Statistical Science
2012, Vol. 27, No. 4, 558–575
DOI: 10.1214/12-STS393

Sparse Estimation by Exponential Weighting

Philippe Rigollet and Alexandre B. Tsybakov

Abstract. Consider a regression model with fixed design and Gaussian noise where the regression function can potentially be well approximated by a function that admits a sparse representation in a given dictionary. This paper resorts to exponential weights to exploit this underlying sparsity by implementing the principle of *sparsity pattern aggregation*. This model selection take on sparse estimation allows us to derive sparsity oracle inequalities in several popular frameworks, including ordinary sparsity, fused sparsity and group sparsity. One striking aspect of these theoretical results is that they hold under *no condition in the dictionary*. Moreover, we describe an efficient implementation of the sparsity pattern aggregation principle that compares favorably to state-of-the-art procedures on some basic numerical examples.

Key words and phrases: High-dimensional regression, exponential weights, sparsity, fused sparsity, group sparsity, sparsity oracle inequalities, sparsity pattern aggregation, sparsity prior, sparse regression.

1. INTRODUCTION

Since the 1990s, the idea of exponential weighting has been successfully used in a variety of statistical problems. In this paper, we review several properties of estimators based on exponential weighting with a particular emphasis on how they can be used to construct optimal and computationally efficient procedures for high-dimensional regression under the sparsity scenario.

Most of the work on exponential weighting deals with a regression learning problem. Some of the results can be extended to other statistical models such as density estimation or classification; cf. Section 6. For the sake of brevity and to make the presentation more transparent, we focus here on the following framework considered in Rigollet and Tsybakov (2011). Let $\mathcal{Z} = \{(x_1, Y_1), \dots, (x_n, Y_n)\}$ be a collection of independent random pairs such that $(x_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is an arbitrary set. Assume the regression model

$$Y_i = \eta(x_i) + \xi_i, \quad i = 1, \dots, n, \tag{1.1}$$

where $\eta : \mathcal{X} \to \mathbb{R}$ is the unknown regression function, and the errors ξ_i are independent Gaussian $\mathcal{N}(0, \sigma^2)$. The covariates are deterministic elements $x_1, \dots, x_n$ of $\mathcal{X}$. For any function $f : \mathcal{X} \to \mathbb{R}$, we define a seminorm $\|\cdot\|$ by[1]

$$\|f\|^2 = \frac{1}{n} \sum_{i=1}^{n} f^2(x_i).$$

We adopt the following learning setup. Let $\mathcal{H} = \{f_1, \dots, f_M\}$, be a dictionary of $M \geq 1$ given functions. For example, f_j can be some basis functions or some preliminary estimators of f constructed from another sample that we consider as frozen; see Section 4 for more details. Our goal is to approximate the regression function η by a linear combination $\mathfrak{f}_\theta(x) = \sum_{j=1}^{M} \theta_j f_j(x)$ with weights $\theta = (\theta_1, \dots, \theta_M)$, where

Philippe Rigollet is Assistant Professor, Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544, USA (e-mail: rigollet@princeton.edu). Alexandre B. Tsybakov is Professor and Head, Laboratoire de Statistique, CREST-ENSAE, 3, av. Pierre Larousse, F-92240 Malakoff Cedex, France (e-mail: alexandre.tsybakov@ensae.fr).

[1] Without loss of generality, in what follows we will associate all the functions with vectors in $\mathbb{R}^n$ since only the values of functions at points $x_1, \dots, x_n$ will appear in the risk. So, $\|\cdot\|$ will be indeed a norm and, with no ambiguity, we will use other related notation such as $\|\mathbf{Y} - f\|$ where $\mathbf{Y}$ is a vector in $\mathbb{R}^n$ with components $Y_1, \dots, Y_n$.

possibly $M \gg n$. The performance of a given estimator $\hat{f}$ of a function η is measured in terms of its averaged squared error

$$R(\hat{f}) = \|\hat{f} - \eta\|^2 := \frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i) - \eta(x_i)]^2.$$

Let Θ be a given subset of $\mathbb{R}^M$. In the aggregation problem, we would ideally wish to find an *aggregated estimator* $\hat{f}$ whose risk $R(\hat{f})$ is as close as possible in a probabilistic sense to the minimum risk $\inf_{\theta\in\Theta} R(\mathrm{f}_\theta)$. Namely, one can construct estimators $\hat{f}$ satisfying the following property:

$$\mathbb{E}R(\hat{f}) \le C \inf_{\theta\in\Theta} R(\mathrm{f}_\theta) + \delta_{n,M}(\Theta), \tag{1.2}$$

where $\delta_{n,M}(\Theta)$ is a small remainder term characterizing the performance of the given aggregate $\hat{f}$ and the complexity of the set Θ, $C \ge 1$ is a constant, and $\mathbb{E}$ denotes the expectation. Bounds of the form (1.2) are called *oracle inequalities*. In some cases, even more general results are available. They have the form

$$\mathbb{E}R(\hat{f}) \le C \inf_{\theta\in\Theta'}\{R(\mathrm{f}_\theta) + \Delta_{n,M}(\theta)\}, \tag{1.3}$$

where $\Delta_{n,M}$ is a remainder term that characterizes the performance of the given aggregate $\hat{f}$ and the complexity of the parameter $\theta \in \Theta' \subseteq \mathbb{R}^M$ (often $\Theta' = \mathbb{R}^M$). To distinguish from (1.2), we will call bounds of the form (1.3) the *balanced oracle inequalities*. If $\Theta \subseteq \Theta'$, then (1.2) is a direct consequence of (1.3) with $\delta_{n,M}(\Theta) = C\sup_{\theta\in\Theta}\Delta_{n,M}(\theta)$.

In this paper, we mainly focus on the case where the complexity of a vector θ is measured as the number of its nonzero coefficients $|\theta|_0$. In this case, inequalities of the form (1.3) are sometimes called *sparsity oracle inequalities*. Other measures of complexity, also related to *sparsity* are considered in Section 5.2. As indicated by the notation and illustrated below, the remainder term $\Delta_{n,M}(\theta)$ depends explicitly on the size M of the dictionary and the sample size n. It reflects the interplay between these two fundamental parameters and also the complexity of θ.

When the linear model is misspecified, that is, where there is no $\theta \in \Theta$ such that $\eta = \mathrm{f}_\theta$ on the set $\{x_1, \dots, x_n\}$, the minimum risk satisfies $\inf_{\theta\in\Theta} R(\mathrm{f}_\theta) > 0$ leading to a systematic bias term. Since this term is unavoidable, we wish to make its contribution as small as possible, and it is therefore important to obtain a leading constant $C = 1$. Many oracle inequalities with leading constant $C > 1$ can be found in the literature for related problems. However, in most of the papers, the set $\Theta = \Theta_n$ depends on the sample size n in such a way that $\inf_{\theta\in\Theta_n} R(\mathrm{f}_\theta)$ tends to 0 as n goes to infinity, under additional regularity assumptions. In this paper, we are interested in the case where Θ is fixed. For this reason, we consider here only oracle inequalities with leading constant $C = 1$ (called *sharp oracle inequalities*). Because they hold for finite M and n, these are truly finite sample results.

One salient feature of the oracle approach as opposed to standard statistical reasoning, is that it does not rely on an underlying model. Indeed, the goal is not to estimate the parameters of an underlying "true" model but rather to construct an estimator that mimics, in terms of an appropriate oracle inequality, the performance of the best model in a given class, whether this model is true or not. From a statistical viewpoint, this difference is significant since performance cannot be evaluated in terms of parameters. Indeed, there is no true parameter. However, we can still compare the risk of the estimator with the optimum value. Oracle inequalities offer a tool for such a comparison.

A particular choice of Θ corresponds to the problem of *model selection aggregation*. Let $\Theta = \Theta^{\mathrm{MC}}$ be the set of M canonical basis vectors of $\mathbb{R}^M$. Then the set of linear combinations $\{\mathrm{f}_\theta, \theta \in \Theta^{\mathrm{MC}}\}$ coincides with the initial dictionary of functions $\mathcal{H} = \{f_1, \dots, f_M\}$, so that the goal of model selection is to mimic the best function in the dictionary in the sense of the risk measure $R(\cdot)$. This can be done in different ways, leading to different rates $\delta_{n,M}(\Theta^{\mathrm{MC}})$; however one is mostly interested in the methods that attain the rate $\delta^*_{n,M}(\Theta^{\mathrm{MC}}) \asymp (\log M)/n$ which is known to be minimax optimal (see Tsybakov, 2003; Bunea, Tsybakov and Wegkamp, 2007; Rigollet, 2012). The first sharp oracle inequalities with this rate for a setting different from the one considered here were obtained by Catoni (1999) (see also Catoni, 2004), who used the progressive mixture method based on exponential weighting. Other methods of model selection for aggregation consist in selecting a function in the dictionary by minimizing a (penalized) empirical risk (see, e.g., Nemirovski, 2000; Wegkamp, 2003; Tsybakov, 2003; Lecué, 2012). One of the major novelties offered by exponential weighting is to *combine* (average) the functions in the dictionary using a convex combination, and not simply to *select* one of them. From the theoretical point of view, selection of one of the functions has a fundamental drawback since it does not attain the optimal rate $(\log M)/n$; cf. Section 2.

