Submitted to the Bernoulli

Mirror averaging with sparsity priors

ARNAK S. DALALYAN[1] and ALEXANDRE B. TSYBAKOV[2]

[1]*LIGM/IMAGINE, Ecole des Ponts ParisTech, Universit Paris Est*
E-mail: dalalyan@imagine.enpc.fr

[2]*Laboratoire de Statistique, CREST and LPMA, Université Paris 6*
E-mail: Alexandre.Tsybakov@ensae.fr

We consider the problem of aggregating the elements of a possibly infinite dictionary for building a decision procedure that aims at minimizing a given criterion. Along with the dictionary, an independent identically distributed training sample is available, on which the performance of a given procedure can be tested. In a fairly general set-up, we establish an oracle inequality for the Mirror Averaging aggregate with any prior distribution. By choosing an appropriate prior, we apply this oracle inequality in the context of prediction under sparsity assumption for the problems of regression with random design, density estimation and binary classification.

Keywords: Mirror averaging, sparsity, aggregation of estimators, oracle inequalities.

1. Introduction

In recent years several methods of estimation and selection under the sparsity scenario have been discussed in the literature. The ℓ_1-penalized least squares (Lasso) is by far the most studied one and its statistical properties are now well understood (cf., e.g., [6, 11, 12, 13, 38, 49, 54, 57, 58] and the references cited therein). Several other estimators are closely related to the Lasso, such as the Elastic net [59], the Dantzig selector [15], the adaptive Lasso [60], the least squares with entropy or $\ell_{1+\delta}$ penalization [33, 34], etc. These estimators are obtained as solutions of convex or linear programming problems and are attractive by their low computational cost. However, they have good theoretical properties only under rather restrictive assumptions, such as the mutual coherence assumption [24], the uniform uncertainty/restricted isometry principle [15], the irrepresentable [58] or the restricted eigenvalue [6] conditions. Roughly speaking, these conditions mean that, for example, in the linear regression context one should assume that the Gram matrix of the predictors is not too far from the identity matrix. Such type of assumption is natural if we want to identify the parameters or to retrieve the sparsity pattern, but it is not necessary if we are interested only in the prediction ability.

Indeed, at least in theory, there exist estimators attaining sufficiently good accuracy of prediction under almost no assumption on the Gram matrix. This is, in particular, the case for the ℓ_0-penalized least squares estimator [10, Thm. 3.6], [12, Thm. 3.1]. However, in practice this estimator can be unstable (cf. [8]). Furthermore, its computation is an NP-hard problem, and there is a challenge to find a method realizing a compromise between the theoretical optimality and computational efficiency. Motivated by this, we proposed in [20, 21, 22, 23] an approach to estimation under the sparsity scenario, which is quite different from the ℓ_1 penalization techniques. The idea is to use an exponentially weighted aggregate (EWA) with a properly chosen sparsity-favoring prior. Let us note that there exists an extensive literature on EWA, which does not discuss the sparsity issue. Thus, procedures with exponential weighting are quite common in the context of on-line learning with deterministic data, see [18, 32, 50], the monograph [19] and the references cited therein. Statistical properties of various versions of EWA are discussed in [2, 3, 9, 16, 17, 28, 30, 31, 39, 51, 52, 53, 56].

On the difference from these works, we focus in [20, 21, 22, 23] on the ability of EWA to deal with the sparsity issue. Specifically, we prove that EWA with a properly chosen prior satisfies sparsity oracle inequalities (SOI), which are comparable with those for the ℓ_0-penalized techniques and are even better in some aspects. At the same time, on the difference from the ℓ_0-penalized methods, our method is computationally feasible for relatively large dimensions of the problem, cf. [23]. Furthermore, our estimator has theoretical advantages as compared to the ℓ_1-penalized methods, since it satisfies oracle inequalities with leading constant 1 that hold with almost no assumption on the dictionary/Gram matrix (cf. detailed comparison with the ℓ_1 based methods in Section 8 below).

The results of [20, 21, 22, 23] are established for the linear regression model with fixed design. The aim of this paper is to show that similar ideas can be successfully implemented for a large scope of statistical problems with i.i.d. data, in particular, for regression with random design, density estimation and classification. The procedure that we propose is mirror averaging (MA) with sparsity priors. The difference from the EWA considered in [20, 21, 22, 23] is that we compute the exponential weights recursively and then average them out.

This paper is organized as follows. In Section 2 we introduce some notation and formulate main assumptions. Section 3 contains the definition of the MA estimator and a general PAC-Bayesian risk bound in expectation. In Section 4 we introduce our sparsity prior and obtain our main SOI as a corollary of the PAC-Bayesian bound. Sections 5, 6 and 7 consider applications of this result to specific models, namely, to nonparametric regression with random design, density estimation and classification. In Section 8 we briefly discuss computational aspects of the MA aggregate and compare it to other methods of sparse estimation. Technical proofs are given in the appendix.

2. Notation and assumptions

Let $(\mathcal{Z}, \mathfrak{F})$ be a measurable space and let $\{P_f,\ f \in \mathcal{F}\}$ be a collection of probability measures on $(\mathcal{Z}, \mathfrak{F})$ indexed by some set $\mathcal{F}$. We are interested the estimation of f based on an i.i.d. sample $Z_1, \ldots, Z_n$ drawn from the probability distribution P_f. We will assume that f is a "functional" parameter, that is $\mathcal{F}$ is a subset of a vector space $\mathcal{E} = \{f : \mathcal{X} \to \mathbb{R}^d\}$ for some set $\mathcal{X}$ and for some positive integer d. From now on, we denote by $\mathbf{E}_f$ the expectation w.r.t. P_f and by $\mathbf{Z}$ the random vector $(Z_1, \ldots, Z_n) \in \mathcal{Z}^n$.

To further specify the settings, let $\ell : \mathcal{E} \times \mathcal{F} \to \mathbb{R}_+$ be a general loss function. An estimator of f is any mapping $\tilde{f} : \mathcal{Z}^n \to \mathcal{E}$ such that the mapping $\mathbf{z} \mapsto \ell(\tilde{f}(\mathbf{z}), f)$, defined on $(\mathcal{Z}^n, \mathfrak{F}^n)$ and with values in $\mathbb{R}_+$, is measurable for every $f \in \mathcal{F}$. The performance of an estimator $\tilde{f}$ is quantified by the risk

$$\mathbf{E}_f[\ell(\tilde{f}(\mathbf{Z}), f)] := \int_{\mathcal{Z}^n} \ell(\tilde{f}(\mathbf{z}), f)\, P_f^n(d\mathbf{z}).$$

Here P_f^n stands for the product measure $P_f \otimes \ldots \otimes P_f$ on $(\mathcal{Z}^n, \mathfrak{F}^n)$. We will assume the following.

Assumption Q1: There exists a mapping $Q : \mathcal{Z} \times \mathcal{E} \to \mathbb{R}$ such that, for every $f \in \mathcal{F}$,

- the mapping $z \mapsto Q(z, g)$ is measurable and P_f-integrable for every $g \in \mathcal{E}$,
- $\Delta(f) \triangleq \int_{\mathcal{Z}} Q(z, g)\, P_f(dz) - \ell(g, f)$ is independent of g and finite for any $f \in \mathcal{F}$.

Assumption Q1 is fulfilled in a number of settings; detailed discussion is given in Sections 5-7. For example, in the case of regression with squared loss, one has $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and $\ell(g, f) = \int_{\mathcal{X}} (g - f)^2 dP_X$, where P_X stands for the distribution of the design and f is the regression function. Assumption Q1 is then fulfilled with $Q(z, g) = (y - g(x))^2$. In simple words, assumption Q1 requires the existence of an unbiased estimator of the risk $\ell(g, f)$, up to a summand depending excusively on f, where f is the unknwon parameter and g is a known function. It is worth noting that under assumption Q1 the minimizer of the loss function $g \mapsto \ell(g, f)$ coincides with the minimizer of the expectation $g \mapsto \int Q(Z, g) P_f(dZ)$. This property is crucial in what follows.

Since, in general, there is no estimator having the smallest possible risk among all possible estimators, we will pursue a more realistic goal, which consists in finding an estimator whose risk, for every f, is nearly as small as the minimal risk $\min_{g \in \mathcal{F}_\Lambda} \ell(g, f)$ over a pre-specified subset $\mathcal{F}_\Lambda$ of $\mathcal{E}$, i.e., we will follow the oracle approach. To make this approach sensible, the subfamily $\mathcal{F}_\Lambda$ should not be too large. On the other hand, it should be chosen large enough to contain a good approximation to the (unknown) "true" function f.

The set $\mathcal{F}_\Lambda$ is indexed by the elements of some measurable space $(\Lambda, \mathfrak{L})$. More precisely, we define $\mathcal{F}_\Lambda = \{f_{\boldsymbol{\lambda}}, \boldsymbol{\lambda} \in \Lambda\} \subset \mathcal{E}$ as a collection of functions (dictionary) such that, for every $x \in \mathcal{X}$ and $z \in \mathcal{Z}$, the mappings $\boldsymbol{\lambda} \mapsto f_{\boldsymbol{\lambda}}(x)$, $\boldsymbol{\lambda} \mapsto Q(z, f_{\boldsymbol{\lambda}})$ and $\boldsymbol{\lambda} \mapsto \ell(f_{\boldsymbol{\lambda}}, f)$ from Λ to $\mathbb{R}$ are measurable. The elements of the dictionary $\mathcal{F}_\Lambda$ can be interpreted as candidate estimators of f. Define $\mathcal{P}_\Lambda$ as the set of all probability measures on $(\Lambda, \mathfrak{L})$ and $\mathcal{P}_1(\mathcal{F}_\Lambda)$ as the set of all measures $\mu \in \mathcal{P}_\Lambda$ such that $\int_\Lambda |f_{\boldsymbol{\lambda}}(x)|\, \mu(d\boldsymbol{\lambda}) < \infty$ for every $x \in \mathcal{X}$. We define for every $\mu \in \mathcal{P}_1(\mathcal{F}_\Lambda)$,

$$\mathsf{f}_\mu = \int_\Lambda f_{\boldsymbol{\lambda}} \mu(d\boldsymbol{\lambda}) \qquad \Big(\mathsf{f}_\mu(x) = \int_\Lambda f_{\boldsymbol{\lambda}}(x) \mu(d\boldsymbol{\lambda}),\ \forall x \in \mathcal{X}\Big).$$

We say that f_μ is a convex aggregate of functions $f_{\boldsymbol{\lambda}}$ with μ being the mixing measure or the measure of aggregation. The estimators we study in the present work are convex aggregates with data-dependent mixing measures.

In what follows, we denote by $\mathcal{C}(\mathcal{F}_\Lambda)$ the set of all convex aggregates of functions $f_{\boldsymbol{\lambda}}$, that is

$$\mathcal{C}(\mathcal{F}_\Lambda) = \big\{g : \mathcal{X} \to \mathbb{R} \ \text{ s.t. } g = \mathsf{f}_\mu \text{ for some } \mu \in \mathcal{P}_1(\mathcal{F}_\Lambda)\big\}.$$

It is clear that $\mathcal{C}(\mathcal{F}_\Lambda)$ is a convex set containing $\mathcal{F}_\Lambda$. For our main result we need the following condition on the function Q appearing in Assumption Q1.

Assumption Q2: There exist $\beta > 0$ and a mapping $\Psi_\beta : \mathcal{C}(\mathcal{F}_\Lambda) \times \mathcal{C}(\mathcal{F}_\Lambda) \to \mathbb{R}_+$ such that

i) $\Psi_\beta(g, g) = 1$ for all $g \in \mathcal{C}(\mathcal{F}_\Lambda)$,
ii) the mapping $g \mapsto \Psi_\beta(g, \tilde g)$ is concave on $\mathcal{C}(\mathcal{F}_\Lambda)$ for every fixed $\tilde g \in \mathcal{C}(\mathcal{F}_\Lambda)$,
iii) the inequality

$$\int_{\mathcal{Z}} \exp\big(-\beta^{-1}\big\{Q(z, g) - Q(z, \tilde g)\big\}\big)\, P_f(dz) \leqslant \Psi_\beta(g, \tilde g)$$

holds for every $g, \tilde g \in \mathcal{C}(\mathcal{F}_\Lambda)$.

At first sight, this assumption seems cumbersome but we will show that it holds for a number of settings which are of central interest in nonparametric statistics. For example, in the model of regression with random design and additive Gaussian noise, Assumption Q2 is fulfilled for $\beta \geqslant 2\sigma^2 + 2\sup_\lambda \|f_{\boldsymbol\lambda} - f\|_\infty^2$, where σ^2 is the noise variance and f is the unknown regression function. Assumption Q2 has been first introduced in [30, Theorem 4.2] for finite dictionaries and a variant of it has been used in [3, Corollary 5.1].

Note also that if Assumption Q2 is satisfied for some (β, Ψ_β), then it is so for $(\beta', \Psi_\beta^{\beta/\beta'})$ with any $\beta' > \beta$. In fact, condition ii) is ensured due to the concavity of the function $t \mapsto t^{\beta/\beta'}$ on $[0, \infty)$, while iii) can be checked using the Hölder inequality.

3. Mirror averaging and a PAC-Bayesian bound in expectation

We now introduce the mirror averaging (MA) estimator. First, we fix a prior $\pi \in \mathcal{P}_1(\mathcal{F}_\Lambda)$, a "temperature" parameter $\beta > 0$, and set

$$\widehat\theta_{m,\boldsymbol\lambda}(\mathbf{Z}) = \frac{\exp\big\{ -\frac{1}{\beta}\sum_{i=1}^m Q(Z_i, f_{\boldsymbol\lambda})\big\}}{\int_\Lambda \exp\big\{ -\frac{1}{\beta}\sum_{i=1}^m Q(Z_i, f_w)\big\}\pi(dw)},$$

$$\widehat\theta_{\boldsymbol\lambda} = \widehat\theta_{\boldsymbol\lambda}(\mathbf{Z}) = \frac{1}{n+1}\sum_{m=0}^n \widehat\theta_{m,\boldsymbol\lambda}(\mathbf{Z})$$

with $\widehat\theta_{0,\boldsymbol\lambda}(\mathbf{Z}) \equiv 1$. For every fixed $\mathbf{Z}$, $\widehat\theta_{\boldsymbol\lambda}$ is a probability density on Λ with respect to the probability measure π. Let $\widehat\mu_n$ be the probability measure on $(\Lambda, \mathfrak{L})$ having $\widehat\theta_{\boldsymbol\lambda}$ as density w.r.t. π. By analogy with the Bayesian context, one can call $\widehat\theta_{\boldsymbol\lambda}$ and $\widehat\mu_n$ the posterior density and the posterior probability, respectively. Following [30] where the case of discrete π was considered (see also [41]), we define the *MA aggregate* as the corresponding posterior mean $\widehat f_n = \mathsf{f}_{\widehat\mu_n}$, that is $\widehat\mu_n(d\boldsymbol\lambda) = \frac{1}{n+1}\sum_{m=0}^n \widehat\theta_{m,\boldsymbol\lambda}(\mathbf{Z})\,\pi(d\boldsymbol\lambda)$ and

$$\widehat f_n(\mathbf{Z}, x) = \int_\Lambda f_{\boldsymbol\lambda}(x)\,\widehat\theta_{\boldsymbol\lambda}(\mathbf{Z})\,\pi(d\boldsymbol\lambda) = \frac{1}{n+1}\sum_{m=0}^n \int_\Lambda f_{\boldsymbol\lambda}(x)\,\widehat\theta_{m,\boldsymbol\lambda}(\mathbf{Z})\,\pi(d\boldsymbol\lambda). \tag{1}$$

To simplify the notation we suppress the dependence of $\widehat f_n$ on $\mathbf{Z}$ and x when it causes no ambiguity.

