}^{(\ell)} - \boldsymbol{\pi}^{(\ell')}\|_1^2 \ge \frac{\overline{\kappa}_{\boldsymbol{\Sigma}}(2D, 0)}{64V^2 M} \times \frac{\varepsilon^2}{d}. \tag{6.14}$$

Let us choose

$$\varepsilon^2 = \frac{d^2 \log(1 + eK/d)}{nA_1}. \tag{6.15}$$

It follows from (6.12) that $\varepsilon \le 1$. Inserting this value of ε in (6.14), we get

$$\mathrm{KL}(f_{\boldsymbol{\pi}^{(\ell)}} \| f_{\boldsymbol{\pi}^{(\ell')}}) \ge 2C_2 \frac{d \log(1 + eK/d)}{n}. \tag{6.16}$$

This inequality shows that condition (i) of Theorem 6.1 is satisfied with

$$s = C_2 \, (d/n) \log(1 + eK/d).$$

For the second condition of the same theorem, we have

$$\max_{\ell \in [L]} \mathrm{KL}(f_\ell^n \| f_1^n) = n \max_\ell \mathrm{KL}(f_{\boldsymbol{\pi}^{(\ell)}} \| f_{\boldsymbol{\pi}^{(1)}}) \tag{6.17}$$

$$\le \frac{nV^2 \lambda_{\max, \boldsymbol{\Sigma}}(2d)}{2m} \max_\ell \|\boldsymbol{\pi}^{(\ell)} - \boldsymbol{\pi}^{(1)}\|_2^2 \tag{6.18}$$

$$\leq \frac{nV^2\lambda_{\max,\mathbf{\Sigma}}(2D)}{m} \times \frac{\varepsilon^2}{d}, \tag{6.19}$$

since one can check that

$$\|\boldsymbol{\pi}^{(\ell)} - \boldsymbol{\pi}^{(1)}\|_2^2 \leq (\varepsilon/d)^2\|\boldsymbol{\omega}^{(\ell)} - \boldsymbol{\omega}^{(1)}\|_1 \leq 2\varepsilon^2/d.$$

Therefore, using the definition of ε, we get

$$\max_{\ell\in[L]} \mathrm{KL}(f_\ell^n\|f_1^n) \leq \frac{nV^2\lambda_{\max,\mathbf{\Sigma}}(2D)}{m} \times \frac{C_1 dm\log(1+eK/d)}{16nV^2\lambda_{\max,\mathbf{\Sigma}}(2D)} \tag{6.20}$$

$$= \frac{C_1 d\log(1+eK/d)}{16} \leq \frac{\log L}{16}. \tag{6.21}$$

Theorem 6.1 implies that

$$\inf_{\hat{f}} \sup_{f\in\mathcal{H}_{\mathcal{F}}(0,D)} \mathbf{P}_f\left(\mathrm{KL}(f\|\hat{f}) \geq C_2\frac{d\log(1+eK/d)}{n}\right) \geq 0.17. \tag{6.22}$$

We use the fact that d is the largest integer satisfying (6.12). Therefore, either $d+1 > D$ or

$$(d+1)^2\log\left(1+\frac{eK}{d+1}\right) \geq A_1 n. \tag{6.23}$$

If $d \geq D$, then the claim of the proposition follows from (6.22), since

$$d\log(1+eK/d) \geq D\log(1+eK/D).$$

On the other hand, if (6.23) is true, then

$$d\log(1+eK/d) \geq \frac{1}{2}(d+1)\log(1+eK/(d+1))$$

$$\geq \frac{1}{2}\left(A_1 n\log(1+eK/(d+1))\right)^{1/2}. \tag{6.24}$$

In addition, $d^2\log(1+eK/d) \leq A_1 n$ implies that $(d+1)^2 \leq A_1 n$. Combining the last two inequalities, we get the inequality

$$d\log(1+eK/d) \geq \frac{1}{2}\left(A_1 n\log(1+eK/\sqrt{A_1 n})\right)^{1/2} \geq \left(n\log(1+eK/\sqrt{n})\right)^{1/2}.$$

Therefore, in view of (6.22), we get the claim of the proposition. $\square$

6.2. Lower bound on $\mathcal{H}_{\mathcal{F}}(\gamma,1)$. Next result shows that the lower bound

$$\frac{\gamma^2}{n}\log\left(1+\frac{K}{\gamma\sqrt{n}}\right)$$

holds for the worst case error when f^* belongs to the set $\mathcal{H}_{\mathcal{F}}(\gamma,1)$.

