

# Optimal exponential bounds for aggregation of density estimators

PIERRE C. BELLEC<sup>1,2</sup>

<sup>1</sup>CREST-ENSAE, 3 avenue Pierre Larousse, 92245 Malakoff Cedex, France.

<sup>2</sup>CMAF, Ecole Polytechnique, Route de Saclay, 91120 Palaiseau, France.

E-mail: pierre.bellec@ensae.fr

We consider the problem of model selection type aggregation in the context of density estimation. We first show that empirical risk minimization is sub-optimal for this problem and it shares this property with the exponential weights aggregate, empirical risk minimization over the convex hull of the dictionary functions, and all selectors. Using a penalty inspired by recent works on the  $Q$ -aggregation procedure, we derive a sharp oracle inequality in deviation under a simple boundedness assumption and we show that the rate is optimal in a minimax sense. Unlike the procedures based on exponential weights, this estimator is fully adaptive under the uniform prior. In particular, its construction does not rely on the sup-norm of the unknown density. By providing lower bounds with exponential tails, we show that the deviation term appearing in the sharp oracle inequalities cannot be improved.

**Keywords:** aggregation; concentration inequality; density estimation; minimax lower bounds; minimax optimality; model selection; sharp oracle inequality

## 1. Introduction

We study the problem of estimation of an unknown density from observations. Let  $(\mathcal{X}, \mu)$  be a measurable space. We are interested in estimating an unknown density  $f$  with respect to the measure  $\mu$  given  $n$  independent observations  $X_1, \dots, X_n$  drawn from  $f$ . We measure the quality of estimation of  $f$  by the  $L^2$  squared distance

$$\|\hat{g} - f\|^2 = \int (f - \hat{g})^2 d\mu = \|\hat{g}\|^2 - 2 \int \hat{g} f d\mu + \|f\|^2, \quad (1.1)$$

for any  $\hat{g} \in L^2(\mu)$  possibly dependent on the data  $X_1, \dots, X_n$ . Since the term  $\|f\|^2$  is constant for all  $\hat{g}$ , we will consider throughout the paper the risk

$$R(\hat{g}) = \|\hat{g}\|^2 - 2 \int \hat{g} f d\mu. \quad (1.2)$$

An estimator  $\hat{g}$  minimizes  $R(\cdot)$  if and only if it minimizes (1.1).

Given  $M$  functions  $f_1, \dots, f_M \in L^2(\mu)$ , we would like to construct a measurable function  $\hat{g}$  of the observations  $X_1, \dots, X_n$  that is almost as good as the best function among  $f_1, \dots, f_M$ . The model may be misspecified, which means that  $f$  may not be one of the functions  $f_1, \dots, f_M$ .

We are interested in deriving oracle inequalities, either in expectation

$$\mathbb{E}R(\hat{g}) \leq C \min_{j=1,\dots,M} R(f_j) + \delta_{n,M},$$

or with high probability, that is, for all  $\varepsilon > 0$ , with probability greater than  $1 - \varepsilon$

$$R(\hat{g}) \leq C \min_{j=1,\dots,M} R(f_j) + \delta_{n,M} + d_{n,M}(\varepsilon),$$

where  $\delta_{n,M}$  is a small quantity and  $d_{n,M}(\cdot)$  is a function of  $\varepsilon$  that we call the deviation term. We are only interested in sharp oracle inequalities, that is, oracle inequalities where the leading constant is  $C = 1$ , since it is essential to derive minimax optimality results.

We consider only deterministic functions for  $f_1, \dots, f_M$ . They cannot depend on the data  $X_1, \dots, X_n$ . A standard application of this setting was introduced in Wegkamp [26]: given  $m + n$  i.i.d. observations drawn from  $f$ , use the first  $m$  observations to build  $M$  estimators  $\hat{f}_1, \dots, \hat{f}_M$ , and in a second step use the remaining  $n$  observations to select the best among the preliminary estimators  $\hat{f}_1, \dots, \hat{f}_M$ . A related problem is selecting the best estimator from a family  $\hat{f}_1, \dots, \hat{f}_M$  where these estimators are built using the same data used for model selection or aggregation. Such problems were recently considered in Dalalyan and Salmon [4] and Dai *et al.* [2] for the regression model with fixed design.

We are also interested in deriving sharp oracle inequalities with prior weights on the model  $\{f_1, \dots, f_M\}$ . To be more precise, for some prior probability distribution  $\pi_1, \dots, \pi_M$  over the finite set  $\{f_1, \dots, f_M\}$  and any  $\varepsilon > 0$ , our estimator  $\hat{f}_n$  should satisfy with probability greater than  $1 - \varepsilon$

$$R(\hat{f}_n) \leq \min_{j=1,\dots,M} \left( R(f_j) + \frac{\beta}{n} \log \frac{1}{\pi_j} \right) + d_{n,M}(\varepsilon), \quad (1.3)$$

for some positive constant  $\beta$  and some deviation term  $d_{n,M}(\cdot)$ . The Mirror Averaging algorithm [6,8] is known to achieve a similar oracle inequality in expectation. The analysis of Juditsky *et al.* [8] shows that the constant  $\beta$  scales linearly with the sup-norm of the unknown density, which is also the case for the results presented here. Model selection techniques with prior weights were used in order to derive sparsity oracle inequalities using sparsity pattern aggregation [6,21,23].

Another related learning problem is that of model selection when the model is finite dimensional with a specific shape, for example a linear span of  $M$  functions or the convex hull of  $M$  functions. This is the aggregation framework and it has received a lot of attention in the last decade to construct adaptive estimators that achieve the minimax optimal rates, especially for the regression problem [11,17,21,23,24] but also for density estimation [10,22,27].

The main contribution of the present paper is the following.

- We provide sharp oracle inequalities and the corresponding tight lower bounds for two procedures: empirical risk minimization over the discrete set  $\{f_1, \dots, f_M\}$  and the penalized procedure (3.2) with the penalty (3.3). Here, tight means that neither the rate nor the deviation term of the sharp oracle inequalities can be improved. The sharp oracle inequalities are given in Theorem 2.2 and Corollary 3.1 and the tight lower bounds are given in Theorems 2.1 and 3.2. These results lead to a definition of minimax optimality in deviation, which is discussed in Section 4.

While proving the above results, we extend several aggregation results that are known for the regression model to the density estimation setting. Let us relate these results of the present paper to the existing literature on the regression model:

- In Theorem 2.2, we derive a sharp oracle inequality in deviation for the empirical risk minimizer over the discrete set  $\{f_1, \dots, f_M\}$ . This is new in the context of density estimation, and an analogous result is known for the regression model [23].
- In Theorem 3.1, we derive a sharp oracle inequality in deviation for penalized empirical risk minimization with the penalty (3.3). With the uniform prior, this yields the correct rate  $(\log M)/n$  of model selection type aggregation. This penalty is inspired by recent works on the  $Q$ -aggregation procedure [3,15] where similar oracle inequalities in deviation were obtained for the regression model. The first sharp oracle inequalities that achieve the correct rate of model selection type aggregation were obtained in expectation for the regression model in [1,27].
- We extend several lower bounds known for the regression model to the density estimation setting. We show that any procedure that selects a dictionary function cannot achieve a better rate than  $\sqrt{(\log M)/n}$  and that the rate of model selection type aggregation is of order  $(\log M)/n$ . We also show that the exponential weights aggregate and the empirical risk minimizer over the convex hull of the dictionary functions cannot be optimal in deviation, with an unavoidable error term of order  $1/\sqrt{n}$ . Earlier results for the regression model can be found in [23,24] for lower bounds on model selection type aggregation and the performance of selectors, while [3,12,14] contain earlier lower bounds on the performance of exponential weights and empirical risk minimization over the convex hull of the dictionary.

An aspect of our results is not present in the previous works on the regression model. In the literature on aggregation in the regression model, lower bounds are proved either in expectation or in probability in the form

$$\mathbb{P}\left(R(\hat{T}_n) > \min_{j=1,\dots,M} R(f_j) + \psi_{n,M}\right) > c, \quad (1.4)$$

for any estimator  $\hat{T}_n$ , a risk function  $R(\cdot)$ , a rate  $\psi_{n,M}$  and some absolute constant  $c > 0$ , usually  $c = 1/2$ . The tight lower bounds presented in Theorems 2.1 and 3.2 contrast with lower bounds of the form (1.4) as they yield for any estimator  $\hat{T}_n$ ,

$$\forall x > 0, \quad \mathbb{P}\left(R(\hat{T}_n) > \min_{j=1,\dots,M} R(f_j) + \psi_{n,M} + \frac{x}{n}\right) > c \exp(-x), \quad (1.5)$$

that is, they provide lower bounds for any probability estimate in an interval  $(0, 1/c)$  where  $c > 0$  is an absolute constant. Moreover, these lower bounds show that the exponential tail of the excess risk of the estimators from Theorems 2.2 and 3.1 cannot be improved. The tools used in the present paper to prove lower bounds of the form (1.5), in particular Lemma 5.1, can be used to prove similar results for regression model. The tight lower bounds of the present paper contrast with the existing literature on the regression model, since to our knowledge, there is no lower bound of the form (1.5) available for regression.

In the regression model with random design, given a class of functions  $G$ , a penalty  $\text{pen}(\cdot)$ , a coefficient  $\nu > 0$  and observations  $(X_1, Y_1), \dots, (X_n, Y_n)$ , penalized empirical risk minimization solves the optimization problem

$$\min_{g \in G} \frac{1}{n} \sum_{i=1}^n (g(X_i) - Y_i)^2 + \nu \text{pen}(g). \quad (1.6)$$

But if the distribution of the design is known, the statistician can compute the quantity  $\mathbb{E}[g(X)^2]$  for all  $g \in G$  and solve the following minimization problem that slightly differs from (1.6):

$$\min_{g \in G} \mathbb{E}[g(X)^2] - \frac{2}{n} \sum_{i=1}^n g(X_i)Y_i + \nu \text{pen}(g). \quad (1.7)$$

In the regression model, the distribution of the design is rarely known so the penalized ERM that solves (1.7) has not received as much attention as the procedure (1.6) when the distribution of the design is not known. The density estimation setting studied in the present paper is closer to the regression setting with known design (1.7) than to the regression setting with unknown design (1.6) studied in [15]. There are differences with respect to the choice of coefficient of the penalty (3.3), and to the form of the empirical process that appears in the analysis. These differences are more thoroughly discussed in Section 3.4.

The paper is organized as follows. In Section 2, we show that empirical risk minimization achieves a sharp oracle inequality with slow rate, but this rate cannot be improved among selectors. Two classical estimators, the exponential weights aggregate and empirical risk minimization over the convex hull of the dictionary functions, are shown to be suboptimal in deviation. In Section 3, we define a penalized procedure that achieves the optimal rate  $\frac{\log M}{n}$  in deviation, and we provide a lower bound that shows that neither the rate nor the deviation term can be improved. Section 4 proposes a definition of minimax optimality in deviation and shows that it is satisfied by the procedures given in Sections 2 and 3. Section 5 is devoted to the proofs.

## 2. Sub-optimality of selectors, ERM and exponential weights

### 2.1. Selectors

Define a selector as a function of the form  $f_{\hat{J}}$  where  $\hat{J}$  is measurable with respect to  $X_1, \dots, X_n$  with values in  $\{1, \dots, M\}$ . It was shown in the regression framework [8,23] that selectors are suboptimal and cannot achieve a better rate than  $\sigma \sqrt{\frac{\log M}{n}}$  where  $\sigma^2$  is the variance of the regression noise. The following theorem extends this lower bound for selectors to density estimation. The underlying measure  $\mu$  is the Lebesgue measure on  $\mathbf{R}^d$  for  $d \geq 1$ .