Theorem 1 (PAC-Bayesian bound in expectation)**.** *If Assumptions Q1 and Q2 are fulfilled, then the MA aggregate $\widehat{f}_n$ satisfies the following oracle inequality*

$$\mathbf{E}_f[\ell(\widehat{f}_n, f)] \leqslant \inf_{p\in\mathcal{P}_\Lambda} \Big(\int_\Lambda \ell(f_{\boldsymbol{\lambda}}, f)\, p(d\boldsymbol{\lambda}) + \frac{\beta\mathcal{K}(p,\pi)}{n+1}\Big), \tag{2}$$

where $\mathcal{K}(p,\pi)$ stands for the Kullback-Leibler divergence

$$\mathcal{K}(p,\pi) = \begin{cases} \int_\Lambda \log\big(\frac{dp}{d\pi}(\boldsymbol{\lambda})\big)\, p(d\boldsymbol{\lambda}), & \text{if } p \ll \pi, \\ +\infty, & \text{otherwise} \end{cases}.$$

Proof of Theorem 1 is given in the appendix. It is based on a cancellation argument that can be traced back to Barron [4].

The oracle inequality of Theorem 1 is in the line of the PAC-Bayesian bounds initiated in [42] and is applicable to a large variety of models. Some particularly relevant examples will be treated in Sections 5-7. An interesting feature of Theorem 1 is that it is valid for a large class of prior distributions.

The fact that (2) holds true for convex mappings $g \mapsto Q(Z, g)$ has been discussed informally in [3], p. 1606, as a consequence of an oracle inequality for a randomized estimator. A difference of Theorem 1 from the approach in [3] is that the convexity of the loss function is not required.

***Remark* 1.** *If the cardinality of $\mathcal{F}_\Lambda$ is finite, say $\mathrm{card}(\mathcal{F}_\Lambda) = N$ and $\Lambda = \{1,\ldots,N\}$, inequality (2) implies that*

$$\mathbf{E}_f[\ell(\widehat{f}_n, f)] \leqslant \min_{j=1,\ldots,N} \Big(\ell(f_j, f) + \frac{\beta\log\pi_j}{n+1}\Big).$$

Oracle inequalities of this type and similar under different sets of assumptions were established earlier by several authors (cf. [16, 17, 51, 52, 53, 9, 30, 3] and the references therein for closely related results). Our PAC-Bayesian bound (2) generalizes the oracle inequality of [30, Thm. 3.2] to arbitrary, not necessarily finite, family $\mathcal{F}_\Lambda$. In the settings that we study below it is crucial to consider uncountable $\mathcal{F}_\Lambda$. As we will see later, this generalization allows us to take advantage of sparsity and suggests a powerful alternative to the classical model selection approach.

***Remark* 2.** *For the regression model with additive noise and deterministic design, PAC-Bayesian bounds in expectation on the empirical l_2-norm similar to (2) have been obtained in [20, 21, 22, 23] for an EWA, which does not contain the step of averaging. Earlier [39] proved a similar result for the special case of finite $\mathrm{card}(\mathcal{F}_\Lambda)$ and Gaussian errors. In the notation of the present paper, the aggregate studied in those works is of the form $\check{f}_n = \int_\Lambda f_{\boldsymbol{\lambda}}\widehat{\theta}_{n,\boldsymbol{\lambda}}\pi(d\boldsymbol{\lambda})$. Interestingly, in a very recent paper Lecué and Mendelson [37] proved that $\check{f}_n$ does not satisfy inequality (2) in the case of i.i.d. observations.*

Finally, we note that the results of this work hold only for proper priors. However, it is very likely that Theorem 1 extends to the case of improper priors under some additional assumption ensuring, for instance, that the integral $\int_\Lambda \exp\big\{-\frac{1}{\beta}\sum_{i=1}^m Q(Z_i, f_w)\big\}\pi(dw)$ appearing in the definition of the MA estimator is finite.

4. Sparsity oracle inequality

In this section we introduce a prior π that we recommend to use for the MA aggregate under the sparsity scenario. Then we prove a sparsity oracle inequality (SOI) leading to some natural choices of the tuning parameters of the prior.

4.1. Sparsity prior and SOI

In what follows we assume that $\Lambda \subseteq \mathbb{R}^M$ for some integer $M \geqslant 2$. We will use bold face letters to denote vectors and, in particular, the elements of Λ. We denote by $\mathrm{Tr}(\mathsf{A})$ the trace of a square matrix A. To deal with integrals of the type $\int_\Lambda \ell(f_{\boldsymbol{\lambda}}, f)\, p(d\boldsymbol{\lambda})$ we introduce the following additional assumption.

Assumption L: For every fixed $f \in \mathcal{F}$, there exists a measurable set $\Lambda_0 \subset \Lambda$ such that $\Lambda \setminus \Lambda_0$ has zero Lebesgue measure and the mapping $L_f : \Lambda_0 \to \mathbb{R}$, where $L_f(\boldsymbol{\lambda}) = \ell(f_{\boldsymbol{\lambda}}, f)$, is twice differentiable. Furthermore, there exists a symmetric $M \times M$ matrix $\mathcal{M}$ such that $\mathcal{M} - \nabla^2 L_f(\boldsymbol{\lambda})$ is positive semi-definite for every $\boldsymbol{\lambda} \in \Lambda_0$, where $\nabla^2 L_f(\boldsymbol{\lambda})$ stands for the Hessian matrix.

We are interested in covering the case of large M, possibly much larger than the sample size n. We will be working under the sparsity assumption, i.e., when there exists $\boldsymbol{\lambda}^* \in \mathbb{R}^M$ such that f is close to $f_{\boldsymbol{\lambda}^*}$ and $\boldsymbol{\lambda}^*$ has a very small number of non-zero components. We argue that an efficient way for handling this situation is based on a suitable choice of the prior π. To be more precise, our results will show how to take advantage of sparsity for the purpose of prediction and not for accurate estimation of the parameters or selection of the sparsity pattern. Thus, if the underlying model is sparse, we do not prove that our estimated model is sparse as well, but we claim that it has a small prediction risk under very mild assumptions. Nevertheless, we have a numerical evidence that our method can also recover very accurately the true sparsity pattern [22, 23]. We observed this in examples where the restrictive assumptions mentioned in the Introduction are satisfied.

Let τ and R be positive numbers. The *sparsity prior* is defined by

$$\pi(d\boldsymbol{\lambda}) = \frac{1}{C_{\tau,R}} \Big\{ \prod_{j=1}^{M} (\tau^2 + \lambda_j^2)^{-2} \Big\} \mathbb{1}(\|\boldsymbol{\lambda}\|_1 \leqslant R)\, d\boldsymbol{\lambda}, \tag{3}$$

where $\|\boldsymbol{\lambda}\|_1 = \sum_j |\lambda_j|$ stands for the ℓ_1-norm, $\mathbb{1}(\cdot)$ denotes the indicator function, and $C_{\tau,R}$ is a normalizing constant such that π is a probability density.

The prior (3) has a simple heuristical interpretation. Note first that R is a regularization parameter, which is typically very large. So, in a rough approximation we may consider that the factor $\mathbb{1}(\|\boldsymbol{\lambda}\|_1 \leqslant R)$ is almost equal to one. Thus, π is essentially a product of M rescaled Student's distributions. Precisely, we deal with the distribution of $\sqrt{2}\tau\mathbf{Y}$, where $\mathbf{Y}$ is a random vector with i.i.d. coordinates drawn from Student's t with three degrees of freedom. In the examples below we choose a very small τ, smaller than $1/n$. Therefore, most of the coordinates of $\tau\mathbf{Y}$ are very close to zero. On the other hand, since Student's distribution has heavy tails, there exists a small portion of coordinates of $\tau\mathbf{Y}$ that are quite far from zero.

The relevance of heavy tailed priors for dealing with sparsity has been emphasized by several authors (see [46, Section 2.1] and references therein). Most of this work is focused on logarithmically concave priors, such as the multivariate Laplace distribution. Also in wavelet estimation on classes of "sparse" functions [29] and [44] invoke quasi-Cauchy and Pareto priors respectively. Bayes estimators with heavy-tailed priors in sparse Gaussian shift models are discussed in [1].

We are now in a position to state the SOI for the MA aggregate with the sparsity prior. The result is even more general because it holds not only for the MA aggregate but for any estimator satisfying (2) with the sparsity prior.

Theorem 2. *Let $\widehat{f}_n$ be any estimator satisfying inequality (2), where the loss function ℓ satisfies Assumption L and π is the sparsity prior defined as above. Assume that Λ contains the set $B_1(R) = \big\{\boldsymbol{\lambda} \in \mathbb{R}^M \mid \|\boldsymbol{\lambda}\|_1 \leqslant R\big\}$ with $R > 2M\tau$. Then for all $\boldsymbol{\lambda}^*$ such that $\|\boldsymbol{\lambda}^*\|_1 \leqslant R-2M\tau$ we have*

$$\mathbf{E}_f[\ell(\widehat{f}_n, f)] \leqslant \ell(f_{\boldsymbol{\lambda}^*}, f) + \frac{4\beta}{n+1}\sum_{j=1}^M \log(1+\tau^{-1}|\lambda_j^*|) + \mathsf{R}(M,\tau), \tag{4}$$

where the residual term is $\mathsf{R}(M,\tau) = 4\tau^2\mathrm{Tr}(\mathcal{M}) + \frac{\beta}{n+1}$.

Proof of Theorem 2 is deferred to Section 9.3 of the appendix.

As follows from (4), the main term of the excess risk $\mathbf{E}_f[\ell(\widehat{f}_n, f)] - \ell(f_{\boldsymbol{\lambda}^*}, f)$ is proportional to $\sum_{j=1}^M \log(1+\tau^{-1}|\lambda_j^*|)$. Importantly, the number of nonzero elements in this sum is equal to the number of nonzero components of $\boldsymbol{\lambda}^*$ that we will further denote by $\|\boldsymbol{\lambda}^*\|_0$. Therefore, for sparse vectors $\boldsymbol{\lambda}^*$ this term is rather small. But still, in all the examples that we consider below, it dominates the remainder term $\mathsf{R}(M,\tau)$, which is made negligible by choosing a sufficiently small τ, for instance, $\tau = O((\mathrm{Tr}(\mathcal{M})n)^{-1/2})$.

Theorem 2 implies the following bound involving only the ℓ_0 norm and the upper bound R on the ℓ_1 norm of $\boldsymbol{\lambda}^*$.

Corollary 1. *If some estimator $\widehat{f}_n$ satisfies the oracle inequality of Theorem 2, then*

$$\mathbf{E}_f[\ell(\widehat{f}_n, f)] \leqslant \ell(f_{\boldsymbol{\lambda}^*}, f) + \frac{4\beta\|\boldsymbol{\lambda}^*\|_0 \log(1+R\tau^{-1})}{n+1} + \mathsf{R}(M,\tau),$$

where $\boldsymbol{\lambda}^*$ and $\mathsf{R}(M,\tau)$ are as in Theorem 2.

Proof. Set $M^* = \|\boldsymbol{\lambda}^*\|_0$ for brevity. Using Jensen's inequality, we get

$$\frac{1}{M^*}\sum_{j=1}^M \log(1+\tau^{-1}|\lambda_j^*|) \leqslant \log\big(1+(\tau M^*)^{-1}\|\boldsymbol{\lambda}^*\|_1\big).$$

Using the inequalities $\|\boldsymbol{\lambda}^*\|_1 \leqslant R$ and $M^* \geqslant 1$, the desired inequality follows. $\square$

Note that the sparsity oracle inequalities (SOI) stated in this section are valid not only for the MA aggregate but for any other estimator (whose definition involves a prior π) satisfying a PAC-Bayesian bound similar to (2), possibly with some additional residual terms that should be then added in the SOI as well. Examples of such estimators can be found in [3].

Remark 3. *Assumption L need not be satisfied exactly. In fact, $L_f(\cdot)$ need not even be differentiable. Inspection of the proof of Theorem 2 reveals that if $L_f(\boldsymbol{\lambda})$ is well approximated by a smooth function $\bar{L}_f(\boldsymbol{\lambda})$, that is $0 \leqslant L_f(\boldsymbol{\lambda}) - \bar{L}_f(\boldsymbol{\lambda}) \leqslant \varepsilon$, $\forall \boldsymbol{\lambda}$, for some small $\varepsilon > 0$ and if $\bar{\mathcal{M}}_\varepsilon - \nabla^2 \bar{L}_f$ is positive semidefinite, then the conclusions of Theorem 2 hold with a modified residual term*

$$\mathsf{R}_\varepsilon(M,\tau) = \varepsilon + 4\tau^2 \mathrm{Tr}(\bar{\mathcal{M}}_\varepsilon) + \frac{\beta}{n+1}.$$

This remark will be useful for studying the problem of classification under the hinge loss where the function L_f is not differentiable, cf. Section 7.