The rest of the paper is organized as follows. In the next section, we discuss some connections between the

exponential weighting schemes and penalized empirical risk minimization. In Section 3, we present the first oracle inequalities that demonstrate how exponential weighting can be used to efficiently combine functions in a dictionary. The results of Section 3 are then extended to the case where one wishes to combine not deterministic functions, but estimators. Oracle inequalities for this problem are discussed in Section 4. They are based on the work of Leung and Barron (2006) and Dalalyan and Salmon (2011). Section 5 shows how these results can be adapted to deal with sparsity. We introduce the principle of *sparsity pattern aggregation*, and we derive sparsity oracle inequalities in several popular frameworks including ordinary sparsity, fused sparsity and group sparsity. Finally, we describe an efficient implementation of the sparsity pattern aggregation principle and compare its performance to state-of-the-art procedures on some basic numerical examples.

2. EXPONENTIAL WEIGHTING AND PENALIZED RISK MINIMIZATION

2.1 Suboptimality of Selectors

A natural candidate to solve the problem of model selection introduced in the previous section is an empirical risk minimizer. Define the empirical risk by

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - f(x_i)]^2 = \|\mathbf{Y} - f\|^2$$

and the empirical risk minimizer by

$$\hat{f}^{\mathrm{ERM}} = \operatorname*{argmin}_{f\in\mathcal{H}} \hat{R}_n(f), \tag{2.1}$$

where ties are broken arbitrarily. However, while this procedure satisfies an exact oracle inequality, it fails to exhibit the optimal rate of order $\delta^*_{n,M}(\Theta^{\mathrm{MC}}) \asymp (\log M)/n$. The following result shows that this defect is intrinsic not only to empirical risk minimization but also to any method that selects only one function in the dictionary $\mathcal{H}$. This includes methods of model selection by penalized empirical risk minimization. We call estimators $\hat{S}_n$, taking values in $\mathcal{H}$ the *selectors*.

THEOREM 2.1. *Assume that* $\|f_j\| \le 1$ *for any* $f_j \in \mathcal{H}$. *Any empirical risk minimizer* $\hat{f}^{\mathrm{ERM}}$ *defined in* (2.1) *satisfies the following oracle inequality*:

$$\mathbb{E}R(\hat{f}^{\mathrm{ERM}}) \le \min_{1\le j\le M} R(f_j) + 4\sigma\sqrt{\frac{2\log M}{n}}. \tag{2.2}$$

Moreover, assume that

$$(\sigma \vee 1)\sqrt{(\log M)/n} \le C_0 \tag{2.3}$$

for $0 < C_0 < 1$ *small enough. Then, there exists a dictionary* $\mathcal{H} = \{f_1, \dots, f_M\}$ *with* $\|f_j\| \le 1$, $j = 1, \dots, M$, *such that the following holds. For any selector* $\hat{S}_n$, *and in particular, for any selector based on penalized empirical risk minimization, there exists a regression function* η *such that* $\|\eta\| \le 1$ *and*

$$\mathbb{E}R(\hat{S}_n) \ge \min_{1\le j\le M} R(f_j) + C_*\sigma\sqrt{\frac{\log M}{n}} \tag{2.4}$$

for some positive constant C_*.

PROOF. See the Appendix. □

It follows from the lower bound (2.4) that *selecting* one of the functions in a finite dictionary $\mathcal{H}$ to solve the problem of model selection is suboptimal in the sense that it exhibits a too large remainder term, of the order $\sqrt{(\log M)/n}$. It turns out that we can do better if we take a *mixture*, that is, a convex combination of the functions in $\mathcal{H}$. We will see in Section 3 [cf. (3.4)] that under a particular choice of weights in this convex combination, namely the *exponential weights*, one can achieve oracle inequalities with much better rate $(\log M)/n$. This rate is known to be optimal in a minimax sense in several regression setups, including the present one (see Tsybakov, 2003; Bunea, Tsybakov and Wegkamp, 2007; Rigollet, 2012).

2.2 Exponential Weighting as a Penalized Procedure

Penalized empirical risk minimization for model selection has received a lot of attention in the literature, and many choices for the penalty can be considered (see, e.g., Birgé and Massart, 2001; Bartlett, Boucheron and Lugosi, 2002; Wegkamp, 2003; Lugosi and Wegkamp, 2004; Bunea, Tsybakov and Wegkamp, 2007) to obtain oracle inequalities with the optimal or near optimal remainder term. However, all these inequalities exhibit a constant $C > 1$ in front of the leading term. This is not surprising as we have proved in the previous section that it is impossible for selectors to satisfy oracle inequalities like (1.2) that are both sharp (i.e., with $C = 1$) and have the optimal remainder term. To overcome this limitation of selectors, we look for estimators obtained as convex combinations of the functions in the dictionary.

The coefficients of convex combinations belong to the flat simplex

$$\Lambda^M := \left\{\lambda \in \mathbb{R}^M : \lambda_j \ge 0, \sum_{j=1}^{M}\lambda_j = 1\right\}.$$

Let us now examine a few ways to obtain potentially good convex combinations. One candidate is a solution of the following penalized empirical risk minimization problem:

$$\min_{\lambda\in\Lambda^M}\{\hat{R}_n(\mathrm{f}_\lambda)+\mathrm{pen}(\lambda)\},$$

where $\mathrm{pen}(\cdot)\geq 0$ is a penalty function. This choice looks quite natural since it provides a proxy of the right-hand side of the oracle inequality (1.3) where the unknown risk $R(\cdot)$ is replaced by its empirical counterpart $\hat{R}_n(\cdot)$. The minimum is taken over the simplex Λ^M because we are looking for a convex combination. Clearly, the penalty $\mathrm{pen}(\cdot)$ should be carefully chosen and ideally should match the best remainder term $\Delta_{n,M}(\cdot)$. Yet, this problem may be difficult to solve as it involves a minimization over Λ^M. Instead, we propose to solve a simpler problem. Consider the following linear upper bound on the empirical risk:

$$\sum_{j=1}^M \lambda_j\hat{R}_n(f_j)\geq \hat{R}_n(\mathrm{f}_\lambda)\quad\forall\lambda\in\Lambda^M$$

and solve the following optimization problem:

$$\min_{\lambda\in\Lambda^M}\left\{\sum_{j=1}^M \lambda_j\hat{R}_n(f_j)+\mathrm{pen}(\lambda)\right\}. \tag{2.5}$$

Note that if $\mathrm{pen}(\lambda)\equiv 0$, the solution $\hat{\lambda}$ of (2.5) is simply the empirical risk minimizer over the vertices of the simplex so that $\mathrm{f}_{\hat{\lambda}}=\hat{f}^{\mathrm{ERM}}$. In general, depending on the penalty function, this problem may be more or less difficult to solve. It turns out that the Kullback–Leibler penalty leads to a particularly simple solution and allows us to approximate the best remainder term $\Delta_{n,M}(\cdot)$ thus adding great flexibility to the resulting estimator.

Observe that vectors in Λ^M can be associated to probability measures on $\{1,\dots,M\}$. Let $\lambda=(\lambda_1,\dots,\lambda_M)$ and $\pi=(\pi_1,\dots,\pi_M)$ be two probability measures on $\{1,\dots,M\}$, and define the Kullback–Leibler divergence between λ and π by

$$\mathcal{K}(\lambda,\pi)=\sum_{j=1}^M\lambda_j\log\left(\frac{\lambda_j}{\pi_j}\right)\geq 0.$$

Here and in the sequel, we adopt the convention that $0\log 0=0$, $0\log(a/0)=0$, and $\log(a/0)=\infty$, for any $a>0$.

Exponential weights can be obtained as the solution of the following minimization problem. Fix $\beta>0$, a *prior* $\pi\in\Lambda^M$, and define the vector $\hat{\lambda}^\pi$ by

$$\hat{\lambda}^\pi=\operatorname*{argmin}_{\lambda\in\Lambda^M}\left\{\sum_{j=1}^M\lambda_j\hat{R}_n(f_j)+\frac{\beta}{n}\mathcal{K}(\lambda,\pi)\right\}. \tag{2.6}$$

This constrained convex optimization problem has a unique solution that can be expressed explicitly. Indeed, it follows from the Karush–Kuhn–Tucker (KKT) conditions that the components $\hat{\lambda}^\pi_j$ of $\hat{\lambda}^\pi$ satisfy

$$n\hat{R}_n(f_j)+\beta\log\left(\frac{\hat{\lambda}^\pi_j}{\pi_j}\right)+\mu-\delta_j=0,\qquad j=1,\dots,M, \tag{2.7}$$

where $\mu,\delta_1,\dots,\delta_M\geq 0$ are Lagrange multipliers, and

$$\hat{\lambda}^\pi_j\geq 0,\quad \delta_j\hat{\lambda}^\pi_j=0,\quad \sum_{j=1}^M\hat{\lambda}^\pi_j=1.$$

Equation (2.7) together with the above constraints lead to the following closed form solution:

$$\hat{\lambda}^\pi_j=\frac{\exp(-n\hat{R}_n(f_j)/\beta)\pi_j}{\sum_{k=1}^M\exp(-n\hat{R}_n(f_k)/\beta)\pi_k},\qquad j=1,\dots,M, \tag{2.8}$$

called the *exponential weights*. We see that one immediate effect of penalizing by the Kullback–Leibler divergence is that the solution of (2.6) is not a selector. As a result, it achieves the desired effect of averaging as opposed to selecting.

3. ORACLE INEQUALITIES

An *aggregate* is an estimator defined as a weighted average of the functions in the dictionary $\mathcal{H}$ with some data-dependent weights. We focus on the aggregate with exponential weights,

$$\hat{f}^\pi=\sum_{j=1}^M\hat{\lambda}^\pi_j f_j,$$

where $\hat{\lambda}^\pi_j$ is given in (2.8). This estimator satisfies the following oracle inequality.