**Theorem 2.1 (Lower bounds for selectors).** *Let  $L > 0$ , and  $M \geq 2, n \geq 1, d \geq 1$  be integers. Let  $\mathcal{F}$  be the class of all densities  $f$  with respect to the Lebesgue measure on  $\mathbf{R}^d$  such that*

$\|f\|_\infty \leq L$ . Let  $x \geq 0$  satisfying

$$\frac{\log(M) + x}{n} < 3.$$

Then there exist  $f_1, \dots, f_M \in L^2(\mathbf{R}^d)$  with  $\|f_j\|_\infty \leq L$  such that the following lower bound holds:

$$\inf_{\hat{S}_n} \sup_{f \in \mathcal{F}} \mathbb{P}_f \left( \|\hat{S}_n - f\|^2 - \inf_{j=1, \dots, M} \|f_j - f\|^2 \geq \frac{L}{\sqrt{3}} \sqrt{\frac{x + \log M}{n}} \right) \geq \frac{1}{24} \exp(-x),$$

where  $\mathbb{P}_f$  denotes the probability with respect to  $n$  i.i.d. observations with density  $f$  and the infimum is taken over all selectors  $\hat{S}_n$ .

The proof of Theorem 2.1 is given in Section 5. It can be extended to other measures as soon as the underlying measurable space allows the construction of an orthogonal system such as the one described in Proposition 5.4 below.

For any  $g \in L^2(\mu)$ , define the empirical risk

$$R_n(g) = \|g\|^2 - \frac{2}{n} \sum_{j=1}^M g(X_j). \quad (2.1)$$

The empirical risk (2.1) is an unbiased estimator of the risk (1.2). In order to explain the idea behind the proof of our main result described in Theorem 3.1, it is useful to prove the following oracle inequality for the empirical risk minimizer over the discrete set  $\{f_1, \dots, f_M\}$ .

**Theorem 2.2.** Assume that the functions  $f_1, \dots, f_M \in L^2(\mu)$  satisfy  $\|f_j\|_\infty \leq L_0$  for all  $j = 1, \dots, M$ . Define

$$\hat{J} \in \operatorname{argmin}_{j=1, \dots, M} \left( \|f_j\|^2 - \frac{2}{n} \sum_{i=1}^n f_j(X_i) \right).$$

Then for any  $x > 0$ , with probability greater than  $1 - \exp(-x)$ ,

$$R(f_{\hat{J}}) \leq \min_{j=1, \dots, M} R(f_j) + L_0 \left( 4\sqrt{2} \sqrt{\frac{x + \log M}{n}} + \frac{8(x + \log M)}{3n} \right).$$

Together with Theorem 2.1, Theorem 2.2 shows that empirical risk minimization is optimal among selectors. Unlike the oracle inequality of Theorem 3.1 below, this result applies for any density  $f$ , with possibly  $\|f\|_\infty = \infty$ . Its proof relies on the concentration of  $R_n(g) - R(g)$  around 0 for fixed functions  $g$  with  $\|g\|_\infty \leq L_0$ .

**Proof of Theorem 2.2.** We will use the following notation that is common in the literature on empirical processes. For any  $g \in L^2(\mu)$ , define

$$\begin{aligned} P g &= \int g f d\mu, \\ P_n g &= \frac{1}{n} \sum_{i=1}^n g(X_i). \end{aligned} \tag{2.2}$$

With this notation, the difference between the real risk (1.2) and the empirical risk (2.1) can be rewritten

$$R(g) - R_n(g) = (P - P_n)(-2g). \tag{2.3}$$

Let  $J^*$  be such that  $R(f_{J^*}) = \min_{j=1, \dots, M} R(f_j)$ . The definition of  $\hat{J}$  yields  $R_n(f_{\hat{J}}) \leq R_n(f_{J^*})$ . Using (2.3), it can be rewritten

$$R(f_{\hat{J}}) - R(f_{J^*}) \leq (P - P_n)(-2f_{\hat{J}} + 2f_{J^*}).$$

We can control the right-hand side of the last display using the concentration inequality (5.2) with a union bound over  $j = 1, \dots, M$ . For any  $t > 0$ , with probability greater than  $1 - M \exp(-t)$ ,

$$\begin{aligned} (P - P_n)(-2f_{\hat{J}} + 2f_{J^*}) &\leq \max_{j=1, \dots, M} (P - P_n)(-2f_j + 2f_{J^*}) \\ &\leq \sigma \sqrt{\frac{2t}{n}} + \frac{8L_0 t}{3n}, \end{aligned}$$

where  $\sigma^2 = \max_{j=1, \dots, M} P(-2f_j + 2f_{J^*})^2 \leq 16L_0^2$ . Setting  $x = t - \log M$  yields the desired oracle inequality.  $\square$

By inspecting the short proof above, we see that the slow rate term  $\sqrt{\frac{x + \log M}{n}}$  comes from the variance term in the concentration inequality (5.2).

We can draw two conclusions from Theorems 2.1 and 2.2.

- In order to achieve faster rates than  $\sqrt{\frac{\log M}{n}}$ , we need to look for estimators taking values beyond the discrete set  $\{f_1, \dots, f_M\}$ . In Section 3, we will consider estimators taking values in the convex hull of this discrete set.
- The proof of Theorem 2.2 suggests that a possible way to derive an oracle inequality with fast rates is to cancel the variance term in the concentration inequality (5.2). In order to do this, we need some positive gain on the empirical risk of our estimator. Namely, for some oracle  $J^*$  we would like our estimator  $\hat{f}_n$  to satisfy  $R_n(\hat{f}_n) \leq R_n(f_{J^*})$  minus some positive value. This value is given by the strong convexity of the empirical objective in Proposition 3.1.

Define the simplex in  $\mathbf{R}^M$ :

$$\Lambda^M = \left\{ \theta \in \mathbf{R}^M, \sum_{j=1}^M \theta_j = 1, \forall j = 1, \dots, M, \theta_j \geq 0 \right\}. \quad (2.4)$$

Given a finite set or *dictionary*  $\{f_1, \dots, f_M\}$ , define for any  $\theta \in \Lambda^M$

$$f_\theta = \sum_{j=1}^M \theta_j f_j. \quad (2.5)$$

In particular,  $f_j = f_{e_j}$  where  $e_1, \dots, e_M$  are the vectors of the canonical basis in  $\mathbf{R}^M$ .

Two classical estimators, the ERM over the convex hull of  $f_1, \dots, f_M$  and the exponential weights aggregate, are known to be sub-optimal in the regression setting [3,12–14]. In the following we show that the same conclusions hold for density estimation with the  $L^2$  risk.

## 2.2. ERM over the convex hull

A first natural estimator valued in the convex hull of the dictionary functions is the ERM. However, as in the regression setting [12], this estimator is suboptimal with an unavoidable error term or order  $1/\sqrt{n}$ .

**Proposition 2.1.** *Let  $\mathcal{X} = \mathbf{R}$  and  $\mu$  be the Lebesgue measure on  $\mathbf{R}$ . There exist absolute constants  $C_0, C_1, C_2, C_3 > 0$  such that the following holds. Let  $L > 0$ . For any integer  $n \geq 1$ , there exist a density  $f$  bounded by  $L$  and a dictionary  $\{f_1, \dots, f_M\}$  of functions bounded by  $2L$ , with  $C_0\sqrt{n} \leq M \leq C_1\sqrt{n}$ , such that with probability greater than  $1 - 12\exp(-C_2M)$ ,*

$$\|f_{\hat{\theta}^{\text{ERM}}} - f\|^2 \geq \min_{j=1,\dots,M} \|f_j - f\|^2 + \frac{C_3L}{\sqrt{n}},$$

where  $\hat{\theta}^{\text{ERM}} := \arg \min_{\theta \in \Lambda^M} R_n(f_\theta)$ .

The proof of Proposition 2.1 can be found in Section 5.5.2.

## 2.3. Exponential weights

The exponential weights aggregate is known to achieve optimal oracle inequalities in expectation when the temperature parameter  $\beta > 0$  is chosen carefully [5,8,16]. Given prior weights  $(\pi_1, \dots, \pi_M)^T \in \Lambda^M$ , it can be defined as follows:

$$\hat{f}_\beta^{\text{EW}} = \sum_{j=1}^M \hat{\theta}_j^{\text{EW},\beta} f_j, \quad \hat{\theta}^{\text{EW},\beta} \in \Lambda^M, \quad \hat{\theta}_j^{\text{EW},\beta} \propto \pi_j \exp\left(-\frac{n}{\beta} R_n(f_j)\right).$$

The following proposition shows that it is suboptimal in deviation for any temperature, with a error term of order at least  $1/\sqrt{n}$ . This phenomenon was observed in the regression setting [3, 12], and Proposition 2.2 shows that it also holds for density estimation. As opposed to [3], the following lower bound requires only 3 dictionary functions.

**Proposition 2.2.** *There exist absolute constants  $C_0, C_1, N_0 > 0$  such that the following holds. Let  $\mathcal{X} = \mathbf{R}$  and  $\mu$  be the Lebesgue measure on  $\mathbf{R}$ . For all  $n \geq N_0, L > 0$ , there exist a probability density  $f$  with respect to  $\mu$ , a dictionary  $\{f_1, f_2, f_3\}$  and prior weights  $(\pi_1, \pi_2, \pi_3) \in \Lambda^3$  such that with probability greater than  $C_0$ ,*

$$\|\hat{f}_\beta^{\text{EW}} - f\|^2 \geq \min_{j=1,2,3} \|f_j - f\|^2 + \frac{C_1 L}{\sqrt{n}}.$$

Furthermore,  $\|f\|_\infty \leq L$ , and  $\|f_j\|_\infty \leq 3L$  for  $j = 1, 2, 3$ .

The following proposition shows that the optimality in expectation cannot hold if the temperature is below a constant, extending a result from [12] to the density estimation setting.

**Proposition 2.3.** *Let  $\mathcal{X} = \mathbf{R}$  and  $\mu$  be the Lebesgue measure on  $\mathbf{R}$ . There exist absolute constants  $c_0, c_1, c_2 > 0$  such that the following holds. Let  $L > 0$ . For any odd integer  $n \geq c_0$ , there exist a probability density  $f$  with respect to  $\mu$  with  $\|f\|_\infty \leq L$ , and a dictionary  $\{f_1, f_2\}$  with  $f_j : \mathcal{X} \rightarrow \mathbf{R}$  and  $\|f_j\|_\infty \leq L$  for  $j = 1, 2$  for which the following holds:*

$$\mathbb{E} \|\hat{f}_\beta^{\text{EW}} - f\|^2 \geq \min_{j=1,2} \|f_j - f\|^2 + \frac{c_2 L}{\sqrt{n}} \text{ if } \beta \leq c_1 L.$$

The proofs of Propositions 2.2 and 2.3 can be found in Section 5.5.3.

### 3. Optimal exponential bounds for a penalized procedure

#### 3.1. From strong convexity to a sharp oracle inequality

In this section, we derive a sharp oracle inequality for the estimator  $f_{\hat{\theta}}$  where  $\hat{\theta}$  is defined in (3.2). Define the empirical objective  $H_n$  and the estimator  $\hat{\theta}$  by

$$H_n(\theta) = \left( \|f_\theta\|^2 - \frac{2}{n} \sum_{i=1}^n f_\theta(X_i) \right) + \frac{1}{2} \text{pen}(\theta) + \frac{\beta}{n} \sum_{j=1}^M \theta_j \log \frac{1}{\pi_j}, \quad (3.1)$$

$$\hat{\theta} \in \underset{\theta \in \Lambda^M}{\text{argmin}} H_n(\theta), \quad (3.2)$$

for some positive constant  $\beta$  and

$$\forall \theta \in \Lambda^M, \quad \text{pen}(\theta) = \sum_{j=1}^M \theta_j \|f_\theta - f_j\|^2. \quad (3.3)$$



The simplex  $\Lambda^M$  and  $f_\theta$  are defined in (2.4) and (2.5).