4.2. Choice of the tuning parameters

The above sparsity oracle inequalities suggest some guidelines for the choice of tuning parameters τ and R:

1. Parameter τ should be chosen very carefully : It should be small enough to guarantee the negligibility of the residual term but not exponentially small to prevent the explosion of the main term of the risk. A reasonable choice (which is not the only possible) for τ is

$$\tau = \min\Big(\frac{\sqrt{\beta}}{\sqrt{Mn}}\,, \frac{R}{4M}\Big)\,. \tag{5}$$

   For this choice of τ we have:
   (a) the residual term $\mathsf{R}(M,\tau)$ is at most of order β/n,
   (b) the terms $\log(1+|\lambda_j^*|/\tau)$ increase at most logarithmically in M and in n under the condition that $\mathrm{Tr}(\mathcal{M})$ increases not faster than a power of M. Note that $\mathrm{Tr}(\mathcal{M}) = O(M)$ in all the examples that we consider below.
   (c) the MA aggregate is accurate enough if there exists a sparse vector $\boldsymbol{\lambda}^*$, with ℓ_1-norm bounded by $R/2$ which provides a good approximation $f_{\boldsymbol{\lambda}^*}$ of f,
2. It is clear that one should choose R as large as possible in order to cover the broadest class of possible values $\boldsymbol{\lambda}^*$. However, we are not aware of any example where Assumption Q2 holds with finite β for $R = +\infty$ or, equivalently, for $\Lambda = \mathbb{R}^M$. Therefore, we assume that R is an a priori chosen large parameter and interpret the above results as follows: If there is a sparse vector $\boldsymbol{\lambda}^*$ such that $\ell(f_{\boldsymbol{\lambda}^*}, f)$ is small and $\|\boldsymbol{\lambda}^*\|_1 \leqslant R - 2M\tau$, then the MA aggregate has a small prediction risk.

Remark 4. *The choice $\tau = \min\Big(\frac{\sqrt{\beta}}{\sqrt{\mathrm{Tr}(\mathcal{M})n}}\,, \frac{R}{4M}\Big)$ ensures that the estimator $\widehat{f}_n$ is invariant with respect to an overall scaling of $\boldsymbol{\lambda}$. More precisely, if instead of considering the parametrization $\{f_{\boldsymbol{\lambda}} : \|\boldsymbol{\lambda}\|_1 \leqslant R\}$ we consider the parametrization $\{\tilde{f}_{\boldsymbol{\omega}} : \|\boldsymbol{\omega}\|_1 \leqslant R/s\}$ with $\tilde{f}_{\boldsymbol{\omega}} = f_{s\boldsymbol{\omega}}$*

for some $s > 0$, then the MA aggregate based on the prior defined by (3) remains unchanged. This can be easily checked by the change of variables using the relation $\tilde{\mathcal{M}} = s^2\mathcal{M}$ where $\tilde{\mathcal{M}}$ denotes the Hessian matrix analogous to $\mathcal{M}$ for the dictionary $\{\tilde{f_{\boldsymbol{\omega}}}\}$.

Along with choosing the parameters (τ, R) of the prior, one needs to choose the "temperature" parameter β. A model-free choice of β seems to be impossible. In fact, even the existence of β such that Assumption Q2 holds is not ensured for every model. Some more discussion of the choice of β is given in Remark 7 below.

5. Application to regression with random design

5.1. Regression estimation in L^2-norm

Let $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and we have the i.i.d. observations $Z_i = (X_i, Y_i)$, $i = 1, \dots, n$ with $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$. We define the regression function by $f(x) = \mathbf{E}(Y_1|X_1 = x)$, $\forall x \in \mathcal{X}$, and assume that the errors

$$\xi_i = Y_i - f(X_i), \qquad i = 1, \dots, n,$$

are such that $\mathbf{E}[\xi_1^2] < \infty$. Then $\mathbf{E}(\xi_i|X_i) = 0$. Let P_X denote the distribution of X_1. For $s \in [1, \infty]$ we denote by $\|\cdot\|_{P_X,s}$ the L^s-norm with respect to P_X. We also denote by $\langle\cdot,\cdot\rangle_{P_X}$ to the scalar product in $L^2(\mathcal{X}, P_X)$. Throughout this section we consider the integrated squared loss $\ell(f,g) = \|f - g\|^2_{P_X,2}$. Then it is easy to check that Assumption Q1 is fulfilled with

$$Q(z,g) = (y - g(x))^2, \qquad z = (x,y) \in \mathcal{Z}.$$

Furthermore, we focus on the particular case where $\mathcal{F}_\Lambda$ is a convex subset of the vector space spanned by a finite number of measurable functions $\{\phi_j\}_{j=1,\dots,M} \subset L^2(\mathcal{X}, P_X)$, that is

$$\mathcal{F}_\Lambda = \Big\{ f_{\boldsymbol{\lambda}} = \sum_{j=1}^{M} \lambda_j \phi_j \;\Big|\; \boldsymbol{\lambda} \in \mathbb{R}^M \text{with } \|\boldsymbol{\lambda}\|_1 \leqslant R \Big\} \tag{6}$$

for some $R > 0$. Then assumption L holds with $\mathcal{M}$ being the matrix with entries $\langle\phi_j, \phi_{j'}\rangle_{P_X}$, which will be referred to as the Gram matrix. This definition of $\mathcal{M}$ will be used throughout this section. The collection of functions $\{\phi_1, \dots, \phi_M\}$ will be called the *dictionary*.

***Remark* 5.** *The value of the parameter τ presented in (5) does not allow us to take into account the possible inhomogeneity of functions ϕ_j. One way of dealing with the inhomogeneity is to let τ depend on j in the definition of the sparsity prior π. In this paper we consider for brevity a less general approach, which is common in the literature on sparsity. Namely, we normalize the functions ϕ_j in advance and use the same τ for all coordinates of $\boldsymbol{\lambda}$. The normalization is done by rescaling the functions ϕ_j so that all the diagonal entries of the Gram matrix $\mathcal{M}$ are equal to one.*

Following this remark, we assume that the functions ϕ_j are such that $\|\phi_j\|_{P_X,2} = 1$ for every j. Therefore, $\mathrm{Tr}(\mathcal{M}) = M$.

Proposition 1. *Assume that for some constant $L_\phi > 0$ we have $\max_{j=1,\dots,M}\|\phi_j\|_{P_X,\infty} \leqslant L_\phi$. If, in addition, the errors ξ_i have a bounded exponential moment:*

$$\exists\, b, \sigma^2 > 0 \quad \textit{such that} \quad \mathbf{E}(e^{t\xi_1}|X_1) \leqslant e^{\sigma^2 t^2/2}, \quad \forall\, |t| \leqslant b, \qquad P_X\textit{-a.s.}, \tag{7}$$

then, for every $\beta \geqslant \max\big(2\sigma^2 + 2\sup_{\boldsymbol{\lambda}\in\Lambda}\|f_{\boldsymbol{\lambda}} - f\|^2_{P_X,\infty}, 4RL_\phi/b\big)$, the MA aggregate $\widehat{f}_n$ defined by (1) with the sparsity prior (3) satisfies

$$\mathbf{E}_f[\|\widehat{f}_n - f\|^2_{P_X,2}] \leqslant \inf_{\boldsymbol{\lambda}^*}\left\{\|f_{\boldsymbol{\lambda}^*} - f\|^2_{P_X,2} + \frac{4\beta}{n+1}\sum_{j=1}^M \log(1+\tau^{-1}|\lambda^*_j|)\right\} + \mathsf{R}(M,\tau) \tag{8}$$

where the $\inf$ *is taken over all $\boldsymbol{\lambda}^*$ such that $\|\boldsymbol{\lambda}^*\|_1 \leqslant R - 2M\tau$ and $\mathsf{R}(M,\tau) = 4\tau^2 M + \frac{\beta}{n+1}$.*

Proof of Proposition 1. In view of Theorem 2, it suffices to check that Assumption Q2 is fulfilled for $\beta \geqslant \max\big(2\sigma^2 + 2\sup_{\boldsymbol{\lambda}\in\Lambda}\|f_{\boldsymbol{\lambda}} - f\|^2_{P_X,\infty}, 4RL_\phi/b\big)$. This is done along the lines of the proof of [30, Corollary 5.5]. We omit the details. □

Proposition 1 can be used in signal denoising under the sparsity assumption. A typical issue studied in statistical literature, as well as in the literature on signal processing, is to estimate a signal f based on its noisy version recorded at some points $X_1, \dots, X_n$, under the assumption that f admits a sparse representation w.r.t. some given dictionary $\{\phi_j; j = 1,\dots,M\}$. By sparse representation we mean a linear combination of a small number of functions ϕ_j. Assume for the moment that the noise satisfies (7) with $b = +\infty$ and some known $\sigma \in [0,\infty)$ and that the unknown signal is bounded by some constant that can be assumed to be equal to 1. The latter assumption is fulfilled in many applications, as for example in image processing.

The method that we suggest for estimating a sparse representation of f, under the assumption $M \geqslant n$, consists of:

a) normalizing the functions ϕ_j,
b) fixing a parameter $R > 0$,
c) setting
$$\beta = 2\sigma^2 + 2(RL_\phi + 1)^2, \quad \tau = \min\Big(\frac{\sqrt{\beta}}{\sqrt{\mathrm{Tr}(\mathcal{M})n}}, \frac{R}{4M}\Big), \tag{9}$$
d) computing the MA aggregate $\widehat{f}_n = \sum_{j=1}^M \widehat{\lambda}_j\phi_j$ with coefficients $\widehat{\lambda}_j = \int_{\mathbb{R}^M}\lambda_j\widehat{\theta}_{\boldsymbol{\lambda}}\pi(d\boldsymbol{\lambda})$ based on the sparsity prior (3) and the posterior density
$$\widehat{\theta}_{\boldsymbol{\lambda}} = \frac{1}{n+1}\sum_{m=0}^{n+1}\frac{\exp\big\{-\frac{1}{\beta}\sum_{i=1}^m (Y_i - f_{\boldsymbol{\lambda}}(X_i))^2\big\}}{\int_\Lambda \exp\big\{-\frac{1}{\beta}\sum_{i=1}^m (Y_i - f_{\boldsymbol{w}}(X_i))^2\big\}\pi(d\boldsymbol{w})}\ .$$

In view of Proposition 1, if we run this procedure with some value $R > 0$, we will get accurate estimates for signals that are well approximated by a sparse linear combination of functions ϕ_j, provided that the coefficients of this linear combination have an ℓ_1-norm bounded by $R - 2M\tau$. In most of the problems arising in signal or image processing the ℓ_1-norm of the best sparse approximation to the signal is unknown. It is therefore important to make a data-driven choice of R. Let us outline one possible way to do this. Consider that only the signals formed by a linear combination of at most M^* functions ϕ_j are of interest, and assume that the dictionary $\{\phi_j\}$ satisfies the restricted isometry property (RIP) of order M^*, see equation (1.3) in [15] for the definition. In other terms, assume that $f \approx f_{\boldsymbol{\lambda}^*}$ with $\|\boldsymbol{\lambda}^*\|_0 \leqslant M^*$ and $\|f_{\boldsymbol{\lambda}^*}\|_{P_X,2} \geqslant \frac{1}{2}\|\boldsymbol{\lambda}^*\|_2$ where $\|\cdot\|_2$ is the Euclidean norm. Then we can bound the ℓ_1-norm of $\boldsymbol{\lambda}^*$ as follows:

$$\|\boldsymbol{\lambda}^*\|_1 \leqslant \sqrt{M^*}\|\boldsymbol{\lambda}^*\|_2 \leqslant 2\sqrt{M^*}\|f_{\boldsymbol{\lambda}^*}\|_{P_X,2} \approx 2\sqrt{M^*}\|f\|_{P_X,2}.$$

We can estimate $\|f\|^2_{P_X,2}$ consistently by $\frac{1}{n}\sum_{i=1}^n (Y_i^2 - \sigma^2)$. Based on these estimates, we suggest the following data-driven choice of R:

$$\widehat{R} = 4\Big[\frac{\widehat{M}^*}{n}\sum_{i=1}^n (Y_i^2 - \sigma^2)\Big]_+^{1/2},$$

where $x_+ = \max(x, 0)$ and $\widehat{M}^*$ a prior approximation of the sparsity index of the signal f.

***Remark* 6.** *The choice of β in (9) requires the knowledge of σ^2, which characterizes the magnitude of the noise. This value may not be available in practice. Then it is natural to consider β as a tuning parameter and to select it by a data-driven method, for example, by a suitably adapted version of cross-validation. This point deserves a special attention and is beyond the scope of the present paper.*

***Remark* 7.** *If the distribution P_X of the design is unknown, it is impossible to normalize the dictionary functions ϕ_j. In such a situation, i.e., when the functions ϕ_j do not necessarily satisfy $\|\phi_j\|_{P_X,2} = 1$, the claim of Proposition 1 continues to hold true with the modified residual term $\mathsf{R}(M,\tau) = 4\tau^2\mathrm{Tr}(\mathcal{M}) + \frac{\beta}{n+1}$, which can be bounded by $4\tau^2 M L_\phi^2 + \frac{\beta}{n+1}$. Thus, once again, choosing τ as in (9) makes the residual term $\mathsf{R}(M,\tau)$ negligible w.r.t. the main terms of the risk bound.*

***Remark* 8.** *Proposition 1 is in agreement with the main principles of the theory of compressive sampling and sparse recovery, cf., e.g., [14]. Indeed, if the tuning parameters are well-chosen, the prediction done by $\widehat{f}_n$ can be quite accurate even if the sample size is relatively small with respect to the dimension M. This happens if the signal admits a M^*-sparse representation in a possibly overcomplete dictionary of cardinality M. Then the number of observations sufficient for an accurate prediction is of order M^* up to a logarithmic factor. Proposition 1 is also in perfect agreement with the principle of incoherent sampling (see, for instance, [14], page 10). In fact, in our setting, the incoherence of the sampling is ensured by the fact that $\phi_j \in L^2(\mathcal{X}, P_X)$ satisfy $\|\phi_j\|_{P_X,2} = 1$.*

Before closing this section, let us mention the recent work [26], where some interesting results on the aggregation of estimators in sparse regression are obtained.

5.2. Linear regression with random design

Consider now the case of linear regression. Assume that the i.i.d. observations $(\boldsymbol{X}_i, Y_i)$, $i = 1, \dots, n$, are drawn from the linear model

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\lambda}^* + \xi_i, \qquad i = 1, \dots, n, \tag{10}$$

where $\boldsymbol{X}_i \in \mathbb{R}^M$ are i.i.d. covariates and $\boldsymbol{\lambda}^* \in \mathbb{R}^M$ is the parameter of interest. Then our method reduces to estimating $\boldsymbol{\lambda}^*$ by

$$\widehat{\boldsymbol{\lambda}}_n = \frac{1}{n+1}\sum_{m=0}^{n+1}\int_{\mathbb{R}^M} \boldsymbol{\lambda}\,\widehat{\theta}_{m,\boldsymbol{\lambda}}\,\pi(d\boldsymbol{\lambda}),$$

where π is the sparsity prior and

$$\widehat{\theta}_{m,\boldsymbol{\lambda}} = \frac{\exp\big\{-\beta^{-1}\sum_{i=1}^m (Y_i - \boldsymbol{X}_i^\top\boldsymbol{\lambda})^2\big\}}{\int_{\mathbb{R}^M}\exp\big\{-\beta^{-1}\sum_{i=1}^m (Y_i - \boldsymbol{X}_i^\top\boldsymbol{\omega})^2\big\}\pi(d\boldsymbol{\omega})}.$$

Then the following result holds.