Proposition 6.5. *Assume that*

$$\left(\frac{\log(1 + eK)}{n}\right)^{1/2} \leq 2\gamma. \tag{6.25}$$

Then, for the constant

$$C_3 = \frac{C_1 m \bar{\kappa}_{\Sigma}(2D, 0)}{2^{12} V^4 M \lambda_{\max, \Sigma}(2D)},$$

it holds that

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma, 1)} \mathbf{P}_f\left(\mathrm{KL}(f \| \hat{f}) \geq C_3 \left\{\frac{\gamma^2}{n} \log\left(1 + \frac{K}{\gamma \sqrt{n}}\right)\right\}^{1/2}\right) \geq 0.17. \tag{6.26}$$

Proof. Let $C > 2$ be a constant the precise value of which will be specified later. Denote by d the largest integer satisfying

$$d \sqrt{\log(1 + eK/d)} \leq C \gamma \sqrt{n}. \tag{6.27}$$

Note that $d \geq 1$ in view of the condition $(\log(1 + eK)/n)^{1/2} \leq 2\gamma$ of the proposition. This readily implies that $d \leq C \gamma \sqrt{n}$ and, therefore,

$$\frac{\gamma}{d} \geq C^{-1}\left\{\frac{1}{n} \log\left(1 + \frac{eK}{C\gamma\sqrt{n}}\right)\right\}^{1/2} \geq 2C^{-2}\left\{\frac{1}{n} \log\left(1 + \frac{K}{\gamma\sqrt{n}}\right)\right\}^{1/2}. \tag{6.28}$$

Let us first consider the case $d \leq (K-1)/2$. According to Lemma 6.2, there exists a subset $\Omega \subset \Omega_d^{K-1}$ of cardinality L satisfying

$$\log L \geq C_1 d \log\left(1 + \frac{e(K-1)}{d}\right) \quad \text{and} \quad \|\boldsymbol{\omega}^{(\ell)} - \boldsymbol{\omega}^{(\ell')}\|_1 \geq d/4$$

for any pair of distinct elements $\boldsymbol{\omega}, \boldsymbol{\omega}'$ taken from Ω. With these binary vectors in hand, we define the set $\mathcal{D} \subset \mathbb{B}_+^K$ of cardinality L as follows:

$$\mathcal{D} = \left\{\boldsymbol{\pi} = \left(1 - \gamma, \gamma \boldsymbol{\omega}/d\right) : \omega \in \Omega\right\}. \tag{6.29}$$

It is clear that all the vectors of $\mathcal{D}$ belong to $\mathcal{H}_{\mathcal{F}}(\gamma, 1)$. Let us fix now an element of $\mathcal{D}$ and denote it by $\boldsymbol{\pi}^1$, the corresponding element of Ω being denoted by $\boldsymbol{\omega}^1$. We have

$$\max_{\boldsymbol{\pi} \in \mathcal{D}} \mathrm{KL}(f_{\boldsymbol{\pi}}^n \| f_{\boldsymbol{\pi}^1}^n) \leq \frac{nV^2}{2m} \max_{\boldsymbol{\pi} \in \mathcal{D}} \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\pi} - \boldsymbol{\pi}^1)\|_2^2 \tag{6.30}$$

$$\leq \frac{nV^2 \lambda_{\max, \boldsymbol{\Sigma}}(2d)\gamma^2}{2md^2} \max_{\boldsymbol{\omega} \in \Omega} \|\boldsymbol{\omega} - \boldsymbol{\omega}^1\|_2^2 \tag{6.31}$$

$$\leq \frac{nV^2 \lambda_{\max, \boldsymbol{\Sigma}}(2d)\gamma^2}{md}. \tag{6.32}$$