The term

$$\frac{\beta}{n} \sum_{j=1}^M \theta_j \log \frac{1}{\pi_j}$$

is a penalty that assigns different weights to the functions  $f_j$  according to some prior knowledge given by  $\pi_1, \dots, \pi_M$ , in order to achieve an oracle inequality such as (1.3).

The penalty (3.3) as well as the present procedure are inspired by recent works on Q-aggregation in regression models [3,15,20]. The choice of the coefficient  $\frac{1}{2}$  for the penalty (3.3) is explained in Remark 3.1 below. An intuitive interpretation of the penalty (3.3) can be as follows. A point  $f_\theta$  is in the convex hull of  $\{f_1, \dots, f_M\}$  if and only if it is the expectation of a random variable taking values in  $\{f_1, \dots, f_M\}$ . The penalty (3.3) can be seen as the variance of such a random variable whose distribution is given by  $\theta$ . More precisely, let  $\eta$  be a random variable with  $\mathbb{P}(\eta = j) = \theta_j$  for all  $j = 1, \dots, M$ . Denote by  $\mathbb{E}_\theta$  the expectation with respect to the random variable  $\eta$ . Then  $\mathbb{E}_\theta[f_\eta] = f_\theta$  and

$$\text{pen}(\theta) = \mathbb{E}_\theta \|f_\eta - \mathbb{E}_\theta[f_\eta]\|^2,$$

which is the variance of the random point  $f_\eta$ . The penalty (3.3) vanishes at the extreme points:

$$\forall j = 1, \dots, M, \quad \text{pen}(e_j) = 0,$$

and  $\text{pen}(\theta)$  increases as  $\theta$  moves away from an extreme point  $e_j$ . Thus we convexify the optimization problem over the discrete set  $\{f_1, \dots, f_M\}$  by considering the convex set  $\{\mathbb{E}_\theta[f_\eta], \theta \in \Lambda^M\}$  which is exactly the convex hull of  $\{f_1, \dots, f_M\}$ , and we penalize by the variance of the random point  $f_\eta$ .

It is also possible to describe the level sets of the penalty (3.3). Assume only in this paragraph that the Gram matrix of  $f_1, \dots, f_M$  is invertible and let  $c \in L^2(\mu)$  be in the linear span of  $f_1, \dots, f_M$  such that for all  $j = 1, \dots, M$ ,  $\int 2cf_j d\mu = \|f_j\|^2$ . Then simple algebra yields

$$\text{pen}(\theta) = \|c\|^2 - \|c - f_\theta\|^2.$$

Thus the level sets of the penalty (3.3) are euclidean balls centered at  $c$ .

Last, note that  $f_{\hat{\theta}}$  coincides with the Q-aggregation procedure from [3] since

$$\left( \|f_{\hat{\theta}}\|^2 - \frac{2}{n} \sum_{i=1}^n f_{\hat{\theta}}(X_i) \right) + \frac{1}{2} \text{pen}(\theta) = R_n(\theta) + \frac{1}{2} \text{pen}(\theta) = \frac{1}{2} \left( R_n(\theta) + \sum_{j=1}^M \theta_j R_n(f_j) \right).$$

We propose an estimator  $f_{\hat{\theta}}$  based on penalized empirical risk minimization over the simplex, with  $\hat{\theta}$  defined in (3.2). This estimator satisfies the following oracle inequality.

**Theorem 3.1.** Assume that the functions  $f_1, \dots, f_M$  satisfy  $\|f_j\|_\infty \leq L_0$  for all  $j = 1, \dots, M$ , and assume that the unknown density  $f$  satisfies  $\|f\|_\infty \leq L$ . Let  $\hat{\theta}$  be defined in (3.2) with

$$\beta = 4L + \frac{8L_0}{3}.$$

Then for any  $x > 0$ , with probability greater than  $1 - \exp(-x)$ ,

$$R(f_{\hat{\theta}}) \leq \min_{j=1, \dots, M} \left( R(f_j) + \frac{\beta}{n} \log \frac{1}{\pi_j} \right) + \frac{\beta x}{n}. \quad (3.4)$$

The following proposition specifies the property of strong convexity of the objective function  $H_n(\cdot)$  defined in (3.1), which is key to prove Theorem 3.1.

**Proposition 3.1 (Strong convexity of  $H_n$ ).** Let  $H_n$  and  $\hat{\theta}$  be defined by (3.1) and (3.2), respectively. Then for any  $\theta \in \Lambda^M$ ,

$$H_n(\hat{\theta}) \leq H_n(\theta) - \frac{1}{2} \|f_\theta - f_{\hat{\theta}}\|^2. \quad (3.5)$$

For any  $\theta \in \Lambda^M$ , empirical risk minimization only grants the simple inequality

$$R_n(\hat{\theta}) \leq R_n(\theta),$$

but with Proposition 3.1 we gain the extra term  $\frac{1}{2} \|f_\theta - f_{\hat{\theta}}\|^2$ . To prove Theorem 3.1, we will use this extra term to compensate the variance term of the concentration inequality (5.3). Strong convexity plays an important role in our proofs, and we believe that our arguments would not work for loss functions that are not strongly convex such as the Hellinger distance, the Total Variation distance or the Kullback–Leibler divergence.

The proof of Proposition 3.1 is given in Section 5.3. We now give the proof of our main result, which is close to the proof of Theorem 2.2 except that we leverage the strong convexity of the empirical objective  $H_n$ .

**Proof of Theorem 3.1.** Note that  $\text{pen}(e_j) = 0$  for  $j = 1, \dots, M$  and let

$$J^* \in \operatorname{argmin}_{j=1, \dots, M} \left( \|f_j\|^2 - 2 \int f_j f d\mu + \frac{\beta}{n} \log \frac{1}{\pi_j} \right) = \operatorname{argmin}_{j=1, \dots, M} \mathbb{E}[H_n(e_j)].$$

Using (3.5) of Proposition 3.1

$$\begin{aligned} H_n(\hat{\theta}) - H_n(e_{J^*}) &\leq -\frac{1}{2} \|f_{J^*} - f_{\hat{\theta}}\|^2, \\ R_n(\hat{\theta}) + \frac{\beta}{n} \sum_{j=1}^M \hat{\theta}_j \log \frac{1}{\pi_j} - R_n(e_{J^*}) - \frac{\beta}{n} \log \frac{1}{\pi_{J^*}} &\leq -\frac{1}{2} \|f_{J^*} - f_{\hat{\theta}}\|^2 - \frac{1}{2} \text{pen}(\hat{\theta}) \\ &= -\frac{1}{2} \sum_{j=1}^M \hat{\theta}_j \|f_j - f_{J^*}\|^2, \end{aligned}$$

where we used Proposition 5.1 with  $g = f_{J^*}$  for the last display. Using (2.3), we get

$$R(f_{\hat{\theta}}) - R(f_{J^*}) - \frac{\beta}{n} \log \frac{1}{\pi_{J^*}} \leq Z_n,$$

where

$$Z_n = (P - P_n)(-2f_{\hat{\theta}} + 2f_{J^*}) - \frac{\beta}{n} \sum_{j=1}^M \hat{\theta}_j \log \frac{1}{\pi_j} - \frac{1}{2} \sum_{j=1}^M \hat{\theta}_j \|f_j - f_{J^*}\|^2$$

and the notation  $P$  and  $P_n$  is defined in (2.2) and (2.3). The quantity  $Z_n$  is affine in  $\theta$  and an affine function over the simplex is maximized at a vertex, so almost surely,

$$\begin{aligned} Z_n &\leq \max_{\theta \in \Lambda^M} \left( -2(P - P_n)(f_{\theta} - f_{J^*}) - \frac{1}{2} \sum_{j=1}^M \theta_j \|f_{J^*} - f_j\|^2 - \frac{\beta}{n} \sum_{j=1}^M \theta_j \log \frac{1}{\pi_j} \right) \\ &= \max_{k=1, \dots, M} \left( -2(P - P_n)(f_k - f_{J^*}) - \frac{1}{2} \|f_k - f_{J^*}\|^2 - \frac{\beta}{n} \log \frac{1}{\pi_k} \right). \end{aligned} \quad (3.6)$$

Let  $k = 1, \dots, M$  fixed. Applying Proposition 5.3 with  $g = -2(f_k - f_{J^*})$  and  $\pi = \pi_k$  yields

$$\mathbb{P} \left( -2(P - P_n)(f_k - f_{J^*}) - \frac{1}{2} \|f_k - f_{J^*}\|^2 - \frac{\beta}{n} \log \frac{1}{\pi_k} > \frac{\beta x}{n} \right) \leq \pi_k \exp(-x).$$

To complete the proof, we use a union bound on  $k = 1, \dots, M$  together with  $\sum_{j=1}^M \pi_j = 1$  and (3.6):

$$\mathbb{P} \left( Z_n > \frac{\beta x}{n} \right) \leq \sum_{k=1}^M \pi_k \exp(-x) = \exp(-x). \quad \square$$

**Remark 3.1 (Choice of the coefficient of the penalty (3.3)).** Let  $\nu \in (0, 1)$ . With minor modifications to the proof of Theorem 3.1, it can be shown that the oracle inequality (3.4) still holds with

$$\begin{aligned} \beta &= \frac{2L}{\min(\nu, 1 - \nu)} + \frac{8L_0}{3}, \\ H_n(\theta) &= \left( \|f_{\theta}\|^2 - \frac{2}{n} \sum_{i=1}^n f_{\theta}(X_i) \right) + \nu \text{pen}(\theta) + \frac{\beta}{n} \sum_{j=1}^M \theta_j \log \frac{1}{\pi_j}, \\ \hat{\theta} &\in \underset{\theta \in \Lambda^M}{\operatorname{argmin}} H_n(\theta). \end{aligned}$$

The oracle inequality (3.4) is best when  $\beta$  is small. Thus, the choice  $\nu = \frac{1}{2}$  is natural since it minimizes the value of  $\beta$ .

The optimization problem (3.2) is a quadratic program, for which efficient algorithms exist. We refer to [3], Section 4, for an analysis of the statistical performance of an algorithm that approximately solves a optimization problem similar to (3.2) in the regression setting.

The estimator  $\hat{\theta}$  of Theorem 3.1 is not adaptive since its construction relies on  $L$ , an upper bound of the sup-norm of the unknown density. However, in the case of the uniform prior  $\pi_j = 1/M$  for all  $j = 1, \dots, M$ , Corollary 3.1 below provides an estimator which is fully adaptive: its construction depends only on the functions  $f_1, \dots, f_M$  and the data  $X_1, \dots, X_n$ . A similar adaptivity property was observed in [15] in the regression setting.

**Corollary 3.1 (Adaptive estimator).** *Assume that the functions  $f_1, \dots, f_M$  satisfy  $\|f_j\|_\infty \leq L_0$  for all  $j = 1, \dots, M$ , and assume that the unknown density  $f$  satisfies  $\|f\|_\infty \leq L$ . Let*

$$\hat{\theta} \in \operatorname{argmin}_{\theta \in \Lambda^M} \left( \|f_\theta\|^2 - \frac{2}{n} \sum_{i=1}^n f_\theta(X_i) \right) + \frac{1}{2} \operatorname{pen}(\theta). \quad (3.7)$$

Then for any  $x > 0$ , with probability greater than  $1 - \exp(-x)$ ,

$$R(f_{\hat{\theta}}) \leq \min_{j=1, \dots, M} R(f_j) + \left( 4L + \frac{8L_0}{3} \right) \frac{\log(M) + x}{n}.$$

**Proof.** With the uniform prior,  $\pi_j = 1/M$  for all  $j = 1, \dots, M$ , the quantity

$$\frac{\beta}{n} \sum_{j=1}^M \theta_j \log \frac{1}{\pi_j} = \frac{\beta}{n} \log M$$

is independent of  $\theta \in \Lambda^M$ . The minimizer (3.7) is also a minimizer of the empirical objective (3.1) used in Theorem 3.1. Thus, the estimator  $f_{\hat{\theta}}$  satisfies (3.4) which completes the proof.  $\square$

Corollary 3.1 is in contrast to methods related to exponential weights such as the mirror averaging algorithm from [8] as these methods rely on the knowledge of the sup-norm of the unknown density. The method presented here is an improvement in two aspects. First, the estimator of Corollary 3.1 is fully data-driven. Second, the sharp oracle inequality is satisfied not only in expectation, but also in deviation.