Proposition 2. *Consider the linear model (10) satisfying the above assumptions. Let the support of the probability distribution of* $\boldsymbol{X}_1$ *be included in* $[-1,1]^M$ *and* $\mathbf{E}[e^{t\xi_1}|\boldsymbol{X}_1] \leqslant e^{\sigma^2t^2/2}$ *for all* $t \in \mathbb{R}$. *Set* $\Sigma_X = \mathbf{E}[\boldsymbol{X}_1\boldsymbol{X}_1^\top]$. *Then for any* $\beta \geqslant 2\sigma^2 + 2(R + \|\boldsymbol{\lambda}^*\|_1)^2$ *and any* $\boldsymbol{\lambda}^*$ *such that* $\|\boldsymbol{\lambda}^*\|_1 \leqslant R - 2M\tau$ *we have*

$$\mathbf{E}[\|\Sigma_X^{1/2}(\widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}^*)\|_2^2] \leqslant \frac{\beta}{n+1}\Big(1 + 4\sum_{j=1}^M \log(1 + \tau^{-1}|\lambda_j^*|)\Big)\Bigg\} + 4\tau^2\mathrm{Tr}(\Sigma_X). \tag{11}$$

This proposition follows directly from Proposition 1 by setting $\phi_j(\boldsymbol{x}) = x_j$ if $|x_j| \leqslant 1$ and $\phi_j(\boldsymbol{x}) = 0$ if $|x_j| > 1$, where $\boldsymbol{x} \in \mathbb{R}^M$ and x_j is its jth coordinate. Note also that here we have $\mathcal{M} = \Sigma_X$.

5.3. Rate optimality

In this section, we discuss the optimality of the rates of aggregation obtained in Proposition 1. We show that the MA aggregate with the sparsity prior attains, up to a logarithmic factor, the optimal rates of aggregation (cf. [47]). Furthermore, $\widehat{f}_n$ is adaptive in the sense that it simultaneously achieves the optimal rates for the Model Selection (MS), Convex (C) and Linear (L) aggregation. In what follows, these rates are denoted respectively by $\psi_n^{\mathbf{MS}}(M)$, $\psi_n^{\mathbf{C}}(M)$ and $\psi_n^{\mathbf{L}}(M)$. It is established in [47] that:

$$\begin{aligned}\psi_n^{\mathbf{MS}}(M) &= n^{-1}\log M,\\ \psi_n^{\mathbf{C}}(M) &= n^{-1}(M \wedge \sqrt{n})\log(1 + Mn^{-1/2}),\\ \psi_n^{\mathbf{L}}(M) &= n^{-1}M.\end{aligned}$$

We wish to compare the risk of the estimator $\widehat{f}_n$ with the sparsity prior π to the smallest error $\|f_{\boldsymbol{\lambda}^*} - f\|_{P_X,2}^2$ where $\boldsymbol{\lambda}^*$ is one of $\boldsymbol{\lambda}^{\mathbf{MS}}$, $\boldsymbol{\lambda}^{\mathbf{C}}$ or $\boldsymbol{\lambda}^{\mathbf{L}}$ such that

$$\begin{aligned}\boldsymbol{\lambda}^{\mathbf{MS}} &= \arg\min_{\|\boldsymbol{\lambda}\|_0 = \|\boldsymbol{\lambda}\|_1 = 1}\|f_{\boldsymbol{\lambda}} - f\|_{P_X,2}^2,\\ \boldsymbol{\lambda}^{\mathbf{C}} &= \arg\min_{\|\boldsymbol{\lambda}\|_1 \leqslant 1}\|f_{\boldsymbol{\lambda}} - f\|_{P_X,2}^2,\\ \boldsymbol{\lambda}^{\mathbf{L}} &= \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^M}\|f_{\boldsymbol{\lambda}} - f\|_{P_X,2}^2.\end{aligned}$$

In the next proposition we denote by c constants which do not depend on M and n.

Proposition 3. *Assume that $\widehat{f}_n$ satisfies (8) with some $\beta > 0$ independent of M and n, and that $\log(M) \leqslant c_0 n$ for some constant c_0. If $R > 4$ and τ satisfies (5) with $\mathrm{Tr}(\mathcal{M}) = M$, then*

$$\mathbf{E}_f[\|\widehat{f}_n - f\|^2_{P_X,2}] \leqslant \|f_{\boldsymbol{\lambda}^{\mathbf{MS}}} - f\|^2_{P_X,2} + c\psi_n^{\mathbf{MS}}(M)\log(1+nM)$$

and

$$\mathbf{E}_f[\|\widehat{f}_n - f\|^2_{P_X,2}] \leqslant \|f_{\boldsymbol{\lambda}^{\mathbf{C}}} - f\|^2_{P_X,2} + c\psi_n^{\mathbf{C}}(M)\log(1+nM).$$

Finally, if $\|\boldsymbol{\lambda}^{\mathbf{L}}\|_1 \leqslant R - 2M\tau$, then

$$\mathbf{E}_f[\|\widehat{f}_n - f\|^2_{P_X,2}] \leqslant \|f_{\boldsymbol{\lambda}^{\mathbf{L}}} - f\|^2_{P_X,2} + c\psi_n^{\mathbf{L}}(M)\log(1+nM).$$

Proof. For model selection and linear aggregation the result follows immediately from (8) by putting there $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^{\mathbf{MS}}$ or $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^{\mathbf{L}}$ and using that $\|\boldsymbol{\lambda}^{\mathbf{MS}}\|_0 = \|\boldsymbol{\lambda}^{\mathbf{MS}}\|_1 = 1$. The case of convex aggregation with $M \leqslant \sqrt{n}$ follows from the bound for the linear aggregation. The case $M > \sqrt{n}$ requires some additional arguments, which are presented below.

Let $s = s_n$ be the integer part of $\sqrt{n}$, denoted by $[\sqrt{n}]$. We assume that $\boldsymbol{\lambda}^{\mathbf{C}}$ has at least s_n non-zero coordinates, the case $\|\boldsymbol{\lambda}^{\mathbf{C}}\|_0 < [\sqrt{n}]$ being a trivial consequence of (8). Using the Maurey randomization argument as in [7, 43], one can show that

$$\min_{\substack{\|\boldsymbol{\lambda}\|_1 \leqslant 1 \\ \|\boldsymbol{\lambda}\|_0 \leqslant s}} \|f_{\boldsymbol{\lambda}} - f\|^2_{P_X,2} \leqslant \|f_{\boldsymbol{\lambda}^{\mathbf{C}}} - f\|^2_{P_X,2} + \frac{\|\boldsymbol{\lambda}^{\mathbf{C}}\|_1^2}{\min(s, \|\boldsymbol{\lambda}^{\mathbf{C}}\|_0)} \leqslant \|f_{\boldsymbol{\lambda}^{\mathbf{C}}} - f\|^2_{P_X,2} + \frac{1}{s}. \tag{12}$$

Let $\boldsymbol{\lambda}^{s,\mathbf{C}}$ be a point where the minimum on the left hand side of (12) is attained. Since $\boldsymbol{\lambda}^{s,\mathbf{C}}$ has not more than s nonzero coordinates and $\|\boldsymbol{\lambda}^{s,\mathbf{C}}\|_1 \leqslant 1$, we have $\sum_j \log(1 + |\lambda_j^{s,\mathbf{C}}/\tau|) \leqslant s\log(1 + \|\boldsymbol{\lambda}^{s,\mathbf{C}}\|_1/\tau) \leqslant s\log(1+\tau^{-1})$. Thus, applying (8) to $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^{s,\mathbf{C}}$ and using (5), we get

$$\mathbf{E}_f[\|\widehat{f}_n - f\|^2_{P_X,2}] \leqslant \|f_{\boldsymbol{\lambda}^{s,\mathbf{C}}} - f\|^2_{P_X,2} + \frac{cs\log(1+\tau^{-1})}{n}, \tag{13}$$

where c is some constant independent of n and M. Recall now that $\|f_{\boldsymbol{\lambda}^{s,\mathbf{C}}} - f\|^2_{P_X,2}$ is equal to the left hand side of (12). This implies

$$\mathbf{E}_f[\|\widehat{f}_n - f\|^2_{P_X,2}] \leqslant \|f_{\boldsymbol{\lambda}^{\mathbf{C}}} - f\|^2_{P_X,2} + \frac{1}{s} + \frac{cs\log(1+\tau^{-1})}{n},$$

which leads to the desired result due to the choice $s = [\sqrt{n}]$ and (5). □

Remark 9. *The theory developed here relies on the fact that the risk is measured by the expected squared loss. In the case of general L_p-loss with $p \geqslant 1$, a universal procedure for aggregation is proposed in [27] and it is proved that the aggregation in L_p for $p > 2$ is more difficult than it is in L_2.*

6. Application to density estimation

Let $X_1, \dots, X_n$ be the observations, which are independent copies of a random variable $X : \Omega \to \mathcal{X}$ whose distribution has a density f with respect to some reference measure μ. We consider the problem of estimating f based on $X_1, \dots, X_n$. We measure the risk of an estimator $\tilde{f}$ of f by the integrated squared error

$$\ell(\tilde{f}, f) = \|\tilde{f} - f\|_{\mu,2}^2 = \int_{\mathcal{X}} \big(\tilde{f}(x) - f(x)\big)^2 \mu(dx).$$

Define the mapping $Q(\cdot, g) : \mathcal{X} \times L^2(\mathcal{X}, \mu) \to \mathbb{R}$ by

$$Q(x, g) = \|g\|_{\mu,2}^2 - 2g(x).$$

It is straightforward that $\mathbf{E}_f Q(X, g) - \ell(g, f) = -\|f\|_{\mu,2}^2$ and, therefore, Assumption Q1 is fulfilled. To further specify the setting, we consider the family $\mathcal{F}_\Lambda$ defined in (6) where the functions ϕ_j are chosen from $L^2(\mathcal{X}, \mu)$ so that $\|\phi_j\|_{\mu,2} = 1$ and $\|\phi_j\|_{\mu,\infty} \leqslant L$, $j = 1, \dots, M$, for some positive constant L. Note that the functions ϕ_j need not be integrable or positive. We have the following result.

Proposition 4. *Let the assumptions given above in this subsection be satisfied and $\|f\|_{\mu,\infty} \leqslant L$. If β is such that*

$$(\beta - 2R^2) e^{-4R(L+\sqrt{L})/\beta} \geqslant 2L + 4RL, \tag{14}$$

then the MA aggregate $\widehat{f}_n$ based on the sparsity prior (3) satisfies

$$\mathbf{E}_f[\|\widehat{f}_n - f\|_{\mu,2}^2] \leqslant \inf_{\boldsymbol{\lambda}^*} \left\{ \|f_{\boldsymbol{\lambda}^*} - f\|_{\mu,2}^2 + \frac{4\beta}{n+1} \sum_{j=1}^{M} \log(1 + \tau^{-1}|\lambda_j^*|) \right\} + \mathsf{R}(M, \tau) \tag{15}$$

where the inf *is taken over all the vectors $\boldsymbol{\lambda}^*$ such that $\|\boldsymbol{\lambda}^*\|_1 \leqslant R - 2M\tau$.*

The proof of this proposition is given in the appendix. It consists in checking that Assumptions Q2 and L are satisfied and then applying Theorem 2. Condition (14) can be significantly simplified in many concrete situations. For example, if we assume that $R = 1$ or $R = 2$, then one can choose $\beta = 12L$ and $\beta = 23L$ respectively, provided that $L \geqslant 2$.

7. Classification

Assume that we have a sample $(X_1, Y_1), \dots, (X_n, Y_n)$, where $X_i \in \mathcal{X}$ and $Y_i \in \{-1, +1\}$ are labels. Here $\mathcal{X}$ is an arbitrary measurable space and (X_i, Y_i) are assumed to be generated independently according to a probability distribution P. The goal of binary classification is to assign a label $+1$ or -1 to a new random point x which is distributed as X_i and independent of $X_1, \dots, X_n$.

The problem of interest in classification is to design a classifier $\widehat{f} : \mathcal{X} \to \mathbb{R}$ having a small misclassification risk $R[\widehat{f}] = \int_{\mathcal{X}\times\{-1,+1\}} \mathbb{1}(\mathrm{sgn}(\widehat{f}(x)) \neq y)\, P(dx, dy)$. Denote by $\eta : \mathcal{X} \to [-1, 1]$ the regression function

$$\eta(x) = \mathbf{E}(Y_1|X_1 = x) = 2\mathbf{P}(Y_1 = 1|X_1 = x) - 1, \qquad \forall\, x \in \mathcal{X}.$$

The Bayes classifier is defined as follows: $f(x) = \mathbb{1}(\eta(x) > 0) - \mathbb{1}(\eta(x) \leqslant 0) = \mathrm{sgn}(\eta(x))$. One easily checks that

$$R[\widehat{f}] - R[f] = \int_{\mathcal{X}} \mathbb{1}(\mathrm{sgn}(\widehat{f}(x)) \neq f(x))|\eta(x)|\, P_X(dx),$$

where P_X is the distribution of X_1. This shows that the Bayes classifier f minimizes the misclassification risk. Clearly, the Bayes classifier is not available in practice because of its dependence on the unknown regression function $\eta(\cdot)$.

This problem is a special case of the general setting of Section 2 if we take there $Z_i = (X_i, Y_i)$ and $\ell(g, f) = R[g] - R[f]$. Assumption Q1 is then fulfilled with $Q(z, g) = \mathbb{1}\big(\mathrm{sgn}(g(x)) = y\big)$ where $z = (x, y)$. However, Assumptions Q2 and L are not satisfied.