THEOREM 3.1. *The aggregate $\hat{f}^\pi$ with $\beta\geq 4\sigma^2$ satisfies the following balanced oracle inequality*

$$\mathbb{E}R(\hat{f}^\pi)\leq\min_{\lambda\in\Lambda^M}\left\{\sum_{j=1}^M\lambda_j R(f_j)+\frac{\beta}{n}\mathcal{K}(\lambda,\pi)\right\}. \tag{3.1}$$

Comparing with (2.6) we see that $\hat{\lambda}^{\pi}$ is the minimizer of the unbiased estimator of the right-hand side of (3.1). The proof of Theorem 3.1 can be found in the papers of Dalalyan and Tsybakov (2007, 2008) containing more general results. In particular, they apply to non-Gaussian distributions of errors ξ_i and to exponential weights with a general (not necessarily discrete) probability distribution π on $\mathbb{R}^M$. Dalalyan and Tsybakov (2007, 2008) show that the corresponding exponentially weighted aggregate $\hat{f}_*^{\pi}$ satisfies the following bound:

$$\mathbb{E}R(\hat{f}_*^{\pi}) \le \inf_p \left\{ \int R(\mathrm{f}_\theta)\, p(\mathrm{d}\theta) + \frac{\beta}{n} \mathcal{K}(p,\pi) \right\}, \tag{3.2}$$

where the infimum is taken over all probability distributions p on $\mathbb{R}^M$, and $\mathcal{K}(p,\pi)$ denotes the Kullback–Leibler divergence between the general probability measures p and π. Bound (3.1) follows immediately from (3.2) by taking p and π as discrete distributions.

A useful consequence of (3.1) can be obtained by restricting the minimum on the right-hand side to the vertices of the simplex Λ^M. These vertices are precisely the vectors $e^{(1)}, \dots, e^{(M)}$ that form the canonical basis of $\mathbb{R}^M$ so that

$$\sum_{j=1}^{M} e_j^{(k)} R(f_j) = R(f_k),$$

where $e_j^{(k)} = \delta_{jk}$ is the jth coordinate of $e^{(k)}$, with δ_{jk} denoting the Kronecker delta. It yields

$$\mathbb{E}R(\hat{f}^{\pi}) \le \min_{1\le j\le M} \left\{ R(f_j) + \frac{\beta}{n} \log(\pi_j^{-1}) \right\}. \tag{3.3}$$

Taking π to be the uniform distribution on $\{1, \dots, M\}$ leads to the following oracle inequality:

$$\mathbb{E}R(\hat{f}^{\pi}) \le \min_{1\le j\le M} R(f_j) + \frac{\beta \log M}{n}, \tag{3.4}$$

that exhibits a remainder term of the optimal order $(\log M)/n$.

The role of the distribution π is to put a prior weight on the functions in the dictionary. When there is no preference, the uniform prior is a common choice. However, we will see in Section 5 that choosing nonuniform weights depending on suitable *sparsity* characteristics can be very useful. Moreover, this methodology can be extended to many cases where one wishes to learn with a prior. It is worth mentioning that while the terminology is reminiscent of a Bayesian setup, this paper deals only with a frequentist setting (the risk is not averaged over the prior).

4. AGGREGATION OF ESTIMATORS

4.1 From Aggregation of Functions to Aggregation of Estimators

Akin to the setting of the previous section, exponential weights were originally introduced to aggregate deterministic functions f_j from a dictionary. These functions can be chosen in essentially two ways. Either they have good approximation properties such as an (overcomplete) basis of functions or they are constructed as preliminary estimators using a hold-out sample. The latter case corresponds to the problem of *aggregation of estimators* originally described in Nemirovski (2000). The idea put forward by Nemirovski (2000) is to obtain two independent samples from the initial one by randomization; estimators are constructed from the first sample while the second is used to perform aggregation. To carry out the analysis of the aggregation step, it is enough to work conditionally on the first sample so that the problem reduces to aggregation of deterministic functions. A limitation is that Nemirovski's randomization only applies to Gaussian model with known variance. Nevertheless, this idea of two-step procedures carries over to models with i.i.d. observations where one can do direct sample splitting (see, e.g., Yang, 2004; Rigollet and Tsybakov, 2007; Lecué, 2007). Thus, in many cases aggregation of estimators can be achieved by reduction to aggregation of functions.

Along with this approach, one can aggregate estimators using the same observations for both estimation and aggregation. While for general estimators this would clearly result in overfitting, the idea proved to be successful for certain types of estimators, first for projection estimators (Leung and Barron, 2006) and more recently for a more general class of linear (affine) estimators (Dalalyan and Salmon, 2011). Our further analysis will be based on this approach. Clearly, direct sample splitting does not apply to independent samples that are not identically distributed as in the present setup. Indeed, the observations in the first sample no longer have the same distribution as those in the second sample. On the other hand, the approach based on Nemirovski's randomization can be still applied, but it leads to somewhat weaker results involving an additional expectation over a randomization distribution and a bigger remainder term than in our oracle inequalities.

4.2 Aggregation of Linear Estimators

Suppose that we are given a finite family $\{\hat{f}_1, \dots, \hat{f}_K\}$ of linear estimators defined by

$$\hat{f}_j(x) = \mathbf{Y}^\top a_j(x), \tag{4.1}$$

where $a_j(\cdot)$ are given functions with values in $\mathbb{R}^n$. This representation is quite general; for example, $\hat{f}_j$ can be ordinary least squares, (kernel) ridge regression estimators or diagonal linear filter estimators; see Kneip (1994); Dalalyan and Salmon (2011) for a longer list of relevant examples. The vector of values $(\hat{f}_j(x_i), i = 1, \dots, n)$ equals to $A_j\mathbf{Y}$ where A_j an $n \times n$ matrix with rows $a_j(x_i), i = 1, \dots, n$.

Now, we would like to consider mixtures of such estimators rather than mixtures of deterministic functions as in the previous sections. For this purpose, exponential weights have to be slightly modified. Indeed, note that in Section 2, the risk of a deterministic function f_j is simply estimated by the empirical risk $\hat{R}_n(f_j)$, which is plugged into the expression for the weights. Clearly, $\mathbb{E}\hat{R}_n(f_j) = R(f_j) + \sigma^2$ so that $\hat{R}_n(f_j)$ is an *unbiased* estimator of the risk $R(f_j)$ of f_j up to an additive constant. For a linear estimator $\hat{f}_j$ defined in (4.1), $\hat{R}_n(\hat{f}_j) - \sigma^2$ is no longer an unbiased estimator of the risk $\mathbb{E}R(\hat{f}_j)$. It is well known that the risk of the linear estimator $\hat{f}_j$ has the form

$$\mathbb{E}R(\hat{f}_j) = \|(A_j - \mathbf{I})\eta\|^2 + \frac{\sigma^2}{n}\operatorname{Tr}[A_j^\top A_j],$$

where $\operatorname{Tr}[A]$ denotes the trace of a matrix A, and $\mathbf{I}$ denotes the $n \times n$ identity matrix. Moreover, an unbiased estimator of $\mathbb{E}R(\hat{f}_j)$ is given by a version of Mallows's C_p,

$$\tilde{R}_n^{\mathrm{unb}}(\hat{f}_j) = \|\mathbf{Y} - \hat{f}_j\|^2 + \frac{2\sigma^2}{n}\operatorname{Tr}[A_j] - \sigma^2. \tag{4.2}$$

Then, for linear estimators, the exponential weights and the corresponding aggregate are modified as follows:

$$\begin{aligned} \hat{\lambda}_j^\pi &= \frac{\exp(-n\tilde{R}_n^{\mathrm{unb}}(\hat{f}_j)/\beta)\pi_j}{\sum_{k=1}^K \exp(-n\tilde{R}_n^{\mathrm{unb}}(\hat{f}_k)/\beta)\pi_k}, \\ \hat{f}^\pi &= \sum_{k=1}^K \hat{\lambda}_k^\pi \hat{f}_k. \end{aligned} \tag{4.3}$$

Note that for deterministic f_j, we naturally define $\tilde{R}_n^{\mathrm{unb}}(f_j) = \hat{R}_n(f_j) - \sigma^2$, so that definition (4.3) remains consistent with (2.8). With this more general definition of exponential weights, Dalalyan and Salmon (2011) prove the following risk bounds for the aggregate $\hat{f}^\pi$.

THEOREM 4.1. *Let $\{\hat{f}_1, \dots, \hat{f}_K\}$ be a family of linear estimators defined in (4.1) such that the matrices A_j are symmetric, positive definite and $A_jA_k = A_kA_j$, for all $1 \le j, k \le K$. Then the exponentially weighted aggregate $\hat{f}^\pi$ defined in (4.3) with $\beta \ge 8\sigma^2$ satisfies*

$$\mathbb{E}R(\hat{f}^\pi) \le \min_{\lambda \in \Lambda^K}\left\{\sum_{j=1}^K \lambda_j \mathbb{E}R(\hat{f}_j) + \frac{\beta}{n}\mathcal{K}(\lambda, \pi)\right\}, \tag{4.4}$$

$$\mathbb{E}R(\hat{f}^\pi) \le \min_{j=1,\dots,K}\left\{\mathbb{E}R(\hat{f}_j) + \frac{\beta}{n}\log(\pi_j^{-1})\right\}. \tag{4.5}$$

If all the A_j are projection matrices ($A_j^\top = A_j$, $A_j^2 = A_j$), then the above inequalities hold with $\beta \ge 4\sigma^2$.