However, the method of Theorem 3.1 loses this adaptivity property when a non-uniform prior is used, and we do not know if it is possible to build an optimal and fully adaptive estimator for non-uniform priors.

### 3.2. A lower bound with exponential tails

The following lower bound shows that the sharp oracle inequality of Corollary 3.1 cannot be improved both in the rate and in the tail of the deviation.

**Theorem 3.2 (Lower bounds with optimal deviation term).** *Let  $M \geq 2, n \geq 1$  be two integers and let a real number  $x \geq 0$  satisfy*

$$\frac{\log(M) + x}{n} < 3.$$

*Let  $L > 0$  and  $d \geq 1$ . Let  $\mathcal{F}$  be the class of densities  $f$  with respect to the Lebesgue measure on  $\mathbf{R}^d$  such that  $\|f\|_\infty \leq L$ .*

*Then there exist  $M$  functions  $f_1, \dots, f_M$  in  $L^2(\mathbf{R}^d)$  with  $\|f_j\|_\infty \leq L$  satisfying*

$$\inf_{\hat{T}_n} \sup_{f \in \mathcal{F}} \mathbb{P}_f \left( \|\hat{T}_n - f\|^2 - \min_{j=1, \dots, M} \|f_j - f\|^2 > \frac{L}{24} \left( \frac{\log(M) + x}{n} \right) \right) \geq \frac{1}{24} \exp(-x),$$

*where the infimum is taken over all estimators  $\hat{T}_n$  and  $\mathbb{P}_f$  denotes the probability with respect to  $n$  i.i.d. observations with density  $f$ .*

Notice that the restriction  $\frac{\log(M)+x}{n} < 3$  is natural since the estimator  $\hat{T}_n^* \equiv 0$  achieves a constant error term and is optimal in the region  $\frac{\log(M)+x}{n} > c$  for some absolute constant  $c$ . Indeed, as the unknown density satisfies  $\|f\|_\infty \leq L$ , we have with probability 1:

$$\begin{aligned} \|\hat{T}_n^* - f\|^2 &= \|f\|^2 \leq L \leq \inf_{j=1, \dots, M} \|f - f_j\|^2 + L, \\ R(\hat{T}_n^*) &\leq \inf_{j=1, \dots, M} R(f_j) + L. \end{aligned} \tag{3.8}$$

Thus it is impossible to get the lower bound of Theorem 3.2 for arbitrarily large  $\frac{x+\log M}{n}$ .

### 3.3. Weighted loss and unboundedness

The previous strategy based on penalized risk minimization over the simplex can be applied to handle unbounded densities or unbounded dictionary functions, if we use a weighted loss.

Let  $w : \mathcal{X} \rightarrow \mathbf{R}^+$  be a measurable function with respect to  $\mu$ . Define the norm (or semi-norm if  $w$  is zero on a set of positive measure)

$$\|g\|_w^2 = \int g^2 w d\mu, \quad \forall g \in L^2(\mu).$$

Then we can define the estimator  $f_{\hat{\theta}}$  where

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Lambda^M} V_n(\theta), \quad V_n(\theta) = P_n \left( \|f_\theta\|_w^2 - \frac{2}{n} \sum_{i=1}^n f_\theta(X_i) w(X_i) + \frac{1}{2} \sum_{j=1}^M \theta_j \|f_j - f_\theta\|_w^2 \right).$$

The function  $V_n$  is strongly convex with respect to the new norm  $\|\cdot\|_w^2$ . As in the proof of Theorem 3.1, this leads to

$$\|f_{\hat{\theta}} - f\|_w^2 \leq \|f_{J^*} - f\|_w^2 + \max_{k=1,\dots,M} \delta_k,$$

$$\delta_k := (P - P_n)(-2(f_{J^*} - f_k)w) - \frac{1}{2}\|f_{J^*} - f_k\|_w^2.$$

If for some  $L, L_0 > 0$ ,  $\|wf\|_\infty \leq L$  and  $\max_{j=1,\dots,M} \|wf_j\|_\infty \leq L_0$ , then

$$\delta_k \leq -2(P - P_n)((f_k - f_{J^*})w) - \frac{1}{2L} \mathbb{E}[(f_k(X) - f_{J^*}(X))^2 w(X)^2].$$

We apply (5.3) to the random variables  $(f_k - f_{J^*})(X_i)w(X_i)$ , which are almost surely bounded by  $L_0$ . Using the union bound on  $k = 1, \dots, M$ , we obtain  $\max_{k=1,\dots,M} \delta_k \leq \beta(x + \log M)/n$  with probability greater than  $1 - \exp(-x)$ . and thus

$$\|f_{\hat{\theta}} - f\|_w^2 \leq \|f_{J^*} - f\|_w^2 + \beta \left( \frac{x + \log M}{n} \right),$$

where  $\beta = c(L + L_0)$  for some numerical constant  $c > 0$ .

### 3.4. Differences and similarities with regression problems

Here, we discuss differences and similarities between aggregation of density and regression estimators. Some notation is needed in order to compare these settings.

We first define some notation related to the Density Estimation (DE) framework studied in the present paper. Let  $X$  be a random variable with density  $f$  absolutely continuous with respect to the measure  $\mu$ , let  $\mathcal{D}^{\text{DE}} = \{f_1, \dots, f_M\}$  be a subset of  $L^2(\mu)$  and define for all  $g \in L^2(\mu)$  and  $x \in \mathcal{X}$ ,

$$\|g\|^2 = \int g^2 d\mu, \quad l_g^{\text{DE}}(x) = \|g\|^2 - 2g(x),$$

$$g^* = \operatorname{argmin}_{g \in \mathcal{D}^{\text{DE}}} \|g - f\|^2 = \operatorname{argmin}_{g \in \mathcal{D}^{\text{DE}}} [l_g^{\text{DE}}(X)].$$

Given  $n$  i.i.d. observations  $X_1, \dots, X_n$  and some fixed function  $g$ , one can use the empirical risk  $P_n(l_g^{\text{DE}}) = \sum_{i=1}^n (1/n) l_g^{\text{DE}}(X_i)$ .

We now define similar notation for the regression problem with the  $L^2$  loss. Let  $(X, Y)$  be a random couple valued in  $\mathcal{X} \times \mathbf{R}$ , let  $P_X$  be the probability measure of  $X$ , let  $f$  be the true regression function defined by  $f(x) = \mathbb{E}[Y|X = x]$ , let  $\mathcal{D}^{\text{R}} = \{f_1, \dots, f_M\}$  be a subset of  $L^2(P_X)$  and define for all  $g \in L^2(P_X)$ ,

$$\|g\|_{P_X}^2 = \mathbb{E}[g(X)^2], \quad g^* = \operatorname{argmin}_{g \in \mathcal{D}^{\text{R}}} \|g - f\|_{P_X}^2.$$

For Regression with Unknown Design (RUD) that is, when the distribution of the design  $X$  is not known to the statistician, a natural choice for the loss function  $l_g$  is

$$l_g^{\text{RUD}}(x, y) = (g(x) - y)^2, \quad \forall x, y \in \mathcal{X} \times \mathbf{R},$$

and the oracle  $g^*$  defined above satisfies  $g^* = \arg \min_{g \in \mathcal{D}^{\mathbf{R}}} \mathbb{E}[l_g^{\text{RUD}}(X, Y)]$ . For Regression with Known Design (RKD), the quantity  $\|g\|_{P_X}^2$  is accessible for all  $g$ . Thus, we can define the loss

$$l_g^{\text{RKD}}(x, y) = \|g\|_{P_X}^2 - 2g(x)y, \quad \forall x, y \in \mathcal{X} \times \mathbf{R},$$

and the oracle  $g^*$  satisfies  $g^* = \arg \min_{g \in \mathcal{D}^{\mathbf{R}}} \mathbb{E}[l_g^{\text{RKD}}(X, Y)]$ . Thus, two natural functions  $l_g$  arise in the regression context, depending on whether the distribution of the design is known or unknown. Given  $n$  i.i.d. observations  $(X_i, Y_i)$  with the same distribution as  $(X, Y)$ , the empirical quantities  $P_n(l_g^{\text{RUD}})$  and  $P_n(l_g^{\text{RKD}})$  can be used to infer the true regression function  $f$ . An estimator constructed using the quantity  $P_n(l_g^{\text{RKD}})$  is used, for example, in [24] for the problem of linear and convex aggregation.

#### Linear or quadratic empirical process

The empirical process  $(P_n - P)(l_g - l_{g^*})$  indexed by  $g$  plays an important role in the proofs of Theorems 2.2 and 3.1. This empirical process also appears in the analysis [15] for regression with unknown design with the loss  $l_g^{\text{RUD}}$ . For density estimation and regression with known design, this empirical process is linear in  $g$ :

$$\begin{aligned} (P_n - P)(l_g^{\text{DE}} - l_{g^*}^{\text{DE}}) &= -2(P_n - P)(g - g^*), \\ (P_n - P)(l_g^{\text{RKD}} - l_{g^*}^{\text{RKD}}) &= -2(P_n - P)[(g - g^*)\dot{y}], \end{aligned}$$

where the function  $\dot{y}(\cdot)$  above is defined by  $\forall x, y \in \mathcal{X} \times \mathbf{R}, \dot{y}(x, y) = y$ . For regression when the design is unknown, the empirical process is quadratic in the class member  $g$ . To control the behavior of this quadratic empirical process, the contraction principle is used in [15], whereas this principle is not needed for density estimation or regression when the distribution of the design is known.

#### The penalty (3.3) and its coefficient

In the regression problem when the distribution is known, given a dictionary of potential regression functions  $\{f_1, \dots, f_M\}$ , the quantity

$$\sum_{j=1}^M \theta_j \|f_j - f_\theta\|_{P_X}^2, \quad (3.9)$$

is accessible and a procedure similar to the one proposed in Theorem 3.1 and Corollary 3.1 can be constructed, with the penalty coefficient  $1/2$  which is a natural choice as explained in

Remark 3.1. For regression with unknown design, the above penalty cannot be computed: the procedure [15] for the  $L^2$  loss is the estimator  $f_{\hat{\theta}}$  where

$$\begin{aligned}\hat{\theta} &= \operatorname{argmin}_{\theta \in \Lambda^M} (P_n(I_{f_\theta}^{\text{RUD}}) + \nu P_n(f_j - f_\theta)^2) \\ &= \operatorname{argmin}_{\theta \in \Lambda^M} \left( \frac{1}{n} \sum_{i=1}^n (Y_i - f_\theta(X_i))^2 + \frac{\nu}{n} \sum_{i=1}^n (f_j - f_\theta)^2(X_i) \right),\end{aligned}$$

for some coefficient  $\nu \in (0, 1)$  and where we chose the uniform prior for clarity. Thus, the procedure [15] can be formulated as a penalized procedure where the penalty is the empirical counterpart of (3.9) with the coefficient  $\nu$ . Although  $1/2$  is a natural choice for regression with known design and density estimation, for regression with unknown design the expression of the optimal coefficient is more intricate [15], minimize  $\beta$  in (1.4).