7.1. Classification under smooth Φ-losses

An alternative approach is to consider the Φ-risk of classifiers. For a fixed convex twice differentiable function $\Phi : \mathbb{R} \to \mathbb{R}_+$, the Φ-risk of a classifier $\widehat{f}$ is defined by

$$R_\Phi[\widehat{f}] = \int_{\mathcal{X}\times\{\pm1\}} \Phi\big(-y\widehat{f}(x)\big)\, P(dx, dy) = \frac{1}{2}\int_{\mathcal{X}} \Big\{\Phi\big(-\widehat{f}(x)\big)\big(1+\eta(x)\big)+\Phi\big(\widehat{f}(x)\big)\big(1-\eta(x)\big)\Big\}\, P_X(dx).$$

In this subsection, we are mainly interested in the four common choices of Φ presented in the top lines of Table 1. For these and other loss functions, sharp relations between the Φ-risk and the misclassification risk of a given classifier $\widehat{f}$ have been established in [55], [5]. In particular, it is proved in these papers that the minimum of Φ-risk is attained at any classifier satisfying

$$f_\Phi(x) \in \arg\min_{u\in\mathbb{R}} \big\{\Phi(-u)\big(1 + \eta(x)\big) + \Phi(u)\big(1 - \eta(x)\big)\big\}, \quad \forall x \in \mathcal{X}.$$

Note however that in practice the computation of f_Φ is impossible because of its dependence on the unknown η.

Our aim here is to design a classifier having a Φ-risk which is nearly as small as the minimal possible Φ-risk. This task can be recast in a problem of estimation where f_Φ is the function to be estimated and the quality of an estimator (classifier) $\widehat{f}$ is measured by the excess risk $R_\Phi[\widehat{f}] - R_\Phi[f_\Phi]$. Therefore, this is a particular case of the setting described in Section 2 with $\ell(g, f) = \ell_\Phi(g, f_\Phi) = R_\Phi[g] - R_\Phi[f_\Phi]$ and $Q(z, g) = \Phi(-yg(x))$ for every $z = (x, y)$. Here Assumption Q1 is obviously satisfied.

In the same spirit as in the previous sections, we assume that we are given a dictionary $\{\phi_j\}_{j=1,\ldots,M}$ of functions on $\mathcal{X}$ with values in $\mathbb{R}$. The family $\mathcal{F}_\Lambda$ is defined as the set of all linear combinations of the functions ϕ_j with coefficients $\lambda_1, \ldots, \lambda_M$, such that the vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)$ belongs to the ℓ_1 ball with radius R, cf. (6). The next proposition shows that a strong sparsity oracle inequality holds for an appropriate choice of β.

Loss	$\boldsymbol{\Phi(u)}$	$\boldsymbol{f_\Phi(x)}$	$\boldsymbol{Q(z,g)}$	$\boldsymbol{\beta_\Phi}$	$\boldsymbol{C_\Phi}$
Squared	$(1+u)^2$	$\eta(x)$	$(1-yg(x))^2$	$2(1+RL_\Phi)^2$	8
Truncated Squared	$\{(1+u)_+\}^2$	$\eta(x)$	$\{\max(1-yg(x),0)\}^2$	$2(1+RL_\Phi)^2$	8
Boosting	e^u	$\frac{1}{2}\log\frac{1+\eta(x)}{1-\eta(x)}$	$e^{-yg(x)}$	e^{RL_Φ}	$4e^{RL_\Phi}$
Logit-Boosting	$\log(1+e^u)$	$\log\frac{1+\eta(x)}{1-\eta(x)}$	$\log(1+e^{-yg(x)})$	e^{RL_Φ}	4
Misclassification	$\mathbb{1}(u=1)$	$\eta(x)$	$\mathbb{1}(g(x)\neq y)$	–	–
Hinge	$(1-u)_+$	$\eta(x)$	$\max(1-yg(x),0)$	–	–

Table 1. Common choices of function Φ; classifiers f_Φ minimizing the Φ-risk; the corresponding functions Q; constants β_Φ and C_Φ appearing in Proposition 5.

Proposition 5. *Assume that for some constant $L_\phi > 0$ we have $\max_{j=1,\dots,M}\|\phi_j\|_{P_X,\infty}\leqslant L_\phi$. Let the function Φ be twice continuously differentiable with*[1]

$$\beta_\Phi := \sup_{|u|\leqslant RL_\phi}\frac{\Phi'(u)^2}{\Phi''(u)} < \infty.$$

Then the MA aggregate defined with $\beta\geqslant\beta_\Phi$ and with the sparsity prior (3) satisfies

$$\mathbf{E}_f[\ell_\Phi(\widehat{f}_n,f)] \leqslant \min_{\|\boldsymbol{\lambda}^*\|_1\leqslant R-2M\tau}\bigg(\ell_\Phi(f_{\boldsymbol{\lambda}^*},f)+\frac{4\beta}{n+1}\sum_{j=1}^M\log(1+\tau^{-1}|\lambda_j^*|)\bigg)+C_\Phi\tau^2\sum_{j=1}^M\|\phi_j\|^2_{P_X,2}+\frac{\beta}{n+1}, \tag{16}$$

where $C_\Phi = 4\max_{|u|\leqslant RL_\phi}\Phi''(u)$.

Proof. We apply Theorems 1 and 2. First, we show that Assumption Q2 is satisfied. Recall that $Q(z,g)=\Phi(-yg(x))$ and set $\Psi_\beta(g,\tilde g)=\int_{\mathcal{X}\times\{\pm1\}}\exp\big(-\beta^{-1}\{Q(z,g)-Q(z,\tilde g)\}\big)\,P(dx,dy)$. Let us show that for $\beta\geqslant\beta_\Phi$ the mapping $g\mapsto\Psi_\beta(g,\tilde g)$ is concave. By standard arguments, this reduces to proving that the function $t\mapsto\phi(t)=\Psi_\beta(tg+(1-t)\bar g,\tilde g)$ is concave on $t\in[0,1]$ for every fixed $g,\bar g$ and $\tilde g$. A simple algebra shows that the second derivative of ϕ is non-positive on $[0,1]$ whenever $\beta\geqslant\Phi'(-yg(x))^2/\Phi''(-yg(x))$ for all $(x,y)\in\mathcal{X}\times\{\pm1\}$ and all $g\in\mathcal{F}_\Lambda$. On this set of x,y,g the value $-yg(x)$ belongs to the interval $[-RL_\phi,RL_\phi]$. Thus, Assumption Q2 is satisfied for $\beta\geqslant\beta_\Phi$ and Theorem 1 can be applied.

To use Theorem 2, it remains to prove that Assumption L is satisfied with $\mathcal{M}$ being the matrix with entries $\big(\frac14 C_\Phi\langle\phi_j,\phi_{j'}\rangle\big)$, where j and j' run over $\{1,\dots,M\}$. From the formula for $R_\Phi[\widehat{f}]$ given at the beginning of this subsection we get

$$\nabla^2 L_f(\boldsymbol{\lambda})=\nabla^2 R_\Phi[f_{\boldsymbol{\lambda}}]=\int_{\mathcal{X}\times\{\pm1\}}\big(\nabla f_{\boldsymbol{\lambda}}(x)\cdot\nabla f_{\boldsymbol{\lambda}}(x)^\top\big)\Phi''\big(-yf_{\boldsymbol{\lambda}}(x)\big)\,P(dx,dy).$$

Since $yf_{\boldsymbol{\lambda}}(x)\in[-RL_\phi,RL_\phi]$ the matrix $\mathcal{M}-\nabla^2L_f(\boldsymbol{\lambda})$, where $\mathcal{M}=\frac14C_\Phi\int(\nabla f_{\boldsymbol{\lambda}}\nabla f_{\boldsymbol{\lambda}}^\top)(x)\,P_X(dx)$, is positive semi-definite. The desired result follows now from the linearity in $\boldsymbol{\lambda}$ of $f_{\boldsymbol{\lambda}}(x)$. $\square$

For the four common choices of Φ presented in the top lines of Table 1 all the conditions of Proposition 5 are satisfied for a properly chosen constant β. The minimal values of β, as well

[1] We use here the convention $0/0=0$.

as the values of the constant C_Φ, for each loss function Φ are reported in the last two columns of Table 1. It is often interesting to use binary classifiers ϕ_j (*i.e.*, functions with values in $\{\pm 1\}$), in which case $L_\phi = 1$. Also note that the expressions for β_Φ suggest to choose R not too large, especially in the case of the boosting and the logit-boosting losses.

7.2. Classification under the hinge loss

One of the key issues in machine learning is classification by support vector machines. They correspond to a penalized Φ-risk classification with the loss $\Phi(u) = \Phi_H(u) = \max(1+u, 0)$, referred to as the hinge loss. A notable feature of the hinge loss is that the classifier $f_{\Phi_H}(x)$ equals $\mathrm{sgn}(\eta(x))$ and therefore coincides with the Bayes classifier for the misclassification risk. However, since the hinge loss does not satisfy Assumptions Q2 and L, Proposition 5 cannot be applied. Furthermore, as shown in [36], no aggregation procedure can attain the fast rate of aggregation (*i.e.*, the rate $1/n$ up to a logarithmic factor) when the risk is measured by the hinge loss.

The reason for the failure of Assumption L is that the hinge loss is not continuously differentiable. One can circumvent this problem by using the smoothing argument of Remark 3. Indeed, let us fix $\varepsilon > 0$ and introduce the function $K_\varepsilon(z) = (\sqrt{\varepsilon^2 + z^2} - \varepsilon)\mathbb{1}(z > 0)$, which is a smooth approximation to the positive part of z. It is easy to see that $K_\varepsilon(z) \leqslant \max(z, 0) \leqslant K_\varepsilon(z) + \varepsilon$ and that $K''_\varepsilon(z) = \varepsilon^2(\varepsilon^2 + z^2)^{-3/2} \in (0, \varepsilon^{-1}]$ for $z > 0$. This allows us to approximate the loss $\ell_{\Phi_H}(g, f)$ by

$$\ell_\varepsilon(g, f) = \frac{1}{2}\int_{\mathcal{X}} \big\{K_\varepsilon(1 - g(x))(1 + \eta(x)) + K_\varepsilon(1 + g(x))(1 - \eta(x))\big\}\, P_X(dx) - R_{\Phi_H}[f].$$

Although Assumption Q2 is not fulfilled, the next proposition shows that it is possible to adapt the argument of Proposition 5 to the hinge loss $\Phi = \Phi_H$. However, unlike Proposition 5 where the rate of convergence is of the order $1/n$ (up to a logarithmic factor), the resulting sparsity oracle inequality has only the rate $1/\sqrt{n}$ (up to a logarithmic factor), cf. also Remark 10 (1) below. This is the best we can get for the hinge loss without imposing any condition on η.

Proposition 6. *Let $\Phi_H(u) = \max(1+u, 0)$ be the hinge loss and $\max_{j=1,\dots,M} \|\phi_j\|_{P_X,\infty} \leqslant L_\phi$ for some $L_\phi > 0$. Then, for every $\beta > 0$ the MA aggregate $\widehat{f}_n$ based on the prior given by (3) satisfies*

$$\mathbf{E}_f[\ell_{\Phi_H}(\widehat{f}_n, f)] \leqslant \min_{\|\boldsymbol{\lambda}^*\|_1 \leqslant R - 2M\tau} \Big(\ell_{\Phi_H}(f_{\boldsymbol{\lambda}^*}, f) + \frac{4\beta}{n+1}\sum_{j=1}^M \log(1 + \tau^{-1}|\lambda^*_j|)\Big) + \frac{2(1 + RL_\phi)^2}{\beta} e^{\frac{1+RL_\phi}{\beta}} + \tilde{\mathsf{R}}(M, \tau),$$

where $\tilde{\mathsf{R}}(M, \tau) = 4\tau L_\phi\sqrt{M} + \beta(n+1)^{-1}$.

The proof of this proposition is given in the appendix.

***Remark* 10.**

1. *Consider the sparsity scenario, i.e., assume that for some vector $\boldsymbol{\lambda}^*$ having at most M^* non-zero coordinates, the excess risk $\ell_{\Phi_H}(f_{\boldsymbol{\lambda}^*}, f)$ is small and $\|\boldsymbol{\lambda}^*\|_1 \leqslant R/2$. Proposition 6 with the choice of $\beta = (1 + RL_\phi)\sqrt{n/M^*}$ and $\tau = \min\big(\frac{1}{\sqrt{nM}}, \frac{R}{4M}\big)$ leads to the sparsity oracle inequality*

$$\mathbf{E}_f[\ell_{\Phi_H}(\widehat{f}_n, f)] \leqslant \min_{\substack{\|\boldsymbol{\lambda}^*\|_1 \leqslant R/2 \\ \|\boldsymbol{\lambda}^*\|_0 \leqslant M^*}} \bigg(\ell_{\Phi_H}(f_{\boldsymbol{\lambda}^*}, f) + \frac{(1 + RL_\phi)\sqrt{M^*}}{\sqrt{n}} \big\{ C + 4\log(1 + \tau^{-1}\|\boldsymbol{\lambda}^*\|_1) \big\} \bigg),$$

   where $C > 0$ is a constant independent on M, M^* and n if $M^* \leqslant n$. This result is valid for arbitrary η. It should be noted that the MA aggregate $\widehat{f}_n$ satisfying this SOI depends on the upper bound M^* on the sparsity level, which is not always available in practice. Constructing a classifier independent of M^* and satisfying the above SOI is an interesting open problem.
2. *An important special case is a dictionary composed from a large number of simple binary classifiers $\phi_j : \mathcal{X} \to \{\pm 1\}$. If we choose $R = 1$, all aggregates $f_{\boldsymbol{\lambda}}$ with $\|\boldsymbol{\lambda}\|_1 \leqslant R$, as well as their mixtures, take values in $[-1, 1]$, and therefore the function $Q(z, f_{\boldsymbol{\lambda}})$ associated with the hinge loss is linear in $\boldsymbol{\lambda}$. This property has two important consequences. The first one is that Assumption L holds with $\mathcal{M} = 0$ and it is no longer necessary to smooth out the function $L_f(\boldsymbol{\lambda})$ and to use Remark 3 in the proof Proposition 6. Thus, the residual term $\tilde{\mathsf{R}}$ is equal to $\beta(n+1)^{-1}$. The second consequence is computational, related to the Langevin Monte-Carlo approximation of the MA aggregate briefly described in Section 8.2 below. Namely, in this case we have strong mixing properties that are independent of the ambient dimension M, due to the independence of the coordinates of the Langevin diffusion.*
3. *According to [36], if the underlying distribution P satisfies the margin assumption of [48], then the rate of aggregation can be substantially improved. It would be interesting to investigate whether this property extends to the sparsity scenario. It is likely that one of the randomized procedures of [3] used in conjunction with our sparsity prior can yield an aggregation rate optimal classifier.*