The definition of d yields

$$(d + 1)\sqrt{\log(1 + eK/(d + 1))} > C\gamma\sqrt{n},$$

which implies that

$$\frac{\gamma^2}{d} \leq 2(d + 1)\frac{\gamma^2}{(d + 1)^2} \leq 2(d + 1)\frac{\log(1 + eK/(d + 1))}{nC^2}$$
$$\leq \frac{4d\log(1 + e(K - 1)/d)}{nC^2}. \tag{6.33}$$

Combined with Equation (6.32), this implies that

$$\max_{\boldsymbol{\pi} \in \mathcal{D}} \mathrm{KL}(f_{\boldsymbol{\pi}}^n \| f_{\boldsymbol{\pi}^1}^n) \leq \frac{nV^2\lambda_{\max,\boldsymbol{\Sigma}}(2d)}{m} \times \frac{4d\log(1 + e(K - 1)/d)}{nC^2} \tag{6.34}$$

$$= \frac{4V^2\lambda_{\max,\boldsymbol{\Sigma}}(2d)}{mC^2} \times d\log(1 + e(K - 1)/d). \tag{6.35}$$

Choosing

$$C^2 = 2 \vee \frac{64V^2\lambda_{\max,\boldsymbol{\Sigma}}(2d)}{C_1 m}$$

we get that

$$\max_{\boldsymbol{\pi} \in \mathcal{D}} \mathrm{KL}(f_{\boldsymbol{\pi}}^n \| f_{\boldsymbol{\pi}^1}^n) \leq \frac{1}{16}C_1 d\log(1 + e(K - 1)/d) \leq \frac{\log L}{16}.$$

Furthermore, for any $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \mathcal{D}$, in view of Lemma 6.3 and (6.28), we have

$$\mathrm{KL}(f_{\boldsymbol{\pi}} \| f_{\boldsymbol{\pi}'}) \geq \frac{\overline{\kappa}_{\boldsymbol{\Sigma}}(2d, 0)}{4V^2 Md}\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_1^2 = \frac{\overline{\kappa}_{\boldsymbol{\Sigma}}(2d, 0)\gamma^2}{4V^2 Md^3}\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1^2 \tag{6.36}$$

$$\geq \frac{\overline{\kappa}_{\boldsymbol{\Sigma}}(2d, 0)}{64V^2 M} \times \frac{\gamma^2}{d} \tag{6.37}$$

$$\geq \frac{\overline{\kappa}_{\boldsymbol{\Sigma}}(2d, 0)}{32V^2 MC^2} \times \left\{\frac{\gamma^2}{n}\log\left(1 + \frac{K}{\gamma\sqrt{n}}\right)\right\}^{1/2}. \tag{6.38}$$

Since $\overline{\kappa}_{\boldsymbol{\Sigma}}(2d, 0)/32V^2 MC^2 = 2C_3$, this implies that Theorem 6.1 can be applied, which leads to the inequality

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma, 1)} \mathbf{P}_f\left(\mathrm{KL}(f \| \hat{f}) \geq C_3\left\{\frac{\gamma^2}{n}\log\left(1 + \frac{K}{\gamma\sqrt{n}}\right)\right\}^{1/2}\right) \geq 0.17. \tag{6.39}$$

To complete the proof of the proposition, we have to consider the case $d > (K-1)/2$. In this case, we can repeat all the previous arguments for $d = K/2$ and get the desired inequality. $\qquad\square$

6.3. Lower bound holding for all densities. Now that we have lower bounds in probability for $\mathcal{H}_{\mathcal{F}}(0, D)$ and $\mathcal{H}_{\mathcal{F}}(\gamma, 1)$, we can derive a lower bound in expectation for $\mathcal{H}_{\mathcal{F}}(\gamma, D)$. In particular, to prove Theorem 3.1, we will use the inequality

$$\mathcal{R}\big(\mathcal{H}_{\mathcal{F}}(\gamma, D)\big) \geq \inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\mathcal{F}}(0,D) \cup \mathcal{H}_{\mathcal{F}}(\gamma,1)} \mathbf{E}[\text{KL}(f^* || \hat{f})]. \qquad (6.40)$$