Here, bound (4.5) follows immediately from (4.4). In the rest of the paper, we mainly use the last part of this theorem concerning projection estimators. The bound (4.5) for this particular case was originally proved in Leung and Barron (2006). The result of Dalalyan and Salmon (2011) is, in fact, more general than Theorem 4.1 covering nondiscrete priors in the spirit of (3.2), and it applies not only to linear, but also to affine estimators $\hat{f}_j$.

5. SPARSE ESTIMATION

The family of projection estimators that we consider in this section is the family of all 2^M least squares estimators, each of which is characterized by its sparsity pattern. We examine properties of these estimators, and show that their mixtures with exponential weights satisfy sparsity oracle inequalities for suitably chosen priors π.

5.1 Sparsity Pattern Aggregation

Assume that we are given a dictionary of functions $\mathcal{H} = \{f_1, \dots, f_M\}$. However, we will not aggregate the elements of the dictionary, but rather the least squares estimators depending on all the f_j. We denote by $\mathbf{X}$, the $n \times M$ *design matrix* with elements $\mathbf{X}_{i,j} = f_j(x_i)$, $i = 1, \dots, n$, $j = 1, \dots, M$.

A *sparsity pattern* is a binary vector $\mathsf{p} \in \mathcal{P} := \{0, 1\}^M$. The terminology comes from the fact that the coordinates p_j of such vectors can be interpreted as indicators of presence ($\mathsf{p}_j = 1$) or absence ($\mathsf{p}_j = 0$) of a given feature indexed by $j \in \{1, \dots, M\}$. We denote by $|\mathsf{p}|$ the number of ones in the sparsity pattern p, and by S^{p} the linear span of canonical basis vectors $e^{(j)}$, such that $\mathsf{p}_j = 1$.

For $\mathsf{p} \in \mathcal{P}$, let $\hat{\theta}_{\mathsf{p}}$ be any least squares estimator on S^{p} defined by

$$\hat{\theta}_{\mathsf{p}} \in \operatorname*{argmin}_{\theta \in S^{\mathsf{p}}} \|\mathbf{Y} - \mathsf{f}_\theta\|^2 \quad \text{with } \mathsf{f}_\theta = \sum_{j=1}^{M} \theta_j f_j. \tag{5.1}$$

The following simple lemma gives an oracle inequality for the least squares estimator. It follows easily from the Pythagorean theorem. Moreover, the random variables $\xi_1, \dots, \xi_n$ need not be Gaussian for the result to hold.

LEMMA 5.1. *Fix $\mathsf{p} \in \mathcal{P}$. Then any least squares estimator $\hat{\theta}_{\mathsf{p}}$ defined in (5.1) satisfies*

$$\begin{aligned} \mathbb{E}\|\mathsf{f}_{\hat{\theta}_{\mathsf{p}}} - \eta\|^2 &= \min_{\theta \in S^{\mathsf{p}}} \|\mathsf{f}_\theta - \eta\|^2 + \sigma^2 \frac{d_{\mathsf{p}}}{n} \\ &\le \min_{\theta \in S^{\mathsf{p}}} \|\mathsf{f}_\theta - \eta\|^2 + \sigma^2 \frac{|\mathsf{p}|}{n}, \end{aligned} \tag{5.2}$$

where d_{p} is the dimension of the linear subspace $\{\mathbf{X}\theta : \theta \in S^{\mathsf{p}}\}$.

Clearly, if $|\mathsf{p}|$ is small compared to n, the oracle inequality gives a good performance guarantee for the least squares aggregate $\mathsf{f}_{\hat{\theta}_{\mathsf{p}}}$. Nevertheless, it may be the case that the approximation error $\min_{\theta \in S^{\mathsf{p}}} \|\mathsf{f}_\theta - \eta\|^2$ is quite large. Hence, we are looking for a sparsity pattern such that $|\mathsf{p}|$ is small and that yields a least squares aggregate with small approximation error. This is clearly a model selection problem, as described in Section 1.

Observe that for each sparsity pattern $\mathsf{p} \in \mathcal{P}$, the function $\mathsf{f}_{\hat{\theta}_{\mathsf{p}}}$ is a projection estimator of the form $\mathsf{f}_{\hat{\theta}_{\mathsf{p}}} = A_{\mathsf{p}}\mathbf{Y}$ where the $n \times n$ matrix A_{p} is the projector onto $\{\mathbf{X}\theta : \theta \in S^{\mathsf{p}}\}$ (as above, we identify the functions $f_j, \mathsf{f}_{\hat{\theta}_{\mathsf{p}}}$ with the vectors of their values at points $x_1, \dots, x_n$ since the risk depends only on these values). Therefore $\operatorname{Tr}[A_{\mathsf{p}}] = d_{\mathsf{p}}$. We have seen in the previous section that, to solve the problem of model selection, projection estimators can be aggregated using exponential weights. Thus, instead of selecting the best sparsity pattern, we resort to taking convex combinations leading to what is called *sparsity pattern aggregation*. For any sparsity pattern $\mathsf{p} \in \mathcal{P}$, define the exponential weights $\hat{\lambda}_{\mathsf{p}}^{\pi}$ and the *sparsity pattern aggregate* $\tilde{f}^{\pi}$, respectively, by

$$\begin{aligned} \hat{\lambda}_{\mathsf{p}}^{\pi} &= \frac{\exp(-n\tilde{R}_n^{\mathrm{unb}}(\mathsf{f}_{\hat{\theta}_{\mathsf{p}}})/\beta)\pi_{\mathsf{p}}}{\sum_{\mathsf{p}' \in \mathcal{P}} \exp(-n\tilde{R}_n^{\mathrm{unb}}(\mathsf{f}_{\hat{\theta}_{\mathsf{p}'}})/\beta)\pi_{\mathsf{p}'}}, \\ \tilde{f}^{\pi} &= \sum_{\mathsf{p} \in \mathcal{P}} \hat{\lambda}_{\mathsf{p}}^{\pi} \mathsf{f}_{\hat{\theta}_{\mathsf{p}}}, \end{aligned}$$

where $\pi = (\pi_{\mathsf{p}})_{\mathsf{p} \in \mathcal{P}}$ is a probability distribution (prior) on the set of sparsity patterns $\mathcal{P}$.

To study the performance of this method, we can now apply the last part of Theorem 4.1 dealing with projection matrices. Let $\mathsf{p}(\theta) \in \mathcal{P}$ be the sparsity pattern of $\theta \in \mathbb{R}^M$, that is, a vector with components $\mathsf{p}_j(\theta) = 1$ if $\theta_j \neq 0$, and $\mathsf{p}_j(\theta) = 0$ otherwise. Note that $|\mathsf{p}(\theta)| = |\theta|_0$. Combining (4.5) and Lemma 5.1 and the fact that $\{\theta : \mathsf{p}(\theta) = \mathsf{p}\} \subset S^{\mathsf{p}}$, we get that for $\beta \ge 4\sigma^2$

$$\begin{aligned} \mathbb{E}R(\tilde{f}^{\pi}) &\le \min_{\mathsf{p} \in \mathcal{P}} \biggl\{ \mathbb{E}R(\mathsf{f}_{\hat{\theta}_{\mathsf{p}}}) + \frac{\beta}{n} \log(\pi_{\mathsf{p}}^{-1}) \biggr\} \\ &\le \min_{\mathsf{p} \in \mathcal{P}} \biggl\{ \min_{\theta\,:\,\mathsf{p}(\theta) = \mathsf{p}} \|\mathsf{f}_\theta - \eta\|^2 + \sigma^2 \frac{|\mathsf{p}|}{n} \\ &\qquad + \frac{\beta}{n} \log(\pi_{\mathsf{p}}^{-1}) \biggr\} \\ &= \min_{\theta \in \mathbb{R}^M} \biggl\{ \|\mathsf{f}_\theta - \eta\|^2 + \sigma^2 \frac{|\theta|_0}{n} \\ &\qquad + \frac{\beta}{n} \log(\pi_{\mathsf{p}(\theta)}^{-1}) \biggr\}, \end{aligned} \tag{5.3}$$

where we have used that $\min_{\theta \in \mathbb{R}^M}$ can be represented as $\min_{\mathsf{p} \in \mathcal{P}} \min_{\theta\,:\,\mathsf{p}(\theta) = \mathsf{p}}$.

The remainder term in the balanced oracle inequality (5.3) depends on the choice of the prior π. Several choices can be considered depending on the information that we have about the oracle, that is, about a potentially good candidate θ that we would like to mimic. For example, we can assume that there exists a good θ that is coordinatewise sparse, group sparse or even that θ is piecewise constant. While this approach to structure the prior knowledge seems to fit in a Bayesian framework, we only pursue a frequentist setup. Indeed, our risk measure is not averaged over a prior. Such priors on good candidates for estimation are often used in a non-Bayesian framework. For example, in nonparametric estimation, it is usually assumed that a good candidate function is smooth. Without such assumptions, one may face difficulties in performing meaningful theoretical analysis.