8. Discussion

8.1. Comparison with other methods of sparse estimation

In this paper we have proved sparsity oracle inequalities (SOI) in a setting, which is important but not much studied in the literature on sparsity. We considered the i.i.d. random sampling and we measured the quality of estimation/prediction by the average loss with respect to the distribution of $Z = (X, Y)$, namely, our main example was the loss $\ell(g, f) = \int_{\mathcal{Z}} Q(z, g)\, P_f(dz)$. Most of the literature on sparse estimation is focused on the high-dimensional linear regression model with fixed design, so the data are not i.i.d. and the empirical prediction loss, rather than the average loss is considered. Notable exceptions are the papers [13, 33, 34, 35, 49] where the framework is similar to ours. Among these, [35] focuses on regression with random design and study the Dantzig selector, while [13, 33, 34, 49] analyze the penalized estimators of the form

$$\widehat{\boldsymbol{\lambda}}_n = \arg\min_{\boldsymbol{\lambda}\in\Lambda}\left(\frac{1}{n}\sum_{i=1}^n Q(Z_i, f_{\boldsymbol{\lambda}}) + \mathrm{Pen}(\boldsymbol{\lambda})\right)$$

where $\mathrm{Pen}(\boldsymbol{\lambda})$ is a penalty, which is equal or close to the ℓ_1-penalty $r\|\boldsymbol{\lambda}\|_1$ with a suitable regularization parameter $r > 0$. For the penalized estimator $\tilde{f}_n = f_{\widehat{\boldsymbol{\lambda}}_n}$ they prove SOI of the form (here we give a "generic" simplified version based on [33]):

$$\ell(\tilde{f}_n, f) \leqslant \min_{\substack{\|\boldsymbol{\lambda}^*\|_1\leqslant R\\ \|\boldsymbol{\lambda}^*\|_0\leqslant M^*}}\left(3\ell(f_{\boldsymbol{\lambda}^*}, f) + \frac{C(1+R^2)M^*}{n\kappa_{n,M}}\mathcal{L}_{n,M}\right) \tag{17}$$

with a probability close to 1, where $C > 0$ is a constant independent of n and M, $\mathcal{L}_{n,M}$ is a factor, which is logarithmic in n and M, and $\kappa_{n,M}$ is minimal sparse eigenvalue appearing in the conditions on the Gram matrix of the dictionary quoted in the Introduction. With the same notation, a "generic" version of our SOI for the MA aggregate $\widehat{f}_n$ is the following:

$$\mathbf{E}[\ell(\widehat{f}_n, f)] \leqslant \min_{\substack{\|\boldsymbol{\lambda}^*\|_1\leqslant R\\ \|\boldsymbol{\lambda}^*\|_0\leqslant M^*}}\left(\ell(f_{\boldsymbol{\lambda}^*}, f) + \frac{C(1+R^2)M^*}{n}\mathcal{L}_{n,M}\right). \tag{18}$$

There are two advantages of (18) with respect to (17). First, (18) is a sharp oracle inequality, since the leading constant is 1, whereas this is not the case in (17). Second and most important, (18) holds under mild assumptions on the dictionary, such as the boundedness of the functions ϕ_j in some norm, whereas (17) requires restrictive assumptions on minimal sparse eigenvalue $\kappa_{n,M}$ which can be very small and appears in the denominator. In particular, (18) is applicable when $\kappa_{n,M} = 0$. Finally, we note that (17) is an oracle inequality "in probability" while (18) is "in expectation". Inequalities in expectation can be derived from the inequalities in probability of the form (17) obtained in [13, 33, 34, 49] only under some additional assumptions. So, strictly speaking, even more assumptions should be imposed in the case of (17) to make possible direct comparison with (18).

In conclusion, we see that the oracle bounds for ℓ_1-penalized methods, such as the Lasso or its modifications can be quite inaccurate as compared to the those that we obtain for the MA aggregate.

The ℓ_0-penalized methods for models with i.i.d. data are less studied. To our knowledge, this is done only for regression with random design [10] and for density estimation [40]. The oracle inequalities in those papers are less accurate than the ours since the leading constant there is greater than 1. Moreover, if we want to make it closer to 1, the remainder term of the oracle inequalities explodes.

Furthermore, as mentioned above, our sparsity oracle inequalities are potentially applicable not only for the MA aggregate, but for any estimator associated to prior distribution π and satisfying a PAC-Bayesian bound in expectation as in Theorem 1.

8.2. Computational aspects

If the dimension M is large the computation of the MA aggregate with the sparsity prior becomes a hard problem. Indeed, its definition contains integrals over a simplex in $\mathbb{R}^M$.

Nevertheless, accurate approximations can be realized by a numerically efficient algorithm based on Langevin Monte-Carlo. This algorithm along with the convergence and simulation studies is discussed in [22, 23]. Here we only sketch some main ideas underlying the numerical procedure. For simplicity, we consider the case of linear regression (cf. Subsection 5.2). The argument is easily extended to other models discussed in the previous section.

Thus, assume that we have a sample $(\boldsymbol{X}_i, Y_i)$, $i = 1, \ldots, n$, and a finite dictionary $\{\phi_j : \mathcal{X} \to \mathbb{R}\}$ of cardinality M. We wish to compute the expression

$$\tilde{\boldsymbol{\lambda}} = \frac{\int_{\mathbb{R}^M} \boldsymbol{\lambda} e^{-\beta^{-1}\|\mathbf{Y}-F_{\boldsymbol{\lambda}}(\boldsymbol{X})\|_2^2} \pi(d\boldsymbol{\lambda})}{\int_{\mathbb{R}^M} e^{-\beta^{-1}\|\mathbf{Y}-F_{\boldsymbol{\lambda}}(\boldsymbol{X})\|_2^2} \pi(d\boldsymbol{\lambda})}, \tag{19}$$

where $F_{\boldsymbol{\lambda}}(\boldsymbol{X}) = (f_{\boldsymbol{\lambda}}(\boldsymbol{X}_1), \ldots, f_{\boldsymbol{\lambda}}(\boldsymbol{X}_n))^\top$ and $f_{\boldsymbol{\lambda}} = \sum_{j=1}^M \lambda_j \phi_j$. A slight modification of the sparsity prior consists in replacing π defined in (3) by

$$\tilde{\pi}(d\boldsymbol{\lambda}) \propto \left(\prod_{j=1}^M \frac{e^{-\varpi(\alpha\lambda_j)}}{(\tau^2 + \lambda_j^2)^2} \right) \mathbb{1}(\|\boldsymbol{\lambda}\|_1 \leqslant R)\, d\boldsymbol{\lambda}, \tag{20}$$

where α is a small parameter and $\varpi : \mathbb{R} \to \mathbb{R}$ is the Huber function: $\varpi(t) = t^2 \mathbb{1}(|t| \leqslant 1) + (2|t| - 1)\mathbb{1}(|t| > 1)$. Introducing the product of $e^{-\varpi(\alpha\lambda_j)}$ in the definition of the prior does not affect its capacity to capture sparse objects, in the sense that the MA aggregate based on the prior (20) can be shown to satisfy a SOI which is quite similar to that of Theorem 2 (cf. [22, 23] where the regression model with fixed design is treated). On the other hand, this modification of the sparsity prior makes it possible to rigorously prove the geometric ergodicity of the Langevin diffusion defined below.

Note that we can equivalently write $\tilde{\boldsymbol{\lambda}}$ in the form

$$\tilde{\boldsymbol{\lambda}} = \frac{\int_{\mathbb{R}^M} \boldsymbol{\lambda} \mathbb{1}(\|\boldsymbol{\lambda}\|_1 \leqslant R) p_V(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}}{\int_{\mathbb{R}^M} \mathbb{1}(\|\boldsymbol{\lambda}\|_1 \leqslant R) p_V(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}}, \tag{21}$$

where $p_V(\boldsymbol{\lambda}) \propto e^{V(\boldsymbol{\lambda})}$ with

$$V(\boldsymbol{\lambda}) = -\beta^{-1}\|\mathbf{Y} - F_{\boldsymbol{\lambda}}(\boldsymbol{X})\|_2^2 - \sum_{j=1}^M 2\Big\{ \log(\tau^2 + \lambda_j^2) + \varpi(\alpha\lambda_j) \Big\}. \tag{22}$$

Consider now the Langevin stochastic differential equation (SDE)

$$d\boldsymbol{L}_t = \nabla V(\boldsymbol{L}_t)\, dt + \sqrt{2}\, d\boldsymbol{W}_t, \qquad \boldsymbol{L}_0 = 0, \qquad t \geqslant 0$$

where $\boldsymbol{W}$ stands for an M-dimensional Brownian motion. For our choice of the potential V this SDE has a unique strong solution. It can be also shown (cf. [22, 23]) that this choice of V guarantees the geometric ergodicity of the solution, which implies that its stationary distribution has the density $p_V(\boldsymbol{\lambda}) \propto e^{V(\boldsymbol{\lambda})}$, $\boldsymbol{\lambda} \in \mathbb{R}^M$. This and (21) suggest the Langevin Monte Carlo procedure of computation of $\tilde{\boldsymbol{\lambda}}$. Indeed, consider the time averages

$$\bar{\boldsymbol{L}}_T = \frac{1}{T}\int_0^T \boldsymbol{L}_t \mathbb{1}(\|\boldsymbol{L}_t\|_1 \leqslant R)\, dt, \qquad S_T = \frac{1}{T}\int_0^T \mathbb{1}(\|\boldsymbol{L}_t\|_1 \leqslant R)\, dt, \qquad T \geqslant 0.$$

According to the above remarks, the ratio of these average values converges, as $T \to \infty$, to the vector $\tilde{\boldsymbol{\lambda}}$ that we want to compute. Note that $\bar{\boldsymbol{L}}_T$ and S_T are one-dimensional integrals over a finite interval and, therefore, are simpler objects than $\tilde{\boldsymbol{\lambda}}$, which is an integral in M dimensions. Still, one cannot compute $\bar{\boldsymbol{L}}_T$ directly, and some discretization is needed. A standard way of doing it is to approximate $\bar{\boldsymbol{L}}_T$ and S_T by the sums

$$\bar{\boldsymbol{L}}^E_{T,h} = \frac{1}{[T/h]} \sum_{k=0}^{[T/h]-1} \boldsymbol{L}^E_k \mathbb{1}(\|\boldsymbol{L}^E_k\|_1 \leqslant R), \qquad S^E_{T,h} = \frac{1}{[T/h]} \sum_{k=0}^{[T/h]-1} \mathbb{1}(\|\boldsymbol{L}^E_k\|_1 \leqslant R),$$

where $\{\boldsymbol{L}^E_k\}$ is the Markov chain defined by the Euler scheme

$$\boldsymbol{L}^E_{k+1} = \boldsymbol{L}^E_k + h\nabla V(\boldsymbol{L}^E_k) + \sqrt{2h}\,\boldsymbol{W}_k, \qquad \boldsymbol{L}^E_0 = 0, \qquad k = 0, 1, \dots, [T/h] - 1.$$

Here $\boldsymbol{W}_1$, $\boldsymbol{W}_2, \dots$ are i.i.d. standard Gaussian random vectors in $\mathbb{R}^M$, $h > 0$ is a step of discretization, and $[x]$ stands for the integer part of $x \in \mathbb{R}$. It can be shown that $\bar{\boldsymbol{L}}^E_{T,h}$ is an accurate approximation of $\bar{\boldsymbol{L}}_T$ for small h. We refer to [22, 23] for further details. The computational complexity is polynomial in M and n. Simulation results in [22, 23], as well as the experiments on image denoising [45], show the fast convergence of the algorithm; it can be easily realized in dimensions M up to several thousands. They also demonstrate nice performance of the exponentially weighted aggregate as compared with the Lasso and other related methods of prediction under the sparsity scenario.

9. Appendix

9.1. Proof of Theorem 1.

First, note that without loss of generality we can set $\beta = 1$. If this is not the case, it suffices to replace Q and ℓ by $\tilde{Q} = \frac{1}{\beta}Q$ and $\tilde{\ell} = \frac{1}{\beta}\ell$, respectively. By Assumption Q1,

$$\mathbf{E}_f[\ell(\widehat{f}_n, f)] = \int_{\mathcal{Z}} \mathbf{E}_f[Q(z, \widehat{f}_n)]\, P_f(dz) - \Delta(f). \tag{23}$$

In the last display we have used Fubini's theorem to interchange the integral and the expectation; this is possible since the integrand is bounded from below. To get the desired result, one needs now to bound the first term on the RHS of (23), which we rewrite as follows

$$\int_{\mathcal{Z}} \mathbf{E}_f[Q(z, \widehat{f}_n)]\, P_f(dz) = -\int_{\mathcal{Z}} \mathbf{E}_f\big[\log\big(\exp\big\{-Q(z, \widehat{f}_n)\big\}\big)\big] P_f(dz). \tag{24}$$

Recall now that $\widehat{f}_n$ is defined as the average of the functions $f_{\boldsymbol{\lambda}}$ w.r.t. the probability measure $\widehat{\mu}_n$. If we knew that the mapping $g \mapsto \exp\big\{-Q(z, g)\big\}$ is concave on the convex hull of $\mathcal{F}_\Lambda$, we could apply Jensen's inequality to get

$$\exp\big\{-Q(z, \widehat{f}_n)\big\} \geqslant \int_\Lambda \exp\big\{-Q(z, f_{\boldsymbol{\lambda}})\big\}\widehat{\mu}_n(d\boldsymbol{\lambda}).$$

As we see below, this would allow us to get inequality (2) by a simple application of the convex duality argument. Unfortunately, the above mentioned concavity property is rather exceptional and therefore the quantity

$$S_1(z,\mathbf{Z}) = \log\Big(\int_\Lambda \exp\big\{-Q(z,f_{\boldsymbol\lambda})\big\}\widehat\mu_n(d\boldsymbol\lambda)\Big) - \log\Big(\exp\big\{-Q(z,\widehat f_n)\big\}\Big)$$

is not necessarily a.s. negative. However, we may write

$$\int_{\mathcal Z} \mathbf{E}_f\big[\log\big(e^{-Q(z,\widehat f_n)}\big)\big]P_f(dz) = \int_{\mathcal Z}\mathbf{E}_f\big[S_0(z,\mathbf{Z}) - S_1(z,\mathbf{Z})\big]P_f(dz) \tag{25}$$

where

$$S_0(z,\mathbf{Z}) = \log\Big(\int_\Lambda \exp\big\{-Q(z,f_{\boldsymbol\lambda})\big\}\,\widehat\mu_n(d\boldsymbol\lambda)\Big).$$