Proof of Theorem 3.1. To ease notation, let us define

$$r(n, K, \gamma, D) = \left[\frac{\gamma^2}{n} \log\left(1 + \frac{K}{\gamma\sqrt{n}}\right) \right]^{1/2} + \frac{D \log(1 + K/D)}{n} \bigwedge \left(\frac{\log(1 + K/\sqrt{n})}{n} \right)^{1/2}. \qquad (6.41)$$

We first consider the case where the dominating term is the first one, that is

$$\left[\frac{\gamma^2}{n} \log\left(1 + \frac{K}{\gamma\sqrt{n}}\right) \right]^{1/2} \geq \frac{3D \log(1 + K/D)}{n}. \qquad (6.42)$$

On the one hand, since $D \geq 1$, we have

$$\frac{3D \log(1 + K/D)}{n} \geq \frac{\log(1 + eK)}{n}. \qquad (6.43)$$

On the other hand, using the inequality $\log(1 + x) \leq x$, we get

$$\left[\frac{\gamma^2}{n} \log\left(1 + \frac{K}{\gamma\sqrt{n}}\right) \right]^{1/2} \leq \frac{\gamma}{\sqrt{n}} \left[\log(1 + eK) + \log\left(1 + \frac{1}{e^2\gamma^2 n}\right) \right]^{1/2} \qquad (6.44)$$

$$\leq \gamma \left[\frac{\log(1 + eK)}{n} \right]^{1/2} + \frac{\gamma}{\sqrt{n}} \left[\frac{1}{e^2\gamma^2 n} \right]^{1/2} \qquad (6.45)$$

$$\leq \gamma \left[\frac{\log(1 + eK)}{n} \right]^{1/2} + \frac{\log(1 + eK)}{2n}. \qquad (6.46)$$

Combining (6.42), (6.43) and (6.46), we get

$$\left(\frac{\log(1 + eK)}{n} \right)^{1/2} \leq 2\gamma. \qquad (6.47)$$

This implies that we can apply Proposition 6.5, which yields

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma, D)} \mathbf{P}_f \left(\text{KL}(f || \hat{f}) \geq C_3 \left\{ \frac{\gamma^2}{n} \log\left(1 + \frac{K}{\gamma\sqrt{n}}\right) \right\}^{1/2} \right) \geq 0.17. \qquad (6.48)$$

In view of (6.42), this implies that

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma, D)} \mathbf{P}_f \left(\text{KL}(f || \hat{f}) \geq \tfrac{3}{4} C_3 \, r(n, K, \gamma, D) \right) \geq 0.17. \qquad (6.49)$$

We now consider the second case, where the dominating term in the rate is the second one, that is

$$\left[\frac{\gamma^2}{n}\log\left(1+\frac{K}{\gamma\sqrt{n}}\right)\right]^{1/2} \leq \frac{3D\log(1+K/D)}{n}\bigwedge\left(\frac{\log(1+K/\sqrt{n})}{n}\right)^{1/2}. \quad (6.50)$$

In view of Proposition 6.4, we have

$$\inf_{\hat{f}}\sup_{f\in\mathcal{H}_{\mathcal{F}}(\gamma,D)}\mathbf{P}_f\left(\mathrm{KL}(f||\hat{f}) \geq C_2\frac{D\log(1+K/D)}{n}\bigwedge\left(\frac{\log\left(1+K/\sqrt{n}\right)}{n}\right)^{1/2}\right)$$
$$\geq 0.17. \quad (6.51)$$

In view of (6.50), we get

$$\inf_{\hat{f}}\sup_{f\in\mathcal{H}_{\mathcal{F}}(\gamma,D)}\mathbf{P}_f\left(\mathrm{KL}(f||\hat{f}) \geq \tfrac{1}{4}C_2\,r(n,K,\gamma,D)\right) \geq 0.17. \quad (6.52)$$

Thus, we have proved that $\log(1+eK)\leq n$ implies that

$$\inf_{\hat{f}}\sup_{f\in\mathcal{H}_{\mathcal{F}}(\gamma,D)}\mathbf{P}_f\left(\mathrm{KL}(f||\hat{f}) \geq C_4\,r(n,K,\gamma,D)\right) \geq 0.17, \quad (6.53)$$

for some constant $C_4 > 0$, whatever the relation between γ and D. The desired lower bound follows now from the Tchebychev inequality

$$\mathbf{E}\left[\mathrm{KL}(f||\hat{f})\right] \geq C_4\,r(n,K,\gamma,D)\mathbf{P}_f\left(\mathrm{KL}(f||\hat{f}) \geq C_4\,r(n,K,\gamma,D)\right). \quad \square$$

A. Concentration inequalities

This section contains some well-known results, which are recalled here for the sake of the self-containedness of the paper.