#### *Sketch of proof for the regression model with known design*

In order to show the similarities between density estimation and regression problems when the design is known, we now give the main ideas to derive an oracle inequality similar to Corollary 3.1 for regression with known design. Note that the framework studied in [15] does not cover the estimator defined below, since the function  $I_g^{\text{RKD}}$  depends on the quantity  $\|g\|_{P_X}^2$ . Given  $n$  i.i.d. observations  $(X_1, Y_1), \dots, (X_n, Y_n)$ , define

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Lambda^M} V_n(\theta), \quad V_n(\theta) = P_n(I_{f_\theta}^{\text{RKD}}) + \frac{1}{2} \sum_{j=1}^M \theta_j \|f_j - f_\theta\|_{P_X}^2.$$

Analogously to the argument of Proposition 3.1, we note that the function  $V_n$  is strongly convex and  $V_n(\hat{\theta}) \leq V_n(e_{J^*}) - \frac{1}{2} \|f_{J^*} - f_{\hat{\theta}}\|_{P_X}^2$  for any  $J^* = 1, \dots, M$ . As in the proof of Theorem 3.1, this leads to

$$\begin{aligned}\|f_{\hat{\theta}} - f\|_{P_X}^2 &\leq \|f_{J^*} - f\|_{P_X}^2 + \max_{k=1, \dots, M} \delta_k, \\ \delta_k &:= (P - P_n)(I_{f_k}^{\text{RKD}} - I_{f_{J^*}}^{\text{RKD}}) - \frac{1}{2} \|f_{J^*} - f_k\|_{P_X}^2.\end{aligned}$$

As explained above, when the distribution of the design is known, the empirical process is linear in  $f_k - f_{J^*}$ :

$$\delta_k = -2(P - P_n)((f_k - f_{J^*})\dot{y}) - \frac{1}{2} \|f_k - f_{J^*}\|_{P_X}^2.$$

If for some  $b > 0$ ,  $|Y| \leq b$  and  $\max_{j=1, \dots, M} |f_j(X)| \leq b$  almost surely, then

$$\delta_k \leq -2(P - P_n)((f_k - f_{J^*})\dot{y}) - \frac{1}{2b^2} \mathbb{E}[Y^2(f_k(X) - f_{J^*}(X))].$$



Using (5.3) and the union bound on  $k = 1, \dots, M$ , we obtain  $\max_{k=1, \dots, M} \delta_k \leq \beta(x + \log M)/n$  with probability greater than  $1 - \exp(-x)$  and thus

$$\|f_{\hat{\theta}} - f\|_{P_X}^2 \leq \|f_{J^*} - f\|_{P_X}^2 + \beta \left( \frac{x + \log M}{n} \right),$$

where  $\beta = cb^2$  for some numerical constant  $c > 0$ .

In conclusion, the density estimation framework studied in the present paper is close to the regression problem when the distribution of the design is known, while it presents several differences with the regression problem when the design is not known.

## 4. Minimax optimality in deviation

The goal of this section is to state a minimax optimality result based on the lower bound of Theorem 3.2 and the sharp oracle inequality of Corollary 3.1. In this section, the underlying measure  $\mu$  is the Lebesgue measure on  $\mathbb{R}^d$  for some integer  $d \geq 1$ .

Minimax optimality in model selection type aggregation is usually defined in expectation [24], by studying the quantity

$$\sup_{j=1, \dots, M} \inf_{\hat{T}_n} \sup_{f \in \mathcal{F}_d} \left( \mathbb{E} R(\hat{T}_n) - \inf_{j=1, \dots, M} R(f_j) \right),$$

where the infimum is taken over all estimators  $\hat{T}_n$ ,  $\mathcal{F}$  is a class of possible functions for the dictionary and  $\mathcal{F}_d$  is the class of all densities satisfying some general constraints.

Let  $\mu$  be the Lebesgue measure on  $\mathbf{R}^d$  and for some  $L > 0$ , let  $\mathcal{F} = \{g \in L^2(\mu), \|g\|_\infty \leq L\}$  and  $\mathcal{F}_d$  be the set of all densities  $f$  with respect to  $\mu$  satisfying  $\|f\|_\infty \leq L$ . Then, by an integration argument, Corollary 3.1 and Theorem 3.2 provide the following bounds for some absolute constant  $c, C > 0$  and any  $M \geq 2, n \geq 1$ :

$$c \frac{L \log M}{n} \leq \sup_{j=1, \dots, M} \inf_{\hat{T}_n} \sup_{f \in \mathcal{F}_d} \left( \mathbb{E} R(\hat{T}_n) - \inf_{j=1, \dots, M} R(f_j) \right) \leq C \frac{L \log M}{n}.$$

This shows that  $\frac{L \log M}{n}$  is the optimal rate of convergence in expectation for model selection type aggregation under the boundedness assumption.

But our results are stronger than the above optimality in expectation since the deviation term in the sharp oracle inequality of Corollary 3.1 and in the lower bound of Theorem 3.2 are the same up to a numerical constant.

The central quantity when dealing with optimality in deviation is, for  $t > 0$ ,

$$\sup_{j=1, \dots, M} \inf_{\hat{T}_n} \sup_{f \in \mathcal{F}_d} \mathbb{P} \left( R(\hat{T}_n) - \inf_{j=1, \dots, M} R(f_j) > t \right).$$

The results of Section 3 provide upper and lower bounds for this quantity.

We propose the following definition of minimax optimality in deviation.

**Definition 4.1 (Minimax optimality in deviation).** Let  $\mathcal{F}$  be a subset of  $L^2(\mu)$  and  $\mathcal{F}_d$  be a set of densities with respect to the measure  $\mu$ . Let  $\mathcal{E}_n$  be a set of estimators. Denote by  $\mathbf{P}_{\mathcal{E}_n, \mathcal{F}, \mathcal{F}_d}^{n, M}(t)$  the quantity

$$\mathbf{P}_{\mathcal{E}_n, \mathcal{F}, \mathcal{F}_d}^{n, M}(t) = \sup_{j=1, \dots, M} \inf_{f_j \in \mathcal{F}} \sup_{\hat{T}_n \in \mathcal{E}_n} \mathbb{P} \left( R(\hat{T}_n) - \inf_{j=1, \dots, M} R(f_j) > t \right).$$

A function  $p_{n, M}(\cdot)$  is called optimal tail distribution over  $(\mathcal{E}_n, \mathcal{F}, \mathcal{F}_d)$  if for any  $n \geq 1, M \geq 2$  and any  $t > 0$ ,

$$cp_{n, M}(c't) \leq \mathbf{P}_{\mathcal{E}_n, \mathcal{F}, \mathcal{F}_d}^{n, M}(t) \leq p_{n, M}(t),$$

where  $c, c' > 0$  are constants independent of  $n, M$  and  $t$ .

The following proposition is a direct consequence of Corollary 3.1 and Theorem 3.2.

**Proposition 4.1.** Let  $M \geq 2, n \geq 1$  and  $L > 0$ . Let  $\mathcal{F} = \{g \in L^2(\mathbf{R}^d), \|g\|_\infty \leq L\}$  and  $\mathcal{F}_d$  be the set of all densities  $f$  with respect to the Lebesgue measure on  $\mathbf{R}^d$  with  $\|f\|_\infty \leq L$ . Let  $\mathcal{E}_n$  be the set of all estimators. Define

$$p_{n, M}(t) = M \exp \left( -\frac{3tn}{20L} \right) \mathbf{1}_{[0, L]}(t),$$

where  $\mathbf{1}_A$  denotes the indicator function of the set  $A$ . Then for all  $t > 0$ ,

$$\frac{1}{24} p_{n, M}(160t) \leq \mathbf{P}_{\mathcal{E}_n, \mathcal{F}, \mathcal{F}_d}^{n, M}(t) \leq p_{n, M}(t).$$

Thus,  $p_{n, M}(\cdot)$  is an optimal tail distribution over  $(\mathcal{E}_n, \mathcal{F}, \mathcal{F}_d)$  according to Definition 4.1.

**Proof.** The regime  $t > L$  corresponds to the trivial case where (3.8) holds and  $\hat{T}_n^* = 0$  is an optimal estimator. In this regime  $p_{n, M}(t) = 0$ .

For  $t \leq L$ , by setting  $t = \beta \frac{\log(M)+x}{n} = \frac{20L}{3} \frac{\log(M)+x}{n}$  in Corollary 3.1, we get

$$\mathbf{P}_{\mathcal{E}_n, \mathcal{F}, \mathcal{F}_d}^{n, M} \leq p_{n, M}(t)$$

while Theorem 3.2 implies that

$$\frac{1}{24} p_{n, M} \left( \frac{24 \cdot 20}{3} t \right) \leq \mathbf{P}_{\mathcal{E}_n, \mathcal{F}, \mathcal{F}_d}^{n, M}(t).$$

□

Similarly, the results of Section 2 imply the following proposition.

**Proposition 4.2.** Let  $M \geq 2, n \geq 1$  and  $L > 0$ . Let  $\mathcal{F} = \{g \in L^2(\mathbf{R}^d), \|g\|_\infty \leq L\}$  and  $\mathcal{F}_d$  be the set of all densities  $f$  with respect to the Lebesgue measure on  $\mathbf{R}^d$  with  $\|f\|_\infty \leq L$ . Let  $\mathcal{S}_n$  be the set of all selectors, that is, the measurable functions valued in the discrete set  $\{f_1, \dots, f_M\}$ . Define

$$q_{n,M}(t) = M \exp\left(-\frac{t^2 n}{L^2(4\sqrt{2} + 8/3)^2}\right) \mathbf{1}_{[0,L]}(t),$$

where  $\mathbf{1}_A$  denotes the indicator function of the set  $A$ . Then for all  $t > 0$ ,

$$\frac{1}{24} q_{n,M}(\sqrt{3}(4\sqrt{2} + 8/3)t) \leq \mathbf{P}_{\mathcal{S}_n, \mathcal{F}, \mathcal{F}_d}^{n,M}(t) \leq q_{n,M}(t).$$

Thus,  $q_{n,M}(\cdot)$  is an optimal tail distribution over  $(\mathcal{S}_n, \mathcal{F}, \mathcal{F}_d)$  according to Definition 4.1.

**Proof.** The regime  $t > L$  can be treated similarly as in the proof of Proposition 4.1.

For  $t \leq L$ , let  $t = L(4\sqrt{2} + 8/3)\sqrt{\frac{x+\log M}{n}}$  in Theorem 2.2. For this definition of  $t$  and  $x$ ,  $1 \geq \sqrt{\frac{x+\log M}{n}} \geq \frac{x+\log M}{n}$ . Then

$$\mathbf{P}_{\mathcal{S}_n, \mathcal{F}, \mathcal{F}_d}^{n,M}(t) \leq q_{n,M}(t)$$

and Theorem 2.1 implies

$$\frac{1}{24} q_{n,M}(\sqrt{3}(4\sqrt{2} + 8/3)t) \leq \mathbf{P}_{\mathcal{S}_n, \mathcal{F}, \mathcal{F}_d}^{n,M}(t). \quad \square$$

## 5. Proofs

### 5.1. Bias-variance decomposition

As discussed in Section 3, the penalty can be viewed as the variance of a random element of the discrete set  $\{f_1, \dots, f_M\}$  and it satisfies the following bias-variance decomposition.

**Proposition 5.1.** For any  $g \in L^2(\mu)$  and  $\theta \in \Lambda^M$ ,

$$\sum_{j=1}^M \theta_j \|f_j - g\|^2 = \|f_\theta - g\|^2 + \text{pen}(\theta), \quad (5.1)$$

where  $\text{pen}(\cdot)$  is the penalty defined in (3.3).

**Proof.** Let  $\eta$  be a random variable with  $\mathbb{P}(\eta = j) = \theta_j$  for all  $j = 1, \dots, M$ . Denote by  $\mathbb{E}_\theta$  the expectation with respect to the random variable  $\eta$ . Then  $\mathbb{E}_\theta[f_\eta] = f_\theta$  and the bias-variance decomposition yields

$$\mathbb{E}_\theta \|f_\theta - g\|^2 = \|g - \mathbb{E}_\theta[f_\eta]\|^2 + \mathbb{E}_\theta \|f_\eta - \mathbb{E}_\theta[f_\eta]\|^2,$$

which is exactly the desired result.  $\square$

## 5.2. Concentration inequalities

**Proposition 5.2.** *Let  $Y_1, \dots, Y_n$  be independent random variables, such that almost surely, for all  $i$ ,  $|Y_i - \mathbb{E}Y_i| \leq b$ . Then for all  $x > 0$ ,*

$$\mathbb{P}\left(\sum_{i=1}^n Y_i - \mathbb{E}Y_i > \sqrt{2xv} + \frac{bx}{3}\right) \leq \exp(-x), \quad (5.2)$$

where  $v = \sum_{i=1}^n \mathbb{V}(Y_i)$ .