By the concavity of the logarithm,

$$S_0(z,\mathbf{Z}) \geqslant \frac{1}{n+1}\sum_{m=0}^{n}\log\Big(\int_\Lambda e^{-Q(z,f_{\boldsymbol\lambda})}\widehat\theta_{m,\boldsymbol\lambda}\,\pi(d\boldsymbol\lambda)\Big).$$

Replacing $\widehat\theta_{m,\boldsymbol\lambda}$ by its explicit expression and taking the integral of both sides of the last display, we get on the RHS a telescoping sum. This leads to the inequality

$$\int_{\mathcal Z}\mathbf{E}_f\big[S_0(z,\mathbf{Z})\big]P_f(dz) \geqslant \frac{1}{n+1}\int_{\mathcal Z^{n+1}}\log\Big(\int_\Lambda e^{-\sum_{i=1}^{n+1}Q(z_i,f_{\boldsymbol\lambda})}\pi(d\boldsymbol\lambda)\Big)P_f^{(n+1)}(d\mathbf{z}).$$

By a convex duality argument (cf., e.g., [25], p.264, or [16], p.160), we get

$$\log\Big(\int_\Lambda e^{-\sum_{i=1}^{n+1}Q(z_i,f_{\boldsymbol\lambda})}\pi(d\boldsymbol\lambda)\Big) \geqslant -\sum_{i=1}^{n+1}\int_\Lambda Q(z_i,f_{\boldsymbol\lambda})\,p(d\boldsymbol\lambda) - \mathcal K(p,\pi),$$

for every $p\in\mathcal P_\Lambda$. Therefore, integrating w.r.t. $z_1,\dots,z_{n+1}$ and using the symmetry, we get

$$\begin{aligned}\int_{\mathcal Z}\mathbf{E}_f\big[S_0(z,\mathbf{Z})\big]P_f(dz) &\geqslant -\int_{\mathcal Z}\int_\Lambda Q(z,f_{\boldsymbol\lambda})\,p(d\boldsymbol\lambda)P_f(dz) - \frac{\mathcal K(p,\pi)}{n+1}\\ &= -\int_\Lambda \ell(f_{\boldsymbol\lambda},f)\,p(d\boldsymbol\lambda) - \Delta(f) - \frac{\mathcal K(p,\pi)}{n+1}.\end{aligned}$$

This and equations (23)-(25) imply

$$\mathbf{E}_f[\ell(\widehat f_n,f)] \leqslant \int_\Lambda \ell(f_{\boldsymbol\lambda},f)\,p(d\boldsymbol\lambda) + \frac{\mathcal K(p,\pi)}{n+1} + \int_{\mathcal Z}\mathbf{E}_f[S_1(z,\mathbf{Z})]\,P_f(dz). \tag{26}$$

Let us show that the last term on the RHS of (26) is non-positive. Rewrite $S_1(z,\mathbf{Z})$ in the form

$$S_1(z,\mathbf{Z}) = \log\int_\Lambda \exp\big(-\big\{Q(z,f_{\boldsymbol\lambda}) - Q(z,\widehat f_n)\big\}\big)\,\widehat\mu_n(d\boldsymbol\lambda).$$

By the Fubini theorem, the concavity of the logarithm and Assumption Q2, we get

$$\int_{\mathcal Z}\mathbf{E}_f[S_1(z,\mathbf{Z})]\,P_f(dz) \leqslant \mathbf{E}_f\Big[\log\int_\Lambda \Psi_1(f_{\boldsymbol\lambda},\widehat f_n)\,\widehat\mu_n(d\boldsymbol\lambda)\Big]$$

(recall that we set $\beta=1$). The concavity of the map $g\mapsto\Psi_1(g,\widehat f_n)$ and Jensen's inequality yield

$$\int_\Lambda \Psi_1(f_{\boldsymbol\lambda},\widehat f_n)\,\widehat\mu_n(d\boldsymbol\lambda) \leqslant \Psi_1\Big(\int_\Lambda f_{\boldsymbol\lambda}\,\widehat\mu_n(d\boldsymbol\lambda),\widehat f_n\Big) = \Psi_\beta(\widehat f_n,\widehat f_n) = 1,$$

and the desired result follows.

9.2. Some lemmas.

We now give some technical results needed in the proofs.

Lemma 1. *For every $M \in \mathbb{N}$ and every $s > M$, the following inequality holds:*

$$\frac{1}{(\pi/2)^M} \int_{\{u:\|u\|_1>s\}} \prod_{j=1}^{M} \frac{du_j}{(1+u_j^2)^2} \leqslant \frac{M}{(s-M)^2}.$$

Proof. Let $U_1, \ldots, U_M$ be iid random variables drawn from the scaled Student $t(3)$ distribution having as density the function $u \mapsto 2/\big[\pi(1+u^2)^2\big]$. One easily checks that $\mathbf{E}[U_1^2] = 1$. Furthermore, with this notation, we have

$$\frac{1}{(\pi/2)^M} \int_{\{u:\|u\|_1>s\}} \prod_{j=1}^{M} \frac{du_j}{(1+u_j^2)^2} = \mathbf{P}\Big(\sum_{j=1}^{M} |U_j| \geqslant s\Big).$$

In view of Chebyshev's inequality the last probability can be bounded as follows:

$$\mathbf{P}\Big(\sum_{j=1}^{M} |U_j| \geqslant s\Big) \leqslant \frac{M\mathbf{E}[U_1^2]}{(s - M\mathbf{E}[|U_1|])^2} \leqslant \frac{M}{(s-M)^2}$$

and the desired inequality follows. □

Lemma 2. *Let the assumptions of Theorem 2 be satisfied and let p_0 be the probability measure defined by (30). If $M \geqslant 2$ then $\int_\Lambda (\lambda_1 - \lambda_1^*)^2 p_0(d\boldsymbol{\lambda}) \leqslant 4\tau^2$.*

Proof. Using the change of variables $u = (\boldsymbol{\lambda} - \boldsymbol{\lambda}^*)/\tau$ we write

$$\int_\Lambda (\lambda_1 - \lambda_1^*)^2 p_0(d\boldsymbol{\lambda}) = C_M \tau^2 \int_{B_1(2M)} u_1^2 \Big(\prod_{j=1}^{M} (1+u_j^2)^{-2}\Big)\, du$$

with

$$C_M = \Big(\int_{B_1(2M)} \Big(\prod_{j=1}^{M} (1+u_j^2)^{-2}\Big)\, du\Big)^{-1} \tag{27}$$

where u_j are the components of u. Extending the integration from $B_1(2M)$ to $\mathbb{R}^M$ and using the inequality $\int_{\mathbb{R}} u_1^2(1+u_1^2)^{-2} du_1 \leqslant \pi$, we get

$$\int_\Lambda (\lambda_1 - \lambda_1^*)^2 p_0(d\boldsymbol{\lambda}) \leqslant C_M \tau^2 \pi \Big(\int_{\mathbb{R}} (1+t^2)^{-2}\, dt\Big)^{M-1} = 2C_M \tau^2 (\pi/2)^M,$$

where we used that the primitive of the function $(1+x^2)^{-2}$ is $\frac{1}{2}\arctan(x) + \frac{x}{2(1+x^2)}$. To bound C_M, we apply Lemma 1 which yields

$$C_M \leqslant (2/\pi)^M \big(1 - 1/M\big)^{-1} \leqslant 2(2/\pi)^M, \tag{28}$$

for $M \geqslant 2$. Combining these estimates we get $\int_\Lambda (\lambda_1 - \lambda_1^*)^2 p_0(d\boldsymbol{\lambda}) \leqslant 4\tau^2$ and the desired inequality follows. □

Lemma 3. *Let the assumptions of Theorem 2 be satisfied and let p_0 be the probability measure defined by (30). Then* $\mathcal{K}(p_0,\pi) \leqslant 4\sum_{j=1}^M \log(1+|\lambda_j^*|/\tau)+1$.

Proof. The definition of π, p_0 and of the Kullback-Leibler divergence imply that

$$\mathcal{K}(p_0,\pi) = \int_{B_1(2M\tau)} \log\left\{\tau^{3M} C_M C_{\tau,R}\prod_{j=1}^M \frac{(\tau^2+\lambda_j^2)^2}{(\tau^2+(\lambda_j-\lambda_j^*)^2)^2}\right\} p_0(d\boldsymbol{\lambda})$$
$$= \log(\tau^{3M}C_M C_{\tau,R}) + 2\sum_{j=1}^M \int_{B_1(2M\tau)} \log\left\{\frac{\tau^2+\lambda_j^2}{\tau^2+(\lambda_j-\lambda_j^*)^2}\right\} p_0(d\boldsymbol{\lambda}). \qquad (29)$$

We now successively evaluate the terms on the RHS of (29). First, in view of (3), we have

$$C_{\tau,R} = \tau^{-3M}\int_{B_1(R/\tau)} \prod_{j=1}^M \frac{1}{(1+u_j^2)^2}\,du_j \leqslant \tau^{-3M}\Big(\int_{\mathbb{R}} (1+u_j^2)^{-2}\,du_j\Big)^M = \tau^{-3M}(\pi/2)^M.$$

This and (28) imply $\log(C_M C_{\tau,R}) \leqslant \log 2 \leqslant 1$.

To evaluate the second term on the RHS of (29) we use that

$$\frac{\tau^2+\lambda_j^2}{\tau^2+(\lambda_j-\lambda_j^*)^2} = 1 + \frac{2\tau(\lambda_j-\lambda_j^*)}{\tau^2+(\lambda_j-\lambda_j^*)^2}(\lambda_j^*/\tau) + \frac{\lambda_j^{*2}}{\tau^2+(\lambda_j-\lambda_j^*)^2}$$
$$\leqslant 1+|\lambda_j^*/\tau| + (\lambda_j^*/\tau)^2 \leqslant (1+|\lambda_j^*/\tau|)^2.$$

This entails that the second term on the RHS of (29) is bounded from above by $\sum_{j=1}^M 2\log(1+|\lambda_j^*|/\tau)$. Combining these inequalities we get the lemma. ☐

9.3. Proof of Theorem 2

In view of inequality (2), we have

$$\mathbf{E}_f[\ell(\widehat{f}_n,f)] \leqslant \int_\Lambda \ell(f_{\boldsymbol{\lambda}},f)\,p(d\boldsymbol{\lambda}) + \frac{\beta\mathcal{K}(p,\pi)}{n+1},$$

for every probability measure p. We choose here $p=p_0$ where p_0 has the following Lebesgue density:

$$\frac{dp_0}{d\boldsymbol{\lambda}}(\boldsymbol{\lambda}) \propto \frac{d\pi}{d\boldsymbol{\lambda}}(\boldsymbol{\lambda}-\boldsymbol{\lambda}^*)\mathbb{1}_{B_1(2M\tau)}(\boldsymbol{\lambda}-\boldsymbol{\lambda}^*). \qquad (30)$$

Here the sign $\propto$ indicates the proportionality of two functions. Since $\|\boldsymbol{\lambda}^*\|_1 \leqslant R-2M\tau$, the condition $\boldsymbol{\lambda}-\boldsymbol{\lambda}^* \in B_1(2M\tau)$ implies that $\boldsymbol{\lambda}\in B_1(R)$ and, therefore, p_0 is absolutely continuous w.r.t. the sparsity prior π. By Taylor's formula and Assumption L we have

$$\ell(f_{\boldsymbol{\lambda}},f) = L_f(\boldsymbol{\lambda}) \leqslant L_f(\boldsymbol{\lambda}^*) + \nabla L_f(\boldsymbol{\lambda}^*)^\top(\boldsymbol{\lambda}-\boldsymbol{\lambda}^*) + (\boldsymbol{\lambda}-\boldsymbol{\lambda}^*)^\top\mathcal{M}(\boldsymbol{\lambda}-\boldsymbol{\lambda}^*),\ \forall\boldsymbol{\lambda}\in\Lambda_0.$$

Integrating both sides of this inequality w.r.t. p_0 and using the fact that the density of π_0 is symmetric about $\boldsymbol{\lambda}^*$ and invariant under permutation of the components we find

$$\int_\Lambda \ell(f_{\boldsymbol{\lambda}},f)\,p_0(d\boldsymbol{\lambda}) \leqslant L_f(\boldsymbol{\lambda}^*) + \mathrm{Tr}(\mathcal{M})\int_\Lambda (\lambda_1-\lambda_1^*)^2 p_0(d\boldsymbol{\lambda}). \qquad (31)$$

Combining this inequality with those stated in Lemmas 2 and 3, we get the desired result.

9.4. Proof of Proposition 4

Note that Assumption Q1 obviously holds and Assumption L is fulfilled with $\mathcal{M}$ being the Gram matrix. The diagonal entries of $\mathcal{M}$ are equal to one since $\|\phi_j\|_{\mu,2} = 1$, and therefore we have $\mathrm{Tr}(\mathcal{M}) = M$.