Theorem A.1. *For each* $s = 1,\ldots,N$, *let* $Y_{1,s},\ldots,Y_{n,s}$ *be* n *independent and zero mean random variables such that for some real numbers* a,b *we have* $\mathbf{P}(Y_{i,s}\in[a,b]) = 1$ *for all* $i\in[n]$ *and* $s\in[N]$. *Then, we have*

$$\mathbf{E}\left[\max_{s\in[N]}\frac{1}{n}\sum_{i=1}^{n}Y_{i,s}\right] \leq (b-a)\left(\frac{\log N}{2n}\right)^{1/2},$$

$$\mathbf{E}\left[\max_{s\in[N]}\left|\frac{1}{n}\sum_{i=1}^{n}Y_{i,s}\right|\right] \leq (b-a)\left(\frac{\log(2N)}{2n}\right)^{1/2}. \quad (A.1)$$

Proof. We denote

$$Z_s = \frac{1}{n}\sum_{i=1}^{n}Y_{i,s} \qquad\qquad \text{for } s = 1,\ldots,N,$$

and

$$Z_s = -\frac{1}{n}\sum_{i=1}^{n}Y_{i,s} \qquad\qquad \text{for } s = N+1,\ldots,2N.$$

For every $s \in [2N]$, the logarithmic moment generating function $\psi_s(\lambda) = \log \mathbf{E}[e^{\lambda Z_s}]$ satisfies

$$\psi_s(\lambda) = \log\Big(\prod_i \mathbf{E}[e^{\lambda Y_{i,s}/n}]\Big) = \sum_{i=1}^n \log \mathbf{E}[e^{\lambda Y_{i,s}/n}] \le \frac{\lambda^2(b-a)^2}{8n}, \qquad (\text{A.2})$$

where the last inequality is a consequence of the Hoeffding lemma (see, for instance, [4, Lemma 2.2]). This means that Z_s is sub-Gaussian with variance-factor $v = (b-a)^2/4n$. Therefore, Theorem 2.5 from [4] yields

$$\mathbf{E}[\max_s Z_s] \le \sqrt{2v \log(2N)},$$

which completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$

We group and state together the bounded differences and the Efron–Stein inequalities ([4], Theorems 6.2 and 3.1, respectively).

Theorem A.2. *Assume that a function f satisfies the bounded difference condition: there exist constants c_i, $i = 1,\ldots,n$ such that for all $i = 1,\ldots,n$, all $X = (X_1,\ldots,X_i,\ldots,X_n)$ and $X' = (X_1,\ldots,X_i',\ldots,X_n)$ where only the ith vector is changed*

$$|f(X) - f(X')| \le c_i. \qquad (\text{A.3})$$

Denote

$$v = \sum_{i=1}^n c_i^2. \qquad (\text{A.4})$$

Let $Z = f(X_1,\ldots,X_n)$ where X_i are independent. Then, for every $\delta \in (0,1)$,

$$\mathbf{P}\Big\{Z \le \mathbf{E}Z + \Big(\frac{v\log(1/\delta)}{2}\Big)^{1/2}\Big\} \ge 1 - \delta, \quad \text{and} \quad \mathbf{Var}[Z] \le \frac{v}{2}. \qquad (\text{A.5})$$

Next we state the contraction principle of [17]; a proof can be found in [4, Theorem 11.6].