5.2 Sparsity Priors

5.2.1 *Coordinatewise sparsity.* This is the basic and most commonly used form of sparsity. The prior π should favor vectors θ that have a small number of nonzero coordinates. Several priors have been suggested for this purpose; cf. Leung and Barron (2006); Giraud (2007); Rigollet and Tsybakov (2011); Alquier and Lounici (2011). We consider here yet another prior, close to that of Giraud (2007). The main difference

is that the prior π^{C} below exponentially downweights sparsity patterns with large $|\mathsf{p}|$, whereas the prior in Giraud (2008) downweights such patterns polynomially. Define

$$\pi_{\mathsf{p}}^{\mathrm{C}} = \left[\binom{M}{|\mathsf{p}|} e^{|\mathsf{p}|} H_M\right]^{-1}, \quad H_M = \sum_{k=0}^{M} e^{-k} \le \frac{e}{e-1}.$$

It can be easily seen that $\sum_{\mathsf{p}\in\mathcal{P}} \pi_{\mathsf{p}}^{\mathrm{C}} = 1$ so that $\pi^{\mathrm{C}} = (\pi_{\mathsf{p}}^{\mathrm{C}}, \mathsf{p} \in \mathcal{P})$ is a probability measure on $\mathcal{P}$. Note that

$$\begin{aligned} \log[(\pi_{\mathsf{p}}^{\mathrm{C}})^{-1}] &= \log\binom{M}{|\mathsf{p}|} + |\mathsf{p}| + \log(H_M) \\ &\le 2|\mathsf{p}| \log\left(\frac{eM}{|\mathsf{p}|}\right) + \frac{1}{2}, \end{aligned} \tag{5.4}$$

where we have used the inequality $\binom{M}{|\mathsf{p}|} \le (\frac{eM}{|\mathsf{p}|})^{|\mathsf{p}|}$ for $|\mathsf{p}| \neq 0$ and the convention $0\log(\infty) = 0$ for $|\mathsf{p}| = 0$. Define the sparsity pattern aggregate

$$\tilde{f}^{\mathrm{C}} = \sum_{\mathsf{p}\in\mathcal{P}} \hat{\lambda}_{\mathsf{p}}^{\pi^{\mathrm{C}}} \mathsf{f}_{\hat{\theta}_{\mathsf{p}}}, \tag{5.5}$$

where $\lambda_{\mathsf{p}}^{\pi^{\mathrm{C}}}$ is the exponential weight given in (4.3), and $\hat{\theta}_{\mathsf{p}}$ is the least squares estimator (5.1).

Plugging (5.4) into (5.3) with $\pi = \pi^{\mathrm{C}}$ and $\beta = 4\sigma^2$ yields the following sparsity oracle inequality:

$$\begin{aligned} \mathbb{E}R(\tilde{f}^{\mathrm{C}}) \le \inf_{\theta\in\mathbb{R}^M} \Bigg\{ &\|\mathsf{f}_\theta - \eta\|^2 \\ &+ \frac{9\sigma^2}{n}|\theta|_0 \log\left(\frac{eM}{|\theta|_0}\right) + \frac{2\sigma^2}{n} \Bigg\}. \end{aligned} \tag{5.6}$$

It is important to note that (5.6) is valid under *no assumption in the dictionary*. This is in contrast to the Lasso and assimilated penalized procedures that are known to have similar properties only under strong conditions on $\mathbf{X}$, such as restricted isometry or restricted eigenvalue conditions (see, e.g., Candes and Tao, 2007; Bickel, Ritov and Tsybakov, 2009; Koltchinskii, Lounici and Tsybakov, 2011).

Another choice for π in the framework of coordinatewise sparsity can be found in Rigollet and Tsybakov (2011) and yields the *exponential screening* estimator. The exponential screening aggregate satisfies an improved version of the above sparsity oracle inequality with $|\theta|_0$ replaced by $\min(|\theta|_0, R)$ where R is the rank of the design matrix $\mathbf{X}$. In particular, if the rank R is small, the exponential screening aggregate adapts to it. Moreover, it is shown in Rigollet and Tsybakov (2011) that the remainder term of the oracle inequality is optimal in a minimax sense.

5.2.2 *Fused sparsity.* When there exists a natural order among the functions $f_1, \dots, f_M$ in the dictionary, it may be appropriate to assume that there exists a "piecewise constant" $\theta \in \mathbb{R}^M$, that is, θ with components taking only a small number of values, such that f_θ has good approximation properties. This property often referred to as *fused sparsity* has been exploited in the image denoising literature for two decades, originating with the classical paper by Rudin, Osher and Fatemi (1992). The *fused Lasso* was introduced in Tibshirani et al. (2005) to deal with the same problem in one dimension instead of two. Here we suggest another method that takes advantage of fused sparsity using the idea of mixing with exponential weights. Its theoretical advantages are demonstrated by the sparsity oracle inequality in Corollary 5.1 below.

At first sight, this problem appears to be different from the one considered above since a good $\theta \in \mathbb{R}^M$ need not be sparse. Yet, the fused sparsity assumption on θ can be reformulated into a coordinatewise sparsity assumption. Indeed, let D be the $M \times M$ matrix defined by the relations $(D\theta)_1 = \theta_1$ and $(D\theta)_j = \theta_j - \theta_{j-1}$ for $j = 2, \dots, M$, where $(D\theta)_j$ is the jth component of $D\theta$. We will call D the "first differences" matrix. Then θ is fused sparse if $|D\theta|_0$ is small.

We now consider a more general setting with an arbitrary invertible matrix D, again declaring θ to be fused sparse if $|D\theta|_0$ is small. Possible definitions of D can be based on higher order differences or combinations of differences of several orders accounting for other types of sparsity. For each sparsity pattern $\mathsf{p} \in \mathcal{P}$, we define the least squares estimator

$$\hat{\theta}_{\mathsf{p}}^{D} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^M :\, D\theta\in S^{\mathsf{p}}} \|\mathbf{Y} - \mathsf{f}_\theta\|^2. \tag{5.7}$$

The corresponding estimator $\mathsf{f}_{\hat{\theta}_{\mathsf{p}}^D}$ (as previously, without loss of generality we consider $\mathsf{f}_{\hat{\theta}_{\mathsf{p}}^D}$ as an n-vector) takes the form $\mathsf{f}_{\hat{\theta}_{\mathsf{p}}^D} = A_{\mathsf{p}}^D \mathbf{Y}$ where A_{p}^D is the projector onto the linear space $\mathcal{L}_{\mathsf{p}} = \{\mathbf{X}\theta : \theta \in \mathbb{R}^M, D\theta \in S^{\mathsf{p}}\}$. In particular, $\mathrm{Tr}[A_{\mathsf{p}}^D] = \dim(\mathcal{L}_{\mathsf{p}})$, where $\dim(\mathcal{L}_{\mathsf{p}})$ is the dimension of $\mathcal{L}_{\mathsf{p}}$. Moreover, it is straightforward to obtain the following result, analogous to Lemma 5.1.

LEMMA 5.2. *Fix $\mathsf{p} \in \mathcal{P}$, and let D be an invertible matrix. Then any least squares estimator $\hat{\theta}_{\mathsf{p}}^D$ defined in (5.7) satisfies*

$$\begin{aligned} \mathbb{E}\|\mathsf{f}_{\hat{\theta}_{\mathsf{p}}^D} - \eta\|^2 &= \min_{\substack{\theta\in\mathbb{R}^M: \\ D\theta\in S^{\mathsf{p}}}} \|\mathsf{f}_\theta - \eta\|^2 + \sigma^2 \frac{\dim(\mathcal{L}_{\mathsf{p}})}{n} \\ &\le \min_{\substack{\theta\in\mathbb{R}^M: \\ D\theta\in S^{\mathsf{p}}}} \|\mathsf{f}_\theta - \eta\|^2 + \sigma^2 \frac{|\mathsf{p}|}{n}. \end{aligned} \tag{5.8}$$

We are therefore in a position to apply the results from Section 4. For example, if p is sparse, and D is the "first differences" matrix, the least squares estimator $\hat{\theta}^D_{\mathsf{p}}$ is piecewise constant with a small number $|\mathsf{p}|$ of jumps.

Now, since the problem has been reduced to coordinatewise sparsity, we can choose the prior π^{C} to favor vectors $\theta \in \mathbb{R}^M$ that are piecewise constant with a small number of jumps. Define the fused sparsity pattern aggregate $\tilde{f}^{\mathrm{F}}$ by

$$\tilde{f}^{\mathrm{F}} = \sum_{\mathsf{p}\in\mathcal{P}} \hat{\lambda}^{\pi^{\mathrm{C}}}_{\mathsf{p}} \mathsf{f}_{\hat{\theta}^D_{\mathsf{p}}}, \tag{5.9}$$

where $\hat{\lambda}^{\pi^{\mathrm{C}}}_{\mathsf{p}}$ is the exponential weight defined in (4.3), and $\hat{\theta}^D_{\mathsf{p}}$ is the least squares estimator defined in (5.7). Note that we can combine (4.5) with Lemma 5.2 in the same way as in (5.3) with the only difference that we use now the relation $\min_{\mathsf{p}\in\mathcal{P}} \min_{\theta\,:\,\mathsf{p}(D\theta)=\mathsf{p}}(\cdot) = \min_{\theta\,:\,D\theta\in\mathbb{R}^M}(\cdot) = \min_{\theta\in\mathbb{R}^M}(\cdot)$. This and (5.4) imply the following bound.