Proposition 5.2 is close to Bennett and Bernstein inequalities. A proof can be found in [18], Section 2.2.3, (2.20) with  $c = b/3$ .

The following one-sided concentration inequality is a direct consequence of Proposition 5.2 and the inequality  $2\sqrt{uv} \leq \frac{u}{a} + av$  for all  $a, u, v > 0$ . Under the same assumptions as Proposition 5.2 above, for all  $x > 0$  and any  $a > 0$ ,

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Y_i - \mathbb{E}Y_i - a\mathbb{V}(Y_i) > \left(\frac{1}{2a} + \frac{b}{3}\right)\frac{x}{n}\right) \leq \exp(-x). \quad (5.3)$$

**Proposition 5.3.** *Let  $X_1, \dots, X_n$  be i.i.d. observations drawn from the density  $f$  with  $\|f\|_\infty \leq L$ . Let  $g \in L^2(\mu)$  with  $\|g\|_\infty \leq 4L_0$ . Let  $\beta = 4L + \frac{8L_0}{3}$ . Define*

$$\zeta_n = (P - P_n)g - \frac{1}{8}\|g\|^2 - \frac{\beta}{n} \log \frac{1}{\pi},$$

where the notation  $P$  and  $P_n$  is defined in (2.2). Then for all  $x > 0$ ,

$$\mathbb{P}\left(\zeta_n > \frac{\beta x}{n}\right) \leq \pi \exp(-x).$$

**Proof.** As the unknown density  $f$  is bounded by  $L$ ,

$$\begin{aligned} \mathbb{V}(g(X_1)) &\leq P(g^2) = \int g^2 f d\mu \leq L\|g\|^2, \\ -\frac{1}{8}\|g\|^2 &\leq -\frac{1}{8L}\mathbb{V}(g(X_1)). \end{aligned}$$

Thus, almost surely

$$\zeta_n \leq (P - P_n)g - \frac{1}{8L}\mathbb{V}(g(X_1)) - \frac{\beta}{n} \log \frac{1}{\pi}.$$

Define  $n$  i.i.d. random variables  $Y_1, \dots, Y_n$  by

$$Y_i = g(X_i).$$

Almost surely,  $|Y_i| \leq 4L_0$  and  $|Y_i - \mathbb{E}Y_i| \leq 8L_0$ . By applying (5.3) to  $Y_1, \dots, Y_n$  with  $b = 8L_0$  and  $a = \frac{1}{8L}$ , we get that for any  $t > 0$  with  $x = t + \log \frac{1}{\pi}$ ,

$$\begin{aligned} \mathbb{P}\left((P - P_n)g - \frac{1}{8L}\mathbb{V}(g(X_1)) > \frac{\beta x}{n}\right) &\leq \exp(-x) \\ \mathbb{P}\left(\zeta_n > \frac{\beta x}{n}\right) &\leq \mathbb{P}\left((P - P_n)g - \frac{1}{8L}\mathbb{V}(g(X_1)) - \frac{\beta}{n} \log \frac{1}{\pi} > \frac{\beta t}{n}\right) \leq \pi \exp(-t). \quad \square \end{aligned}$$

### 5.3. Strong convexity

**Proof of Proposition 3.1.** We will first prove that for any  $\theta, \theta'$ ,

$$H_n(\theta) - H_n(\theta') = \langle \nabla H_n(\theta'), \theta - \theta' \rangle + \frac{1}{2} \|f_\theta - f_{\theta'}\|^2. \quad (5.4)$$

Using the bias-variance decomposition of (5.1) with  $g = 0$ , we get

$$\text{pen}(\theta) = \sum_{j=1}^M \theta_j \|f_\theta - f_j\|^2 = -\|f_\theta\|^2 + \sum_{j=1}^M \theta_j \|f_j\|^2.$$

Thus,  $H_n$  can be rewritten as  $H_n(\theta) = \frac{1}{2} \|f_\theta\|^2 + L(\theta)$  where  $L$  is affine in  $\theta$ . If we can prove  $N(\theta) - N(\theta') = \langle \nabla N(\theta'), \theta - \theta' \rangle + \|f_\theta - f_{\theta'}\|^2$  where  $N(\theta) = \|f_\theta\|^2$ , then (5.4) holds. By simple properties of the norm,

$$\begin{aligned} \|f_\theta\|^2 - \|f_{\theta'}\|^2 &= 2 \int f_{\theta'}(f_\theta - f_{\theta'}) d\mu + \|f_\theta - f_{\theta'}\|^2 \\ &= 2\theta'^T G(\theta - \theta') + \|f_\theta - f_{\theta'}\|^2, \end{aligned}$$

where  $G$  is the Gram matrix with elements  $G_{j,k} = \int f_j f_k d\mu$  for all  $j, k = 1, \dots, M$ . The gradient at  $\theta'$  of the function  $\theta \rightarrow \|f_\theta\|^2$  is exactly  $2G\theta'$  so (5.4) holds.

The function  $H_n$  is convex and differentiable. If  $\hat{\theta}$  minimizes  $H_n$  over the simplex, then for any  $\theta \in \Lambda^M$ ,  $\langle \nabla H_n(\hat{\theta}), \theta - \hat{\theta} \rangle \geq 0$  which proves (3.5).  $\square$

### 5.4. Tools for lower bounds

**Proposition 5.4.** *There exists a countable set of functions  $\varepsilon_1, \varepsilon_2, \dots$  defined on  $[0, 1]$  such that for all  $j, k > 0$  with  $k \neq j$ ,*

$$\forall u \in [0, 1), \quad \varepsilon_j(u) \in \{-1, 1\},$$

$$\int_{[0,1]} \varepsilon_j(x) \varepsilon_k(x) dx = 0,$$

$$\int_{[0,1]} \varepsilon_j^2(x) dx = 1.$$

Furthermore, if  $U$  is uniformly distributed on  $[0, 1]$ , then  $\varepsilon_1(U), \varepsilon_2(U), \dots$  are i.i.d. Rademacher random variables.

See [7], Definition 3.22, for an explicit construction of these functions and a proof of their properties.

If  $P \ll Q$  are two probability measures defined on some measurable space, define their Kullback–Leibler divergence and their  $\chi_2$  divergence by

$$K(P, Q) = \int \log\left(\frac{dP}{dQ}\right) dP, \quad \chi_2(P, Q) = \int \left(\frac{dP}{dQ} - 1\right)^2 dQ.$$

The following comparison holds

$$K(P, Q) \leq \chi_2(P, Q). \quad (5.5)$$

Furthermore, if  $n \geq 1$  and  $P^{\otimes n}$  denotes the  $n$ -product of measures  $P$ ,

$$K(P^{\otimes n}, Q^{\otimes n}) = nK(P, Q). \quad (5.6)$$

The proofs of (5.5) and (5.6) are given in [25], Lemma 2.7 and page 85.

**Lemma 5.1.** *Let  $(\Omega, \mathcal{A})$  be a measurable space and  $m \geq 1$ . Let  $m \geq 1$  and  $A_0, \dots, A_m \in \mathcal{A}$  be disjoint events:  $A_j \cap A_k = \emptyset$  for any  $j \neq k$ . Assume that  $Q_0, \dots, Q_m$  are probability measures on  $(\Omega, \mathcal{A})$  such that*

$$\frac{1}{m} \sum_{j=1}^m K(Q_j, Q_0) \leq \chi < \infty.$$

Then,

$$\max_{j=0, \dots, m} Q_j(\Omega \setminus A_j) \geq \frac{1}{12} \min(1, m \exp(-3\chi)).$$

Lemma 5.1 can be found in [9], Lemma 3. It is a direct consequence of [25], Proposition 2.3, with  $\tau^* = \min(m^{-1}, \exp(-3\chi))$ .

**Corollary 5.1 (Minimax lower bounds).** *Let  $n \geq 1$  be an integer and  $s > 0$  be a positive number. Let  $m \geq 1$  and  $q_0, \dots, q_m$  be a family of densities with respect to the same measure  $\mu$ . Assume that for any  $j \neq k$ ,*

$$\|q_j - q_k\|^2 \geq 4s > 0. \quad (5.7)$$

If  $P_k^{\otimes n}$  denotes the product measure associated with  $n$  i.i.d. observations drawn from the density  $q_k$ , assume that

$$\frac{1}{m} \sum_{j=1}^m K(P_j^{\otimes n}, P_0^{\otimes n}) \leq \chi$$

for some finite  $\chi > 0$ . Then, for any estimator  $\hat{T}_n$ ,

$$\max_{k=0,\dots,m} \mathbb{P}_k^{\otimes n}(\|\hat{T}_n - q_k\|^2 \geq s) \geq \frac{1}{12} \min(1, m \exp(-3\chi)).$$

**Proof.** For any estimator  $\hat{T}_n$ , for any  $j = 0, \dots, m$  define the events

$$A_j = \{\|\hat{T}_n - q_j\|^2 < s\}.$$

These events are disjoint because of the triangle inequality and (5.7). Applying Lemma 5.1 completes the proof.  $\square$

## 5.5. Lower bound theorems

### 5.5.1. Lower bounds with exponential tails

**Proof of Theorem 3.2.** Let  $\varepsilon_2, \dots, \varepsilon_M$  be  $M - 1$  functions from Proposition 5.4. Consider the dictionary  $\{f_1, \dots, f_M\}$  such that for all  $(u_1, \dots, u_d) \in \mathbf{R}^d$

$$f_1(u_1, \dots, u_d) = \frac{L}{2} \mathbf{1}_{[0,1]} \left( \frac{L}{2} u_1 \right) \prod_{q=2}^d \mathbf{1}_{[0,1]}(u_q),$$

and for  $j \geq 2$

$$f_j(u_1, \dots, u_d) = \frac{L}{2} \left( 1 + \sqrt{\frac{\log(M) + x}{3n}} \varepsilon_j \left( \frac{L}{2} u_1 \right) \right) \mathbf{1}_{[0,1]} \left( \frac{L}{2} u_1 \right) \prod_{q=2}^d \mathbf{1}_{[0,1]}(u_q).$$

Since  $\frac{\log M + x}{n} < 3$ , these functions are densities and satisfy  $\|f_j\|_\infty < L$ .