Theorem A.3. *Let $x_1,\ldots,x_n$ be vectors whose real-valued components are indexed by $\mathcal{T}$, that is, $x_i = (x_{i,s})_{s\in\mathcal{T}}$. For each $i = 1,\ldots,n$ let $\varphi_i : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function such that $\varphi_i(0) = 0$. Let $\epsilon_1,\ldots,\epsilon_n$ be independent Rademacher random variables, and let $\Psi:[0,\infty) \to \mathbb{R}$ be a non-decreasing convex function. Then*

$$\mathbf{E}\Big[\Psi\Big(\frac{1}{2}\sup_{s\in\mathcal{T}}\Big|\sum_{i=1}^n \epsilon_i\varphi_i(x_{i,s})\Big|\Big)\Big] \le \mathbf{E}\Big[\Psi\Big(\sup_{s\in\mathcal{T}}\Big|\sum_{i=1}^n \epsilon_i x_{i,s}\Big|\Big)\Big] \qquad (\text{A.6})$$

$$\mathbf{E}\Big[\Psi\Big(\sup_{s\in\mathcal{T}}\sum_{i=1}^n \epsilon_i\varphi_i(x_{i,s})\Big)\Big] \le \mathbf{E}\Big[\Psi\Big(\sup_{s\in\mathcal{T}}\sum_{i=1}^n \epsilon_i x_{i,s}\Big)\Big]. \qquad (\text{A.7})$$

References

[1] P. C. Bellec, Optimal exponential bounds for aggregation of density estimators, *Bernoulli*, **23** (2017), no. 1, 219–248. Zbl 1368.62085 MR 3556772

[2] P. C. Bellec, A. S. Dalalyan, E. Grappin, and Q. Paris, *On the prediction loss of the lasso in the partially labeled setting*, Technical report, June 2016. arXiv:1606.06179

[3] K. Bertin, E. Le Pennec, and V. Rivoirard, Adaptive Dantzig density estimation, *Ann. Inst. Henri Poincaré Probab. Stat.*, **47** (2011), no. 1, 43–74. Zbl 1207.62077 MR 2779396

[4] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, Oxford, 2013. Zbl 1279.60005 MR 3185193

[5] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, Aggregation for gaussian regression, *Ann. Statist.*, **35** (2007), no. 4, 1674–1697. Zbl 1209.62065 MR 2351101

[6] F. Bunea, A. B. Tsybakov, M. H. Wegkamp, and A. Barbu, Spades and mixture models, *Ann. Statist.*, **38** (2010), no. 4, 2525–2558. Zbl 1198.62025 MR 2676897

[7] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer, *Optimal exponential bounds for aggregation of estimators for the Kullback–Leibler loss*, Technical report, Jan. 2016. arXiv:1601.05686

[8] O. Catoni, *The mixture approach to universal model selection*, Technical report, 1997.

[9] D. Dai, P. Rigollet, and T. Zhang, Deviation optimal learning using greedy Q-aggregation, *Ann. Statist.*, **40** (2012), no. 3, 1878–1905. Zbl 1257.62037 MR 3015047

[10] A. S. Dalalyan and A. B. Tsybakov, Mirror averaging with sparsity priors, *Bernoulli*, **18** (2012), no. 3, 914–944. Zbl 1243.62008 MR 2948907

[11] A. Juditsky and A. Nemirovski, Functional aggregation for nonparametric regression, *Ann. Statist.*, **28** (2000), no. 3, 681–712. Zbl 1105.62338 MR 1792783

[12] A. Juditsky, P. Rigollet, and A. B. Tsybakov, Learning by mirror averaging, *Ann. Statist.*, **36** (2008), no. 5, 2183–2206. Zbl 1274.62288 MR 2458184

[13] V. Koltchinskii, *Oracle inequalities in empirical risk minimization and sparse recovery problems*, Lectures from the 38th Probability Summer School held in Saint-Flour (2008), Lecture Notes in Mathematics, 2033, Springer, 2011. Zbl 1223.91002 MR 2829871

[14] G. Lecué, Lower bounds and aggregation in density estimation, *J. Mach. Learn. Res.*, **7** (2006), 971–981. Zbl 1222.62052 MR 2274393

[15] G. Lecué, Empirical risk minimization is optimal for the convex aggregation problem, *Bernoulli*, **19** (2013), no. 5B, 2153–2166. Zbl 06254557 MR 3160549

[16] G. Lecué and S. Mendelson, On the optimality of the empirical risk minimization procedure for the convex aggregation problem, *Ann. Inst. Henri Poincaré Probab. Stat.*, **49** (2013), no. 1, 288–306, 2013. Zbl 06168767 MR 3060158

[17] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*, Springer, Berlin, 1991. Zbl 0748.60004 MR 1102015

[18] J. Q. Li, *Estimation of Mixture Models*, Ph.D. thesis, Yale University, 1999. MR 2699116

[19] J. Q. Li and A. R. Barron, Mixture density estimation, in *Advances in Neural Information Processing Systems*, 279–285, MIT Press, 1999.