COROLLARY 5.1. *Let D be an invertible matrix. The fused sparsity pattern aggregate $\tilde{f}^{\mathrm{F}}$ defined in (5.9) with $\beta = 4\sigma^2$ satisfies*

$$\begin{aligned}\mathbb{E}R(\tilde{f}^{\mathrm{F}}) &\le \inf_{\theta\in\mathbb{R}^M}\biggl\{\|\mathsf{f}_\theta - \eta\|^2 \\ &\qquad + \frac{9\sigma^2}{n}|D\theta|_0 \log\biggl(\frac{eM}{|D\theta|_0}\biggr) + \frac{2\sigma^2}{n}\biggr\}.\end{aligned} \tag{5.10}$$

To our knowledge, analogous bounds for fused Lasso are not available. Furthermore, Corollary 5.1 holds under no assumption on the matrix $\mathbf{X}$, which cannot be the case for the Lasso type methods. Let us also emphasize that Corollary 5.1 is valid for any invertible matrix D, and not only for the standard "first differences" matrix D defined above.

5.2.3 *Group sparsity.* Since recently, estimation under group sparsity has been intensively discussed in the literature. Starting from Yuan and Lin (2006), several estimators have been studied, essentially the Group Lasso and some related penalized techniques. Theoretical properties of the Group Lasso are treated in some generality by Huang and Zhang (2010) and Lounici et al. (2011) where one can find further references. Here we show that one can deal with group sparsity using exponentially weighted aggregates. The new estimator that we propose presents some theoretical advantages as compared to the Group Lasso type methods.

Let $B_1, \dots, B_K$ be given subsets of $\{1, \dots, M\}$ called the groups. We impose no restriction on B_j's; for example, they need not form a partition of $\{1, \dots, M\}$ and can overlap. In this section, we consider $\theta \in \mathbb{R}^M$ such that $\operatorname{supp}(\theta) \subseteq B \triangleq \bigcup_{k=1}^K B_k$ where $\operatorname{supp}(\theta)$ is the support of θ. For any such θ, we denote by $J(\theta)$ the subset of $\{1, \dots, K\}$ of smallest cardinality among all J satisfying $\operatorname{supp}(\theta) \subseteq B_J \triangleq \bigcup_{k\in J} B_k$. We assume without loss of generality that $J(\theta)$ is unique. (If there are several J of same cardinality satisfying this property, we define $J(\theta)$ as the smallest among them with respect to some partial ordering of the subsets of $\{1, \dots, K\}$.) Set

$$g(\theta) = |J(\theta)|, \quad B(\theta) = \bigcup_{k\in J(\theta)} B_k.$$

The group sparsity setup assumes that there exists $\theta \in \mathbb{R}^M$ such that $\|\mathsf{f}_\theta - \eta\|^2$ is small and that θ is supported by a small number of groups, that is, that $g(\theta) \ll K$.

Let now J be a subset $\{1, \dots, K\}$. Denote by p^J the sparsity pattern with coordinates defined by

$$\mathsf{p}^J_j = \begin{cases} 1, & \text{if } j \in B_J, \\ 0, & \text{otherwise,}\end{cases}$$

for $j = 1, \dots, M$. Consider the set of all such sparsity patterns:

$$\mathcal{P}_{\mathrm{G}} = \{\mathsf{p}^J, J \subseteq \{1, \dots, K\}\}.$$

To each sparsity pattern $\mathsf{p}^J \in \mathcal{P}_{\mathrm{G}}$ we assign a least squares estimator $\hat{\theta}_{\mathsf{p}^J}$, cf. (5.1), constrained to having null coordinates outside of $B_J = \bigcup_{k\in J} B_k$.

Define the following prior on $\mathcal{P}_{\mathrm{G}}$:

$$\pi^{\mathrm{G}}_{\mathsf{p}^J} = \biggl[\binom{K}{|J|} e^{|J|} H_K\biggr]^{-1}, \quad J \subseteq \{1, \dots, K\}.$$

This prior enforces group sparsity by favoring the small number of groups $|J|$. As in (5.4), we obtain

$$\log[(\pi^{\mathrm{G}}_{\mathsf{p}^J})^{-1}] \le 2|J| \log\biggl(\frac{eK}{|J|}\biggr) + \frac{1}{2}. \tag{5.11}$$

We introduce now the sparsity pattern aggregate

$$\tilde{f}^{\mathrm{G}} = \sum_{\mathsf{p}\in\mathcal{P}_{\mathrm{G}}} \hat{\lambda}^{\pi^{\mathrm{G}}}_{\mathsf{p}} \mathsf{f}_{\hat{\theta}_{\mathsf{p}}}, \tag{5.12}$$

where $\lambda^{\pi^{\mathrm{G}}}_{\mathsf{p}}$ is the exponential weight defined in (4.3) and $\hat{\theta}_{\mathsf{p}}$ is the least squares estimator defined in (5.1).

For any $\mathsf{p} = \mathsf{p}^J \in \mathcal{P}_{\mathrm{G}}$, we have $V_J \triangleq \{\theta : \operatorname{supp}(\theta) \subseteq B, J(\theta) = J\} \subseteq S^{\mathsf{p}}$. Arguing as in (5.3) with $\mathcal{P}_{\mathrm{G}}$ instead of $\mathcal{P}$, setting $\pi = \pi^{\mathrm{G}}$ and using (5.11) we obtain

that, for $\beta \geq 4\sigma^2$,

$$\begin{aligned}
&\mathbb{E}R(\tilde{f}^{\mathrm{G}}) \\
&\quad \leq \min_{\mathsf{p}\in\mathcal{P}_{\mathrm{G}}} \Bigg\{ \min_{\theta\in S^{\mathsf{p}}} \|\mathsf{f}_\theta - \eta\|^2 + \sigma^2 \frac{|\mathsf{p}|}{n} \\
&\qquad\qquad + \frac{\beta}{n} \log((\pi_{\mathsf{p}}^{\mathrm{G}})^{-1}) \Bigg\} \\
&\quad \leq \min_{J\subseteq\{1,\dots,K\}} \Bigg\{ \min_{\theta\in V_J} \|\mathsf{f}_\theta - \eta\|^2 + \sigma^2 \frac{|B_J|}{n} \\
&\qquad\qquad + \frac{\beta}{n} \log((\pi_{\mathsf{p}^J}^{\mathrm{G}})^{-1}) \Bigg\} \\
&\quad \leq \min_{J\subseteq\{1,\dots,K\}} \min_{\theta\in V_J} \Bigg\{ \|\mathsf{f}_\theta - \eta\|^2 + \sigma^2 \frac{|B(\theta)|}{n} \\
&\qquad\qquad + \frac{2\beta}{n} g(\theta) \log\biggl(\frac{eK}{g(\theta)}\biggr) + \frac{\beta}{2n} \Bigg\}.
\end{aligned}$$

This leads to the following oracle inequality.

COROLLARY 5.2. *The group sparsity pattern aggregate* $\tilde{f}^{\mathrm{G}}$ *defined in* (5.12) *with* $\beta = 4\sigma^2$ *satisfies*

$$\begin{aligned}
\mathbb{E}R(\tilde{f}^{\mathrm{G}}) \leq \inf_{\substack{\theta\in\mathbb{R}^M: \\ \operatorname{supp}(\theta)\subseteq B}} \Bigg\{ &\|\mathsf{f}_\theta - \eta\|^2 + \sigma^2 \frac{|B(\theta)| + 2}{n} \\
&+ \frac{8\sigma^2}{n} g(\theta) \log\biggl(\frac{eK}{g(\theta)}\biggr) \Bigg\}.
\end{aligned} \tag{5.13}$$

We see from Corollary 5.2 that if there exists an ideal "oracle" θ in $\mathbb{R}^M$, such that the approximation error $\|\mathsf{f}_\theta - \eta\|^2$ is small, and θ is sparse in the sense that it is supported by a small number of groups, then the sparsity pattern aggregate $\tilde{f}^{\mathrm{G}}$ mimics the risk of this oracle.

A remarkable fact is that Corollary 5.2 holds for arbitrary choice of groups B_j. They can overlap and not necessarily cover the whole set $\{1,\dots,M\}$.

To illustrate the power of the oracle inequality (5.13), we consider the multi-task learning setup as in Lounici et al. (2011). Namely, assume that all the groups B_j are of the same size T and form a partition of $\{1,\dots,M\}$, so that $M = KT$. We restrict our analysis to the class $\mathcal{F}_s$ of regression functions η such that $\eta = \mathsf{f}_\theta$ for some θ satisfying $g(\theta) \leq s$ where $s \leq K$ is a given integer. Then $|B(\theta)| \leq sT$. Combining these remarks with (5.13) and with the fact that the function $x \mapsto x\log(\frac{eK}{x})$ is increasing, we find that, uniformly over $\eta \in \mathcal{F}_s$,

$$\mathbb{E}R(\tilde{f}^{\mathrm{G}}) \leq \frac{\sigma^2 s}{n}\biggl(T + 8\log\biggl(\frac{eK}{s}\biggr) + \frac{2}{s}\biggr). \tag{5.14}$$

On the other hand, a minimax lower bound on the same class $\mathcal{F}_s$ is available in Lounici et al. (2011). It has exactly the form of the right-hand side of (5.14); cf. equation (6.2) in Lounici et al. (2011). This immediately implies that (i) the lower bound of Lounici et al. (2011) is tight so that $\frac{s}{n}(T + \log(\frac{K}{s}))$ is the optimal rate of convergence on $\mathcal{F}_s$, and (ii) the estimator $\tilde{f}^{\mathrm{G}}$ is rate optimal. To our knowledge, this gives the first example of rate optimal estimator under group sparsity. The upper bounds for the Group Lasso estimators in Huang and Zhang (2010) and Lounici et al. (2011) as well as in the earlier papers cited therein depart from this optimal rate at least by a logarithmic factor. Furthermore, they are obtained under strong assumptions on the dictionary such as restricted isometry or restricted eigenvalue type conditions, while (5.14) is valid under no assumption on the dictionary.