For any  $j \neq k$ ,

$$\|f_j - f_k\|^2 \geq L \frac{\log(M) + x}{6n} \quad (5.8)$$

and (5.8) is true with equality when  $j = 1$ . If  $P_k^{\otimes n}$  denotes the probability with respect to  $n$  i.i.d. random variables with density  $f_j$ , the properties (5.5) and (5.6) give that for any  $k \geq 2$ ,

$$\begin{aligned} K(P_k^{\otimes n}, P_1^{\otimes n}) &= n K(P_k^{\otimes 1}, P_1^{\otimes 1}) \\ &\leq n \chi_2(P_k^{\otimes 1}, P_1^{\otimes 1}) \\ &= n \frac{2}{L} \|f_k - f_1\|^2 \\ &= \frac{\log(M) + x}{3}. \end{aligned}$$

Applying Corollary 5.1 with  $m = M - 1$  yields that for any estimator  $\hat{T}_n$ ,

$$\begin{aligned} \sup_{j=1,\dots,M} P_j^{\otimes n} \left( \|\hat{T}_n - f_j\|^2 \geq L \frac{\log(M) + x}{24n} \right) &\geq \frac{1}{12} \min \left( 1, \frac{M-1}{M} \exp(-x) \right) \\ &\geq \frac{1}{24} \exp(-x). \end{aligned} \quad \square$$

**Proof of Theorem 2.1.** Let  $\varepsilon_1, \dots, \varepsilon_M$  be  $M$  functions from Proposition 5.4.

For  $(u_1, \dots, u_d) \in \mathbf{R}^d$ , we define a dictionary  $\{f_1, \dots, f_M\}$  by

$$f_j(u_1, \dots, u_d) = \frac{L}{2} \left( 1 + \varepsilon_j \left( \frac{L}{2} u_1 \right) \right) \mathbf{1}_{[0,1]} \left( \frac{L}{2} u_1 \right) \prod_{q=2}^d \mathbf{1}_{[0,1]}(u_q),$$

and we define  $M$  densities  $\{d_1, \dots, d_M\}$ :

$$d_j(u_1, \dots, u_d) = \frac{L}{2} \left( 1 + \gamma \varepsilon_j \left( \frac{L}{2} u_1 \right) \right) \mathbf{1}_{[0,1]} \left( \frac{L}{2} u_1 \right) \prod_{q=2}^d \mathbf{1}_{[0,1]}(u_q),$$

for some  $\gamma \in (0, \frac{1}{2})$  that will be specified later. Due to the properties of the  $(\varepsilon_j)$ , the following holds for any  $j \neq k$

$$\begin{aligned} \|f_k - d_j\|^2 &= \frac{L}{2} (1 + \gamma^2), \\ \|f_j - d_j\|^2 &= \frac{L}{2} (1 - \gamma)^2, \\ \|d_j - d_k\|^2 &= L\gamma^2. \end{aligned}$$

Thus if  $\hat{S}_n$  is any selector taking values in the discrete set  $\{f_1, \dots, f_M\}$ :

$$\|\hat{S}_n - d_j\|^2 - \inf_{l=1,\dots,M} \|f_l - d_j\|^2 = \|\hat{S}_n - d_j\|^2 - \|f_j - d_j\|^2 = 2L\gamma \mathbf{1}_{\hat{S}_n \neq f_j}. \quad (5.9)$$

Let  $P_k^{\otimes n}$  be the product measure associated with  $n$  i.i.d. random variables drawn from the density  $d_k$ . Equation (5.9) ensures that with probability  $\mathbb{P}_j^{\otimes n}(\hat{S}_n \neq f_j)$ , the excess risk is  $2L\gamma$ .

For any  $k \neq 1$ , using (5.5) and (5.6), we obtain

$$\begin{aligned} K(P_k^{\otimes n}, P_1^{\otimes n}) &= nK(P_k^{\otimes 1}, P_1^{\otimes 1}) \\ &\leq n\chi_2(P_k^{\otimes 1}, P_1^{\otimes 1}) \\ &\leq \frac{4}{L} n \|d_k - d_1\|^2 \\ &= 4n\gamma^2, \end{aligned}$$

where we used that  $d_1(u_1, \dots, u_d) \geq L/4$  almost surely on the common support of  $d_k$  and  $d_1$ .



Now we choose  $\gamma = \frac{1}{2\sqrt{3}} \sqrt{\frac{x+\log M}{n}}$  such that  $\forall k \neq 1, K(P_k^{\otimes n}, P_1^{\otimes n}) \leq \frac{x+\log M}{3}$ . Let  $\hat{S}_n$  be any estimator with values in the discrete set  $\{f_1, \dots, f_M\}$ . For any  $j = 1, \dots, M$ , define the event  $A_j = \{\hat{S}_n = f_j\}$ . The events are disjoint if  $f_j \neq f_k$  for all  $j \neq k$  (if this is not satisfied, we can always remove the duplicates). By applying Lemma 5.1 with  $m = M - 1$  and  $\chi = \frac{1}{3}(x + \log M)$ , we get

$$\max_{j=1, \dots, M} \mathbb{P}_j^{\otimes n}(\hat{S}_n \neq f_j) \geq \frac{M-1}{12M} \exp(-x).$$

Since  $(M-1)/M \geq 1/2$ ,

$$\begin{aligned} \max_{j=1, \dots, M} \mathbb{P}_j^{\otimes n} \left( \|\hat{S}_n - d_j\|^2 - \inf_{l=1, \dots, M} \|f_l - d_j\|^2 > 2L\gamma \right) &\geq \frac{M-1}{12M} \exp(-x) \\ &\geq \frac{1}{24} \exp(-x). \quad \square \end{aligned}$$

### 5.5.2. ERM over the convex hull

**Proof of Proposition 2.1.** By homogeneity, it is enough to prove the case  $L = 2$ . Let  $\phi_1, \dots, \phi_M, \phi_{M+1}$  be  $M+1$  functions given by Proposition 5.4. Consider the probability density  $f = \mathbf{1}_{[0,1]}$  and the dictionary of  $2M+1$  functions

$$\mathcal{D} = \{\mathbf{1}_{[0,1]}\} \cup \{(1 \pm \phi_j \phi_{M+1})\mathbf{1}_{[0,1]}, j = 1, \dots, M\}.$$

The true density is in the dictionary thus  $\min_{g \in \mathcal{D}} \|f - g\|^2 = 0$ . Also, all the elements of the dictionary are uniformly bounded by  $L = 2$ .

The convex hull of the dictionary is the set

$$\{g_\lambda = (1 + f_\lambda \phi_{M+1})\mathbf{1}_{[0,1]}, \lambda \in \mathbf{R}^M, |\lambda|_1 \leq 1\},$$

where  $f_\lambda = \sum_{j=1}^M \lambda_j \phi_j$  and  $|\cdot|_1$  is the  $\ell_1$  norm in  $\mathbf{R}^M$ .

For all  $\lambda \in \mathbf{R}^M$  with  $|\lambda|_1 \leq 1$ ,  $\|f - g_\lambda\|^2 = |\lambda|_2^2$  where  $|\cdot|_2$  is the  $\ell_2$  norm in  $\mathbf{R}^M$ .

Let  $\mathcal{L}_\lambda := \|g_\lambda\|^2 - 2g_\lambda + 2f - \|f\|^2 = |\lambda|_2^2 - 2f_\lambda \phi_{M+1}$ . Since the empirical process is linear in  $\lambda$ , the proof from [12] can be adapted as follows. Given  $n$  i.i.d. observations  $X_1, \dots, X_n$  generated by the density  $f$ , [12], Lemma 5.4, states that for every  $r > 0$ , with probability greater than  $1 - 6 \exp(-C_2 M)$ ,

$$c_0 \sqrt{\frac{r}{M}} \leq c_1 \sqrt{\frac{rM}{n}} \leq \sup_{\lambda \in \mathbf{R}^M, |\lambda|_2 \leq \sqrt{r}} P_n(f_\lambda \phi_{M+1}) \leq c_2 \sqrt{\frac{rM}{n}} \leq c_3 \sqrt{\frac{r}{M}},$$

where  $c_0, c_1, c_2, c_3 > 0$  are absolute constants.

Let  $r \leq 1/M$  that will be specified later (such that if  $|\lambda|_2 \leq \sqrt{r}$  then  $|\lambda|_1 \leq 1$ ). On the one hand,

$$\inf_{\lambda \in \mathbf{R}^M, |\lambda|_2 \leq \sqrt{r}} P_n \mathcal{L}_\lambda \leq r - 2 \sup_{\lambda \in \mathbf{R}^M, |\lambda|_2 \leq \sqrt{r}} P_n(f_\lambda \phi_{M+1}).$$

Given that  $n \sim M^2$ , using the above high probability estimate, there exists a positive absolute constant  $c_4$  such that for all  $r \leq c_3^2/(4M)$ , with probability greater than  $1 - 6 \exp(-C_2 M)$ ,  $\inf_{\lambda \in \mathbf{R}^M, |\lambda|_2 \leq \sqrt{r}} P_n \mathcal{L}_\lambda \leq \sqrt{r}(\sqrt{r} - c_3/\sqrt{M}) \leq -c_4 \sqrt{r/M}$ , where  $c_4 = c_3/2$ .

On the other hand, if  $\rho \leq 1/M$ , with probability greater than  $1 - 6 \exp(-C_2 M)$ ,

$$\sup_{\lambda \in \mathbf{R}^M, |\lambda|_2 \leq \sqrt{\rho}} |(P_n - P) \mathcal{L}_\lambda| = 2 \sup_{\lambda \in \mathbf{R}^M, |\lambda|_2 \leq \sqrt{\rho}} |(P_n - P) f_\lambda \phi_{M+1}| \leq 2c_3 \sqrt{\frac{\rho}{M}}.$$

Finally, choose  $r, \rho$  such that  $2c_3 \sqrt{\rho/M} < c_4 \sqrt{r/M}$  and  $\rho > c_5/\sqrt{n}$  for some absolute constant  $c_5 > 0$ , then with probability greater than  $1 - 12 \exp(-C_2 M)$ ,

$$\inf_{\lambda, |\lambda|_2 \leq \sqrt{\rho}} P_n \mathcal{L}_\lambda \geq - \sup_{\lambda, |\lambda|_2 \leq \sqrt{\rho}} |(P_n - P) \mathcal{L}_\lambda| \geq -2c_3 \sqrt{\frac{\rho}{M}} > -c_4 \sqrt{\frac{r}{M}} \geq \inf_{\lambda, |\lambda|_2 \leq \sqrt{r}} P_n \mathcal{L}_\lambda.$$

Thus with high probability,  $\inf_{\lambda, |\lambda|_2 \leq \sqrt{\rho}} P_n \mathcal{L}_\lambda > \inf_{\lambda, |\lambda|_1 \leq 1} P_n \mathcal{L}_\lambda$ . The inequality is strict so the empirical risk minimizer has a risk greater than  $\rho$ . As  $\rho$  satisfies  $\rho > C_3/\sqrt{n}$ , the proof is complete.  $\square$

### 5.5.3. Exponential weights

If  $Y_1, \dots, Y_m$  are i.i.d. with  $\mathbb{P}(Y_1 = \pm 1) = 1/2$ , then for all  $u \in [0, \sqrt{m}/4]$ ,

$$\frac{1}{15} \exp(-4u^2) \leq \mathbb{P}(Y_1 + \dots + Y_m \geq u\sqrt{m}) \leq \exp(-u^2/2). \quad (5.10)$$

A proof of the lower bound can be found in [19], Proposition 7.3.2, and a standard Chernoff bound provides the upper bound. The following proof uses arguments similar to [3].

**Proof of Proposition 2.2.** By homogeneity, it is enough to prove the case  $L = 1$ . Let  $\varepsilon_1, \varepsilon_2, \varepsilon_3$  be 3 functions from Proposition 5.4. Let  $f = \mathbf{1}_{[0,1]}$  be the unknown density and let

$$\begin{aligned} f_1 &= f + \varepsilon_1, & f_2 &= f + \left(1 + \frac{1}{\sqrt{n}}\right) \varepsilon_2, & f_3 &= f_2 + \frac{\alpha}{\sqrt{n}} \varepsilon_3, \\ \pi_1 &= 1/(8\sqrt{n}), & \pi_2 &= 1/(8\sqrt{n}), & \pi_3 &= 1 - 1/(4\sqrt{n}), \end{aligned}$$

where  $0 \leq \alpha \leq n^{1/4}$  will be specified later. The best function in the dictionary is  $f_1$ :  $\|f_1 - f\|^2 = \min_{j=1, \dots, M} \|f_j - f\|^2$ .

Let  $E$  be the event  $\{R_n(f_2) + 2/\sqrt{n} \leq R_n(f_1)\}$ . By simple algebra,

$$E = \{1 + 4\sqrt{n} - 2\sqrt{n} P_n(\varepsilon_2) \leq 2n P_n(\varepsilon_2 - \varepsilon_1)\} \supseteq \{7\sqrt{n} \leq 2n P_n(\varepsilon_2 - \varepsilon_1)\},$$

where for the inclusion we used  $1 \leq \sqrt{n}$  and  $|P_n(\varepsilon_2)| \leq 1$ . The  $2n$  random variables  $(\varepsilon_j(X_i))_{j=1,2; i=1, \dots, n}$  are i.i.d. Rademacher random variables, so applying the lower bound of (5.10) with  $m = 2n$  and  $u = 7\sqrt{2}/4$  yields  $\mathbb{P}(E) \geq C_2 > 0$  for some absolute constant  $C_2$ . Now set  $\alpha^2 = 8 \log(2n/C_2)$ , and choose  $N_0$  such that for all  $n \geq N_0$ ,  $8 \log(2n/C_2) > 0$  and  $\alpha^2 \leq \sqrt{n}$ .