[20] K. Lounici, Generalized mirror averaging and D-convex aggregation, *Math. Methods Statist.*, **16** (2007), no. 3, 246–259. Zbl 1231.62046 MR 2356820

[21] A. Rakhlin, D. Panchenko, and S. Mukherjee, Risk bounds for mixture density estimation, *ESAIM Probab. Stat.*, **9** (2005), 220–229. Zbl 1141.62024 MR 2148968

[22] G. Raskutti, M. J. Wainwright, and B. Yu, Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls, *IEEE Trans. Inform. Theory*, **57** (2011), no. 10, 6976–6994. Zbl 1365.62276 MR 2882274

[23] P. Rigollet, *Oracle inequalities, aggregation and adaptation*, Ph.D. thesis, Université Pierre et Marie Curie - Paris VI, Nov. 2006.

[24] P. Rigollet, Kullback-Leibler aggregation and misspecified generalized linear models, *Ann. Statist.*, **40** (2012), no. 2, 639–665. Zbl 1274.62298 MR 2933661

[25] P. Rigollet and A. Tsybakov, Exponential screening and optimal rates of sparse estimation, *Ann. Statist.*, **39** (2011), no. 2, 731–771. Zbl 1215.62043 MR 2816337

[26] P. Rigollet and A. B. Tsybakov, Linear and convex aggregation of density estimators, *Math. Methods Statist.*, **16** (2007), no. 3, 260–280. Zbl 1231.62057 MR 2356821

[27] A. Tsybakov, *Introduction to Nonparametric Estimation*, Springer Series in Statistics, Springer, 2009. Zbl 1176.62032 MR 2724359

[28] A. B. Tsybakov, Optimal rates of aggregation, in *Learning theory and kernel machines. 16th annual conference on learning theory and 7th Kernel workshop, COLT/Kernel 2003, (Washington DC, USA, 2003). Proceedings*, 303–313, Lect. Notes Comput. Sci., 2777, Springer, 2003. Zbl 1208.62073

[29] A. B. Tsybakov, Aggregation and minimax optimality in high-dimensional estimation, in *Proceedings of the International Congress of Mathematicians (Seoul, August 2014)*, 225–246, KM Kyung Moon Sa, 2014. Zbl 06803199

[30] S. van de Geer and P. Bühlmann, On the conditions used to prove oracle results for the Lasso, *Electron. J. Stat.*, **3** (2009), 1360–1392. Zbl 1327.62425 MR 2576316

[31] Z. Wang, S. Paterlini, F. Gao, and Y. Yang, Adaptive minimax regression estimation over sparse ℓ_q-hulls, *J. Mach. Learn. Res.*, **15** (2014), 1675–1711. Zbl 1319.62016 MR 3225246

[32] D. Xia and V. Koltchinskii, Estimation of low rank density matrices: Bounds in Schatten norms and other distances, *Electron. J. Stat.*, **10** (2016), no. 2, 2717–2745. Zbl 06628775 MR 3546973

[33] Y. Yang, Mixing strategies for density estimation, *Ann. Statist.*, **28** (2000), no. 1, 75–87. Zbl 1106.62322 MR 1762904

[34] A. B. Yuditskiĭ, A. V. Nazin, A. B. Tsybakov, and N. Vayatis, Recursive aggregation of estimators by the mirror descent method with averaging, *Problemy Peredachi Informatsii*, **41** (2005), no. 4, 78–96. Zbl 1123.62044 MR 2198228

A. S. Dalalyan, ENSAE-CREST, Université Paris Saclay,
91764 Palaiseau, France

E-mail: arnak.dalalyan@ensae.fr

M. Sebbar, ENSAE-CREST, Université Paris Saclay,
91764 Palaiseau, France

E-mail: mehdi.sebbar@ensae.fr