6. RELATED PROBLEMS

In this paper, we have considered only the Gaussian regression model with fixed design and known variance of the noise. This is a basic setup where the sharpest results, expressed in terms of sparsity oracle inequalities, are now available for exponentially weighted (EW) procedures both in aggregation and sparsity scenarios. Similar but somewhat weaker properties are obtained for exponential weighting in several other models.

Models with i.i.d. observations. Some EW aggregates achieve sparsity oracle inequalities in regression model with random design (Dalalyan and Tsybakov, 2012; Alquier and Lounici, 2011; Gerchinovitz, 2011) as well as in density estimation and classification problems (Dalalyan and Tsybakov, 2012). However, the results differ in several aspects from those of the present paper. First, they do not use aggregation of estimators, but rather EW procedures driven by continuous priors (Dalalyan and Tsybakov, 2012; Gerchinovitz, 2011), or by priors with both discrete and continuous components (Alquier and Lounici, 2011). The developments in Dalalyan and Tsybakov (2012); Alquier and Lounici (2011); Gerchinovitz (2011) start from the general oracle inequalities similar to (3.2), which are sometimes called PAC-bounds; cf. recent overview in Catoni (2007). Sparsity oracle inequalities are then derived from PAC-bounds. However, as opposed to (5.6), they involve not only $|\theta|_0$ but also the ℓ_1-norm of θ. The estimators in Dalalyan and Tsybakov (2012); Gerchinovitz (2011) are defined as an average of exponentially weighted aggregates over the sample sizes from 1 to n. This is related to earlier work on mirror

averaging; cf. Juditsky et al. (2005); Juditsky, Rigollet and Tsybakov (2008), which in turn, is inspired by the concept of mirror descent in optimization due to Nemirovski. Finally, the computational algorithms are also quite different from those that we describe in the next section. For example, under continuous sparsity priors, one of the suggestions is to use Langevin Monte-Carlo; cf. Dalalyan and Tsybakov (2012, 2012).

Unknown variance of the noise, non-Gaussian noise. Modifications of EW procedures and of the corresponding oracle inequalities for the case of unknown variance σ^2 are discussed in Giraud (2007); Gerchinovitz (2011). Moreover, the results can be extended to regression with non-Gaussian noise under deterministic or random design (Dalalyan and Tsybakov, 2007; Dalalyan and Tsybakov, 2008; Dalalyan and Tsybakov, 2012; Gerchinovitz, 2011). In particular, Gerchinovitz, 2011 uses a version of the EW estimator with data-driven truncation to cover a rather general noise structure. The estimator satisfies a balanced oracle inequality but not a sparsity oracle inequality as defined here, since along with $|\theta|_0$, it involves other characteristics of θ and of the target function η.

7. NUMERICAL IMPLEMENTATION

All the sparsity pattern aggregates defined in the previous section are of the form $\mathsf{f}_{\theta^{\exp}}$, where

$$\theta^{\exp} = \sum_{\mathsf{p}\in\mathcal{G}} \hat{\lambda}^{\pi}_{\mathsf{p}} \bar{\theta}_{\mathsf{p}} \tag{7.1}$$

for some $\mathcal{G} \subset \mathcal{P}$, $\lambda^{\pi}_{\mathsf{p}}$ is the exponential weight defined in (4.3), and $\bar{\theta}_{\mathsf{p}}$ is either $\hat{\theta}_{\mathsf{p}}$ defined in (5.1) or $\hat{\theta}^D_{\mathsf{p}}$ defined in (5.7).

From (7.1), it is clear that one needs to add up with some weights 2^M (or 2^K in the case of group sparsity with K groups) least squares estimators to compute $\theta^{\exp}$ *exactly*. In many applications this number is prohibitively large. However, most of the terms in the sum receive an exponentially low weight with the choices of π that we have described. We resort to a numerical approximation that exploits this fact.

Note that $\theta^{\exp}$ is obtained as the expectation of the random variable $\hat{\theta}_{\mathsf{P}}$ or $\hat{\theta}^D_{\mathsf{P}}$ where P is a random variable taking values in $\mathcal{P}$ with probability distribution ν given by

$$\nu_{\mathsf{p}} = \frac{\exp(-n\tilde{R}^{\mathrm{unb}}_n(\mathsf{f}_{\bar{\theta}_{\mathsf{p}}})/\beta)\pi_{\mathsf{p}}}{\sum_{\mathsf{p}'\in\mathcal{G}} \exp(-n\tilde{R}^{\mathrm{unb}}_n(\mathsf{f}_{\bar{\theta}_{\mathsf{p}'}})/\beta)\pi_{\mathsf{p}'}}, \quad \mathsf{p}\in\mathcal{G}\subset\mathcal{P}.$$

This Gibbs-type distribution can be expressed as the stationary distribution of the Markov chain generated by the Metropolis–Hastings (MH) algorithm (see, e.g., Robert and Casella, 2004, Section 7.3). We now describe the MH algorithm employed here. Note that in the examples considered in the previous section, $\mathcal{G}$ is either the hypercube $\mathcal{P}$ or the hypercube $\mathcal{P}_{\mathrm{G}}$. For any $\mathsf{p}\in\mathcal{G}$, define the instrumental distribution $q(\cdot|\mathsf{p})$ as the uniform distribution on the neighbors of p in $\mathcal{G}$, and notice that since each vertex has the same number of neighbors, we have $q(\mathsf{p}|\mathsf{q}) = q(\mathsf{q}|\mathsf{p})$ for any $\mathsf{p},\mathsf{q}\in\mathcal{P}$. The MH algorithm is defined in Figure 1. We use here the uniform instrumental distribution for the sake of simplicity. Our simulations show that it yields satisfactory results both in terms of performance and of speed. Another choice of $q(\cdot|\cdot)$ can potentially further accelerate the convergence of the MH algorithm.

Fix $\mathsf{p}_0 = 0 \in \mathbb{R}^M$. For any $t \ge 0$, given $\mathsf{p}_t \in \mathcal{G}$:

(1) Generate a random variable Q_t with distribution $q(\cdot|\mathsf{p}_t)$.

(2) Generate a random variable

$$\mathsf{P}_{t+1} = \begin{cases} \mathsf{Q}_t, & \text{with probability } r(\mathsf{p}_t,\mathsf{Q}_t), \\ \mathsf{p}_t, & \text{with probability } 1-r(\mathsf{p}_t,\mathsf{Q}_t), \end{cases}$$

where

$$r(\mathsf{p},\mathsf{q}) = \min\left(\frac{\nu_{\mathsf{q}}}{\nu_{\mathsf{p}}}, 1\right).$$

(3) Compute the least squares estimator $\bar{\theta}_{\mathsf{P}_{t+1}}$.

FIG. 1. *The Metropolis–Hastings algorithm on the M-hypercube.*

From the results of Robert and Casella (2004) (see also Rigollet and Tsybakov, 2011, Theorem 7.1) the Markov chain $(\mathsf{P}_t)_{t\ge0}$ defined in Figure 1 is ergodic. In other words, it holds

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \bar{\theta}_{\mathsf{P}_t} = \sum_{\mathsf{p}\in\mathcal{G}} \bar{\theta}_{\mathsf{p}}\nu_{\mathsf{p}}, \quad \text{almost surely,}$$

where $T_0 \ge 0$ is an arbitrary integer.

In view of this result, we approximate $\theta^{\exp} = \sum_{\mathsf{p}\in\mathcal{G}} \bar{\theta}_{\mathsf{p}}\nu_{\mathsf{p}}$ by

$$\tilde{\theta}^{\exp}_T = \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \bar{\theta}_{\mathsf{P}_t},$$

which is close to $\theta^{\exp}$ for sufficiently large T. One remarkable feature of the MH algorithm is that it involves only the ratios $\nu_{\mathsf{q}}/\nu_{\mathsf{p}}$ where p and q are two neighbors in $\mathcal{G}$. Such ratios are easy to compute, at least in the

examples given in the previous section. As a result, the MH algorithm in this case takes the form of a stochastic greedy algorithm with averaging, which measures a trade-off between sparsity and prediction to decide whether to add or remove a variable. In all subsequent examples, we use a pure R implementation of the sparsity pattern aggregates. While the benchmark estimators considered below employ a C based code optimized for speed, we observed that a safe implementation of the MH algorithm (three times more iterations than needed) exhibits an increase of computation time of at most a factor two.

7.1 Numerical Experiments

The aim of this subsection is to illustrate the performance of the sparsity pattern aggregates $\tilde{f}^{\mathrm{C}}$ and $\tilde{f}^{\mathrm{F}}$ defined in (5.5) and (5.9) respectively, on a simulated dataset and to compare it with state-of-the-art procedures in sparse estimation. In our implementation, we replace the prior π^{C} by the *exponential screening* prior employed in Rigollet and Tsybakov (2011). As a result, the following results are about the exponential screening (ES) aggregate defined in Rigollet and Tsybakov (2011). Nevertheless, it presents the same qualitative behavior as the aggregates constructed above.

While our results for the ES estimator hold under no assumption on the dictionary, we compare the behavior of our algorithm in a well-known example where sparse estimation by ℓ_1-penalized techniques is theoretically achievable.