Let  $F := \{R_n(f_3) \leq R_n(f_1)\}$  and define

$$G = \{2(\alpha/\sqrt{n})P_n(\varepsilon_3) \leq \alpha^2/n - 2/\sqrt{n}\}.$$

Since  $R_n(f_3) = R_n(f_2) + \alpha^2/n - 2(\alpha/\sqrt{n})P_n(\varepsilon_3)$ , we have  $E \cap G^c \subseteq F$ . As  $\alpha^2 \leq \sqrt{n}$  holds, we have  $\alpha^2 - 2\sqrt{n} \leq -\alpha^2$  and

$$G \subseteq \{(2\alpha/n)P_n(\varepsilon_3) \leq -\alpha^2/n\} = \{-nP_n(\varepsilon_3) \geq \sqrt{n}\alpha/2\}.$$

The random variable  $-nP_n(\varepsilon_j)$  is the sum of  $n$  independent Rademacher random variables. Applying the upper bound of (5.10) to  $u = \alpha/2$ , we have  $\mathbb{P}(G) \leq \exp(-\alpha^2/8) = C_2/(2n)$  since  $\alpha = 8 \log(2n/C_2)$ . Now as  $F^c \subset E^c \cup G$ ,

$$\mathbb{P}(E^c \cup F^c) \leq \mathbb{P}(E^c \cup G) \leq (1 - C_2) + \frac{C_2}{2n} \leq 1 - C_2/2 < 1.$$

The probability of the event  $E \cap F$  is greater than  $C_0 := C_2/2$ . On this event,  $R_n(f_2) \leq R_n(f_1)$  and  $R_n(f_3) \leq R_n(f_1)$  thus

$$\begin{aligned} \hat{\theta}_1^{\text{EW},\beta} &= \frac{\pi_1 \exp(-R_n(f_1)/\beta)}{\pi_1 \exp(-R_n(f_1)/\beta) + \pi_2 \exp(-R_n(f_2)/\beta) + \pi_3 \exp(-R_n(f_3)/\beta)} \\ &\leq \frac{\pi_1 \exp(-R_n(f_1)/\beta)}{(\pi_1 + \pi_2 + \pi_3) \exp(-R_n(f_1)/\beta)} = \pi_1 = \frac{1}{8\sqrt{n}}. \end{aligned}$$

Let  $\theta_1 = \hat{\theta}_1^{\text{EW},\beta}$  for simplicity. As  $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$  is an orthonormal system,

$$\begin{aligned} \|f_{\hat{\theta}^{\text{EW},\beta}} - f\|^2 - \|f_1 - f\|^2 &\geq \|\theta_1 f_1 + (1 - \theta_1) f_2 - f\|^2 - \|f_1 - f\|^2 \\ &= (1 - \theta_1)^2 \|f_2 - f\|^2 - (1 - \theta_1^2) \|f_1 - f\|^2 \\ &\geq 2(1 - \theta_1)^2 / \sqrt{n} + [(1 - \theta_1)^2 - (1 - \theta_1^2)] \\ &\geq 1/(2\sqrt{n}) - 2\theta_1 \\ &\geq 1/(2\sqrt{n}) - 2/(8\sqrt{n}) \geq 1/(4\sqrt{n}). \end{aligned} \quad \square$$

The proof of Proposition 2.3 is based on estimates from [14] and highlights the similarities between regression with random design and density estimation with the  $L^2$  risk.

**Proof of Proposition 2.3.** By homogeneity, it is enough to prove the case  $L = 1$ . The strategy is to construct an example for density estimation such that the calculations from [14], Proof of Theorem A, can be leveraged. Let  $f_Y$  be the probability density

$$f_Y(x) = \begin{cases} 1/4 + 1/(2\sqrt{n}), & \text{if } x \in [-2, 0), \\ 1/4 - 1/(2\sqrt{n}), & \text{if } x \in (0, 2], \end{cases}$$

and 0 elsewhere. Let  $\{f_1 = \frac{1}{2}\mathbf{1}_{[-2,0)}, f_2 = \frac{1}{2}\mathbf{1}_{[0,2)}\}$  be the dictionary. Let

$$\mathcal{L}_2(y) := \|f_2\|^2 - 2f_2(y) + 2f_1(y) - \|f_1\|^2, \quad \forall y \in \mathbf{R},$$

and observe that  $\mathcal{L}_2(Y) = -X$  where  $X = \mathbf{1}_{[0,2)}(Y) - \mathbf{1}_{[-2,0)}(Y)$  so that  $X$  satisfies

$$X = \begin{cases} 1, & \text{with probability } 1/2 - 1/\sqrt{n}, \\ -1, & \text{with probability } 1/2 + 1/\sqrt{n}. \end{cases}$$

By definition of  $\mathcal{L}_2$ ,

$$P\mathcal{L}_2 = \mathbb{E}\mathcal{L}_2(Y) = \|f_2 - f_Y\|^2 - \|f_1 - f_Y\|^2.$$

As  $P\mathcal{L}_2 = \mathbb{E}[-X] = 2/\sqrt{n} > 0$ ,  $f_1$  is the best function in the dictionary and  $P\mathcal{L}_2$  is the excess risk of  $f_2$ . Finally, let

$$\alpha = \frac{\|f_1 - f_2\|^2}{P\mathcal{L}_2} = \frac{\sqrt{n}}{2}.$$

For any  $\theta \in [0, 1]$ , let  $f_\theta = \theta f_1 + (1 - \theta)f_2$ . An explicit calculation of the excess risk of  $f_\theta$  yields

$$\begin{aligned} \|f_\theta - f_Y\|^2 - \|f_1 - f_Y\|^2 &= \theta^2\|f_1\|^2 + (1 - \theta)^2\|f_2\|^2 - 2\mathbb{E}[f_\theta(Y)] + 2\mathbb{E}[f_1(Y)] - \|f_1\|^2 \\ &= -\theta(1 - \theta)\|f_1 - f_2\|^2 + (1 - \theta)\mathbb{E}[-X] \\ &= (1 - \theta - \theta(1 - \theta)\alpha)P\mathcal{L}_2. \end{aligned}$$

Given  $n$  independent observations  $Y_1, \dots, Y_n$  with common density  $f$ , define  $X_i = \mathbf{1}_{[0,2)}(Y_i) - \mathbf{1}_{[-2,0)}(Y_i)$  as above. The exponential weights estimator with temperature  $\beta$  can be written as

$$\hat{f}_\beta^{\text{EW}} = \hat{\theta}_1 f_1 + (1 - \hat{\theta}_1) f_2, \quad \hat{\theta}_1 := \frac{1}{1 + \exp(-(n/\beta)(1/n) \sum_{i=1}^n [-X_i])},$$

and its excess risk is  $\|\hat{f}_\beta^{\text{EW}} - f_Y\|^2 - \|f_1 - f_Y\|^2 = (1 - \hat{\theta}_1 - \hat{\theta}_1(1 - \hat{\theta}_1)\alpha)P\mathcal{L}_2$ .

The constants  $\alpha$  and  $P\mathcal{L}_2$ , the law of  $X_1, \dots, X_n, \hat{\theta}_1$  are the same as in [14], Proof of Theorem A, thus the lower bounds in expectation and probability of the quantity  $(1 - \hat{\theta}_1 - \hat{\theta}_1(1 - \hat{\theta}_1)\alpha)$  in Lecué and Mendelson [14] also hold here and yield the lower bound of Proposition 2.3.  $\square$

## Acknowledgement

This work was supported by GENES and by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

## References

- [1] Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization. Lecture Notes in Math.* **1851**. Berlin: Springer. [MR2163920](#)
- [2] Dai, D., Rigollet, P., Xia, L. and Zhang, T. (2014). Aggregation of affine estimators. *Electron. J. Stat.* **8** 302–327. [MR3192554](#)
- [3] Dai, D., Rigollet, P. and Zhang, T. (2012). Deviation optimal learning using greedy  $Q$ -aggregation. *Ann. Statist.* **40** 1878–1905. [MR3015047](#)
- [4] Dalalyan, A.S. and Salmon, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40** 2327–2355. [MR3059085](#)
- [5] Dalalyan, A.S. and Tsybakov, A.B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning Theory. Lecture Notes in Computer Science* **4539** 97–111. Berlin: Springer. [MR2397581](#)
- [6] Dalalyan, A.S. and Tsybakov, A.B. (2012). Mirror averaging with sparsity priors. *Bernoulli* **18** 914–944. [MR2948907](#)
- [7] Heil, C. (2011). *A Basis Theory Primer*, Expanded ed. *Applied and Numerical Harmonic Analysis*. New York: Birkhäuser/Springer. [MR2744776](#)
- [8] Juditsky, A., Rigollet, P. and Tsybakov, A.B. (2008). Learning by mirror averaging. *Ann. Statist.* **36** 2183–2206. [MR2458184](#)
- [9] Kerkacharian, G., Tsybakov, A.B., Temlyakov, V., Picard, D. and Koltchinskii, V. (2014). Optimal exponential bounds on the accuracy of classification. *Constr. Approx.* **39** 421–444. [MR3207667](#)
- [10] Lecué, G. (2006). Lower bounds and aggregation in density estimation. *J. Mach. Learn. Res.* **7** 971–981. [MR2274393](#)
- [11] Lecué, G. (2013). Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli* **19** 2153–2166. [MR3160549](#)
- [12] Lecué, G. and Mendelson, S. (2009). Aggregation via empirical risk minimization. *Probab. Theory Related Fields* **145** 591–613. [MR2529440](#)
- [13] Lecué, G. and Mendelson, S. (2010). Sharper lower bounds on the performance of the empirical risk minimization algorithm. *Bernoulli* **16** 605–613. [MR2730641](#)
- [14] Lecué, G. and Mendelson, S. (2013). On the optimality of the aggregate with exponential weights for low temperatures. *Bernoulli* **19** 646–675. [MR3037168](#)
- [15] Lecué, G. and Rigollet, P. (2014). Optimal learning with  $Q$ -aggregation. *Ann. Statist.* **42** 211–224. [MR3178462](#)
- [16] Leung, G. and Barron, A.R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52** 3396–3410. [MR2242356](#)
- [17] Lounici, K. (2007). Generalized mirror averaging and  $D$ -convex aggregation. *Math. Methods Statist.* **16** 246–259. [MR2356820](#)
- [18] Massart, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Berlin: Springer. [MR2319879](#)
- [19] Matousek, J. and Vondrak, J. (2008). The probabilistic method, lecture notes. Available at <http://kam.mff.cuni.cz/~matousek/>.
- [20] Rigollet, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *Ann. Statist.* **40** 639–665. [MR2933661](#)
- [21] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. [MR2816337](#)
- [22] Rigollet, Ph. and Tsybakov, A.B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** 260–280. [MR2356821](#)

- [23] Rigollet, P. and Tsybakov, A.B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* **27** 558–575. [MR3025134](#)
- [24] Tsybakov, A.B. (2003). Optimal rates of aggregation. In *Learning Theory and Kernel Machines* 303–313. Berlin: Springer.
- [25] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Berlin: Springer. [MR2724359](#)
- [26] Wegkamp, M.H. (1999). Quasi-universal bandwidth selection for kernel density estimators. *Canad. J. Statist.* **27** 409–420. [MR1704442](#)
- [27] Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.* **28** 75–87. [MR1762904](#)

*Received January 2015 and revised April 2015*