It remains to check Assumption Q2 in order to apply Theorem 2. Introduce the function

$$\begin{aligned}\Xi(t) &= \exp\Big(- \beta^{-1}\big\{Q(X_1, g_0 + t(g_1 - g_0)) - Q(X_1, \tilde g)\big\}\Big)\\ &= \exp\Big[- \beta^{-1}\big\{\|g_t\|^2_{\mu,2} - \|\tilde g\|^2_{\mu,2} + 2\big(\tilde g(X_1) - g_t(X_1)\big)\big\}\Big], \qquad t \in [0,1]\end{aligned}$$

where g_0, g_1 and $\tilde g$ are functions from the convex set $\mathcal{F}_\Lambda$, and $g_t = g_0 + t(g_1 - g_0) \in \mathcal{F}$. It is not hard to see that Assumption Q2 follows from the fact that the mapping $t \mapsto \mathbf{E}_f[\Xi(t)]$ is concave for any triplet $g_0, g_1, \tilde g \in \mathcal{F}_\Lambda$. Let us prove now this concavity property. Since the functions $g_0, g_1, \tilde g$ are uniformly bounded we get that $\Xi(\cdot)$ is twice continuously differentiable and the differentiation inside the expectation $\mathbf{E}_f[\Xi(t)]$ is legitimate. Therefore,

$$\begin{aligned}\frac{d}{dt}\,\mathbf{E}_f[\Xi(t)] &= -2\beta^{-1}\mathbf{E}_f\Big[\big(\langle g_t, h\rangle - h(X_1)\big)\Xi(t)\Big],\\ \frac{d^2}{dt^2}\,\mathbf{E}_f[\Xi(t)] &= -2\beta^{-2}\mathbf{E}_f\Big[\Big(\beta\|h\|_2^2 - 2\big\{\langle g_t, h\rangle - h(X_1)\big\}^2\Big)\Xi(t)\Big],\end{aligned}$$

where $h = g_1 - g_0$, and

$$\frac{\beta^2}{2}\,\frac{d^2}{dt^2}\,\mathbf{E}_f[\Xi(t)] \leqslant -\big(\beta\|h\|_2^2 - 2\langle g_t, h\rangle^2\big)\mathbf{E}_f\big[\Xi(t)\big] + 2\mathbf{E}_f\Big[\big\{h(X_1)^2 - 2\langle g_t, h\rangle h(X_1)\big\}\Xi(t)\Big].$$

This leads to

$$\Xi(t) \leqslant \exp\Big[- \beta^{-1}\big\{\|g_t\|^2_{\mu,2} - \|\tilde g\|^2_{\mu,2}\big\} + 4RL/\beta\Big] := \Xi_1(t)$$

and

$$\mathbf{E}_f[\Xi(t)] \geqslant \exp\Big[- \beta^{-1}\big\{\|g_t\|^2_{\mu,2} - \|\tilde g\|^2_{\mu,2} + 4\max_{\mathcal{F}_\Lambda}\mathbf{E}_f[|g(X_1)|]\big\}\Big] = \Xi_1(t)e^{-4R(L+\sqrt{L})/\beta}.$$

Combining these estimates with inequalities

$$\mathbf{E}[h(X_1)^2] \leqslant L\|h\|_2^2, \quad |\langle g_t, h\rangle| \leqslant \|g_t\|_2\|h\|_2 \leqslant R\|h\|_2, \quad \mathbf{E}[|\langle g_t, h\rangle h(X_1)|] \leqslant RL\|h\|_2^2,$$

we get

$$\frac{\beta^2}{2}\,\frac{d^2}{dt^2}\,\mathbf{E}_f[\Xi(t)] \leqslant -\|h\|_2^2\Xi_1(t)\big((\beta - 2R^2)e^{-4R(L+\sqrt{L})/\beta} - 2L - 4RL\big) \leqslant 0,$$

whenever $(\beta - 2R^2)e^{-4R(L+\sqrt{L})/\beta} \geqslant 2L + 4RL$. This proves the concavity of $t \mapsto \mathbf{E}_f[\Xi(t)]$, and thus the proposition.

9.5. Proof of Proposition 6

In view of (26), for any prior π and any $\beta > 0$ the MA aggregate $\widehat{f}_n$ satisfies the inequality

$$\mathbf{E}_f[\ell_\Phi(\widehat{f}_n, f)] \leqslant \inf_{p\in\mathcal{P}_\Lambda} \Big(\int_\Lambda \ell_\Phi(f_{\boldsymbol\lambda}, f)\, p(d\boldsymbol\lambda) + \frac{\beta\mathcal{K}(p,\pi)}{n+1}\Big) + \beta \int_{\mathcal{Z}} \mathbf{E}_f[S_1(z,\mathbf{Z})]\, P_f(dz). \qquad (32)$$

with $S_1(z,\mathbf{Z})$ defined by $S_1(z,\mathbf{Z}) = \log\int_{\boldsymbol\lambda} \exp\big(-\beta^{-1}\{Q(z,f_{\boldsymbol\lambda}) - Q(z,\widehat{f}_n)\}\big)\,\widehat{\mu}_n(d\boldsymbol\lambda)$. Let us introduce the function $\psi_{\boldsymbol\lambda}(t) = \exp\big(-t\{Q(z,f_{\boldsymbol\lambda}) - Q(z,\widehat{f}_n)\}\big)$. This function is infinitely differentiable, equals one at the origin and we have $S_1(z,\mathbf{Z}) = \log\int_\Lambda \psi_{\boldsymbol\lambda}(\beta^{-1})\widehat{\mu}_n(d\boldsymbol\lambda)$. Using the Taylor formula, we get

$$\psi_{\boldsymbol\lambda}(t) \leqslant 1 + t\psi'_{\boldsymbol\lambda}(0) + \frac{t^2}{2}\big(Q(z,f_{\boldsymbol\lambda}) - Q(z,\widehat{f}_n)\big)^2 e^{tQ(z,\widehat{f}_n)}, \quad \forall t \geqslant 0.$$

Furthermore, since the hinge loss is convex, the Jensen inequality yields $\int_\Lambda \psi'_{\boldsymbol\lambda}(0)\,\widehat{\mu}_n(d\boldsymbol\lambda) \leqslant 0$. Replacing t by β^{-1} and using that $Q(z,\widehat{f}_n) \leqslant 1 + RL_\phi$, we get the inequalities

$$\begin{aligned} S_1(z,\mathbf{Z}) &= \log\int_\Lambda \psi_{\boldsymbol\lambda}(\beta^{-1})\widehat{\mu}_n(d\boldsymbol\lambda) \leqslant \log\Big(1 + \frac{e^{(1+RL_\phi)/\beta}}{2\beta^2}\int_\Lambda \big(Q(z,f_{\boldsymbol\lambda}) - Q(z,\widehat{f}_n)\big)^2\widehat{\mu}_n(d\boldsymbol\lambda)\Big) \\ &\leqslant \frac{e^{(1+RL_\phi)/\beta}}{2\beta^2}\int_\Lambda \big(Q(z,f_{\boldsymbol\lambda}) - Q(z,\widehat{f}_n)\big)^2\widehat{\mu}_n(d\boldsymbol\lambda) \leqslant \frac{2e^{(1+RL_\phi)/\beta}}{\beta^2}(1+RL_\phi)^2. \end{aligned}$$

Thus we obtain

$$\mathbf{E}_f[\ell_\Phi(\widehat{f}_n, f)] \leqslant \inf_{p\in\mathcal{P}_\Lambda} \Big(\int_\Lambda \ell_\Phi(f_{\boldsymbol\lambda}, f)\, p(d\boldsymbol\lambda) + \frac{\beta\mathcal{K}(p,\pi)}{n+1}\Big) + \frac{2(1+RL_\phi)^2 e^{(1+RL_\phi)/\beta}}{\beta}, \qquad (33)$$

which is valid for any prior π. Note that the term with the infimum in (33) coincides with the right hand side of the oracle inequality of Theorem 1. Therefore, when the sparsity prior is used, this term can be bounded from above using Remark 3 with $\bar{L}_f(\boldsymbol\lambda) = \int_{\mathcal{X}} |\eta(x)| K_\varepsilon\big(f_{\boldsymbol\lambda}(x) - f(x)\big)\, P_X(dx)$. Since also $|\eta(x)| \leqslant 1$, we get

$$\begin{aligned} \mathbf{E}_f[\ell(\widehat{f}_n, f)] \leqslant \min_{\|\boldsymbol\lambda^*\|_1 \leqslant R-2M\tau} \Big(\ell(f_{\boldsymbol\lambda^*}, f) + \frac{2\beta}{n+1}\Big\{\alpha\|\boldsymbol\lambda^*\|_1 + \sum_{j=1}^M \log(1+\tau^{-1}|\lambda_j^*|)\Big\}\Big) + \varepsilon + 4\tau^2\mathrm{Tr}(\bar{\mathcal{M}}_\varepsilon) \\ + \frac{2(1+RL_\phi)^2 e^{(1+RL_\phi)/\beta}}{\beta}, \end{aligned}$$

where the entries of the matrix $\bar{\mathcal{M}}_\varepsilon$ are $\varepsilon^{-1}\int_{\mathcal{X}} |\eta(x)|\phi_j(x)\phi_{j'}(x)P_X(dx)$ with $i,j = 1,\dots,M$. Thus, $\mathrm{Tr}(\bar{\mathcal{M}}_\varepsilon) \leqslant L_\phi^2 M\varepsilon^{-1}$, and we get the result of the proposition by minimizing the right hand side of the last display with respect to $\varepsilon > 0$.

Aknowledgments

The authors acknowledge the financial support by ANR under grant PARCIMONIE.

References

[1] Abramovich, F., Grinshtein, V. and Pensky, M. (2007). On optimality of Bayesian testimation in the normal means problem. *Ann. Statist.* **35** 2261–2286.

[2] Alquier, P. (2008). Pac-Bayesian bounds for randomized empirical risk minimizers. *Math. Methods Statist.* **17** 1–26.

[3] Audibert, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *Ann. Statist.* **37** 1591–1646.

[4] Barron, A. (1987). Are Bayes rules consistent in information? In *Open Problems in Communication and Computation* (T. M. Cover and B. Gopinath, eds.) 85-91. Springer, New York.

[5] Bartlett, P., Jordan, M. and McAuliffe, J. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156.

[6] Bickel, P., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732.

[7] Bickel, P., Ritov, Y. and Tsybakov, A.B. (2010). Hierarchical selection of variables in sparse high-dimensional regression. In *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown. IMS Collections*, **6**, 56-69, Institute of Mathematical Statistics.

[8] Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384.

[9] Bunea, F. and Nobel, A.B. (2008). Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Trans. Inform. Theory* **54** 1725–1735 .

[10] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2004). Aggregation for regression learning. ArXiv:math/0410214.

[11] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2006). Aggregation and sparsity via ℓ_1-penalized least squares. *Learning theory* 379–391, Lecture Notes in Comput. Sci., 4005, Springer.

[12] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007). Aggregation for gaussian regression. *Ann. Statist.* **35** 1674–1697.

[13] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194.

[14] Candès, E. (2006). Compressive sampling. *Int. Congress of Mathematics* **3** 1433–1452, Madrid, Spain.

[15] Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351.

[16] Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization.* Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. Lecture Notes in Mathematics, 1851. Springer-Verlag, Berlin, 2004.

[17] Catoni, O. (2007). *Pac-Bayesian supervised classification: the thermodynamics of statistical learning.* Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, Ohio.

[18] Cesa-Bianchi, N., Conconi, A. and Gentile, G. (2004). On the generalization ability of on-line learning algorithms. *IEEE Trans. Inform. Theory* **50** 2050–2057.

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

[20] Dalalyan A. S. and Tsybakov, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. Learning theory, 97–111, Lecture Notes in Comput. Sci., 4539, Springer, Berlin.

[21] Dalalyan, A. and Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *Machine Learning* **72** 39–61.

[22] Dalalyan, A. and Tsybakov, A. B. (2009). Sparse regression learning by aggregation and Langevin Monte-Carlo. *Proceedings of COLT-2009.* Published online.

[23] Dalalyan, A. and Tsybakov, A. B. (2010). Sparse regression learning by aggregation and Langevin Monte-Carlo. Submitted to *Journal of Computer and System Sciences.* ArXiv:0903.1223(v3).

[24] Donoho, D.L., Elad, M. and Temlyakov, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** 6–18.

[25] Dembo, A. and Zeitouni, O. (1998) *Large Deviations Techniques and Applications.* Springer, New York.

[26] Gaïffas, S. and Lecu, G. (2009) Hyper-sparse optimal aggregation. arXiv:0912.1618

[27] Goldenshluger, A. (2009). A universal procedure for aggregating estimators. *Ann. Statist.* **37** 542-568.

[28] Giraud, C. (2008). Mixing Least-square estimators when the variance is unknown. *Bernoulli* **14** 1089–1107.

[29] Johnstone, I. and Silverman, B.W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752.

[30] Juditsky, A., Rigollet, P., and Tsybakov, A.B. (2008). Learning by mirror averaging. *Ann. Statist.* **36** 2183–2206.

[31] Juditsky, A.B., Nazin, A.V., Tsybakov, A.B. and Vayatis, N. (2005). Recursive aggregation of estimators via the Mirror Descent Algorithm with averaging. *Problems of Information Transmission* **41** 368 – 384.

[32] Haussler, D., Kivinen, J. and Warmuth, M. (1998). Sequential prediction of individual sequences under general loss functions. *IEEE Trans. Inform. Theory* **44** 1906–1925

[33] Koltchinskii, V. (2009). Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.* **45** 7–57.

[34] Koltchinskii, V. (2009). Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.* **37** 1332–1359.

[35] Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, **15** 799–828.

[36] Lecué, G. (2007). Optimal rates of aggregation in classification under low noise assumption. *Bernoulli* **13** 1000–1022.

[37] Lecué, G. and Mendelson, S. On the optimality of the aggregate with exponential weights. *Submitted*, `http://www.maths.anu.edu.au/~mendelso/papers/LM26-03-10.pdf`

[38] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462.

[39] Leung, G., and Barron, A. (2006). Information theory and mixing least-square regressions. *IEEE Trans. Inform. Theory* **52** 3396–3410.

[40] Klemelä, J. (2009). *Smoothing of Multivariate Data.* Wiley, New York.

[41] Lounici, K. (2007). Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist.* **16** 246–259.

[42] McAllester, D. (2003). PAC-Bayesian stochastic model selection. *Machine Learning* **51**

5–21.

[43] Rigollet, Ph. and Tsybakov, A.B. (2010) Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, to appear.

[44] Rivoirard, V. (2006). Non linear estimation over weak Besov spaces and minimax Bayes method. *Bernoulli* **12** 609–632.

[45] Salmon, J. and Le Pennec, E. (2009). NL-Means and aggregation procedures. *Proc. ICIP 2009* 2941–2944.

[46] Seeger, M. (2008). Bayesian Inference and Optimal Design in the Sparse Linear Model *J. Mach. Learn. Res.* **9** 759–813.

[47] Tsybakov, A. B. (2003). Optimal rates of aggregation. Computational Learning Theory and Kernel Machines. B. Schölkopf and M. Warmuth, eds. Lecture Notes in Artificial Intelligence **2777** 303–313. Springer, Heidelberg.

[48] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166.

[49] van de Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614–645.

[50] Vovk, V. (1990). Aggregating Strategies. In: Proceedings of the 3rd Annual Workshop on Computational Learning Theory, COLT1990, CA: Morgan Kaufmann 371–386.

[51] Yang, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc.* **96** 574–588.

[52] Yang, Y. (2003). Regression with multiple candidate models: selecting or mixing? *Statist. Sinica* **13** 783–809.

[53] Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* **10** 25–47.

[54] Zhang, C.-H., Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594.

[55] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* **32** 56–85.

[56] Zhang, T. (2006). From epsilon-entropy to KL-complexity: analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180–2210.

[57] Zhang, T. (2009). Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.* **37** 2109–2144.

[58] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563.

[59] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **67** 301–320.

[60] Zou, H. (2006). The Adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429.

Received January 0000