Consider the model $\mathbf{Y} = \mathbf{X}\theta^* + \sigma\xi$, where $\mathbf{X}$ is an $n \times M$ matrix with independent standard Gaussian entries, and $\xi \in \mathbb{R}^n$ is a vector of independent standard Gaussian random variables and is independent of $\mathbf{X}$. Depending on our sparsity assumption, we choose two different θ^*.

The variance is chosen as $\sigma^2 = \|\mathsf{f}_{\theta^*}\|^2/9 = |\mathbf{X}\theta^*|_2^2/(9n)$ following the numerical experiments of Candes and Tao [(2007), Section 4]. Here $|\cdot|_2$ denotes the ℓ_2 norm. For different values of (n, M, S), we run the ES algorithm on 500 replications of the problem and compare our results with several other popular estimators in the literature on sparse estimation that are readily implemented in R. The considered estimators are:

(1) the Lasso estimator with regularization parameter obtained by ten-fold cross-validation;

(2) the MC+ estimator of Zhang (2010) with regularization parameter obtained by ten-fold cross-validation;

(3) the SCAD estimator of Fan and Li (2001) with regularization parameter obtained by ten-fold cross-validation.

The Lasso estimator is calculated using the `glmnet` package in R (Friedman, Hastie and Tibshirani, 2010). The cross-validated MC+ and SCAD estimators are implemented in the `ncvreg` package in R (Breheny and Huang, 2011).

The performance of each of the four estimators, generically denoted by $\hat{\theta}$ is measured by its prediction error $|\mathbf{X}(\hat{\theta} - \theta^*)|_2^2/n = \|\mathsf{f}_{\hat{\theta}} - \mathsf{f}_{\theta^*}\|^2$. Moreover, even though the estimation error $|\hat{\theta} - \theta^*|_2^2$ is not studied above, we also report its values for a better comparison with other simulation studies.

7.1.1 *Coordinatewise sparsity.* The vector θ^* is given by $\theta_j^* = \mathbb{1}(j \le S)$ for some fixed S so that $|\theta^*|_0 = S$. Here, $\mathbb{1}(\cdot)$ denotes the indicator function.

We considered the cases $(n, M, S) \in \{(100, 200, 10), (200, 500, 20)\}$. The Metropolis approximation $\tilde{\theta}_T^{\exp}$ was computed with $T_0 = 3000, T = 7000$, which should be in the asymptotic regime of the Markov chain since Figure 2 shows that, on a typical example, the right sparsity pattern is recovered after about 2000 iterations.

Figure 3 displays comparative boxplots, and Table 1 reports averages and standard deviations over the 500 repetitions. In particular, it shows that ES outperforms the Lasso estimator and has performance similar to MC+ and SCAD.

Figure 2 illustrates a typical behavior of the ES estimator for one particular realization of $\mathbf{X}$ and ξ. For better visibility, both displays represent only the 50 first coordinates of $\tilde{\theta}_T^{\exp}$, with $T_0 = 3000, T = 7000$. The left-hand side display shows that the sparsity pattern is well recovered and the estimated values are close to one. The right-hand side display illustrates the evolution of the intermediate parameter $\hat{\theta}_{\mathsf{P}_t}$ for $t = 1, \ldots, 5000$. It is clear that the Markov chain that runs on the M-hypercube graph gets "trapped" in the vertex that corresponds to the sparsity pattern of θ^* after only 2000 iterations. As a result, while the ES estimator is not sparse itself, the MH approximation to the ES estimator may output a sparse solution.

FUSED SPARSITY. The vector θ^* is chosen piecewise constant as follows. Fix an integer $S \ge 1$ such that $10S \le M$ and consider the blocks $I_1, \ldots, I_S$ defined by

$$I_j = \{10(j-1) + 1, \ldots, 10j\}, \quad j = 1, \ldots, S.$$

The vector θ^* is defined to take value $(-1)^j$ on I_j, $j = 1, \ldots, S$ and $1/2$ elsewhere. We considered the cases

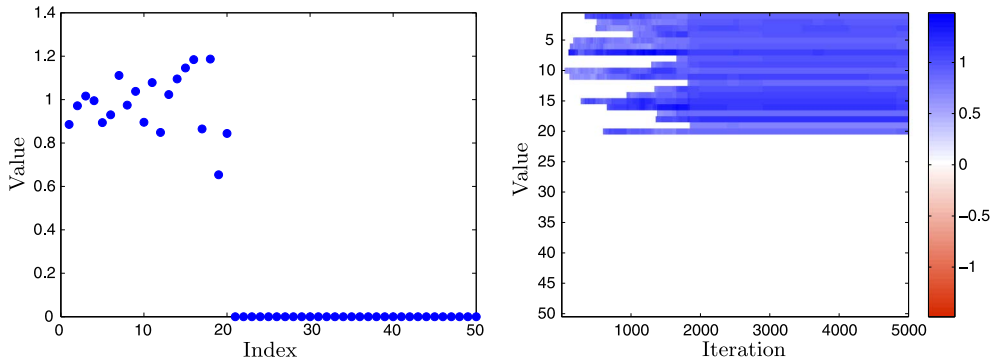

FIG. 2. *Typical realization for* $(M, n, S) = (500, 200, 20)$. *Left: Value of the* $\tilde{\theta}_T^{\exp}$, $T_0 = 3000, T = 7000$. *Right: Value of* $\hat{\theta}_{\mathsf{P}_t}$ *for* $t = 1, \ldots, 5000$. *Only the first* 50 *coordinates are shown for each vector.*

$(M, n, S) \in \{(200, 100, 10), (500, 200, 20)\}$ that are illustrated in Figure 4. Note that in both cases, the vector θ^* is *not* sparse.

The fused versions of Lasso, MC+ and SCAD are not readily available in R, and we implement them as follows. Recall that D is the $M \times M$ matrix defined in Section 5.2 by $(D\theta)_1 = \theta_1$ and $(D\theta)_j = \theta_j - \theta_{j-1}$ for $j = 2, \ldots, M$. The inverse D^{-1} is the $M \times M$ lower triangular matrix with ones on the diagonal and in the lower triangle. To obtain the fused versions of Lasso, MC+ and SCAD, we simply run these algorithms on the design matrix $\mathbf{X}D^{-1}$ to obtain a solution $\hat{\theta}$. We then

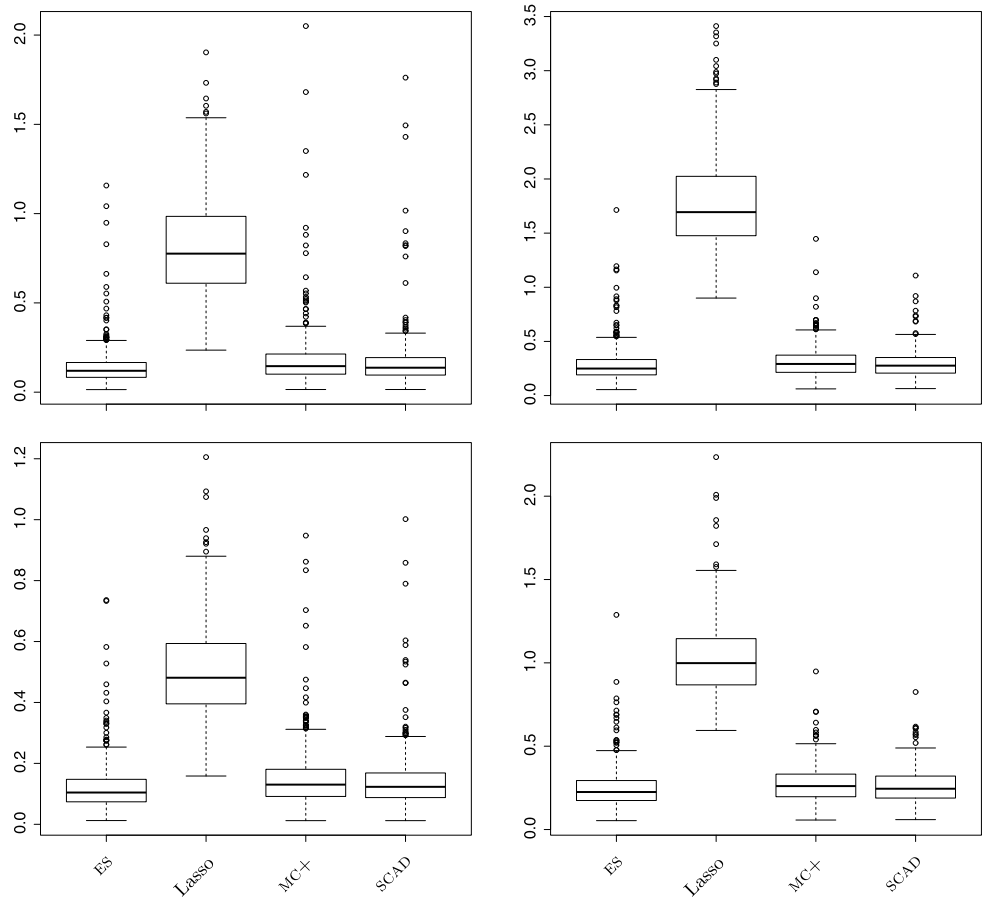

FIG. 3. *Boxplots of performance measure over* 500 *realizations for the* ES, *Lasso*, MC+ *and* SCAD *estimators. Top: estimation performance* $|\hat{\theta} - \theta^*|_2^2$. *Bottom: Prediction performance*: $|\mathbf{X}(\hat{\theta} - \theta^*)|_2^2/n$. *Left:* $(M, n, S) = (200, 100, 10)$. *Right:* $(M, n, S) = (500, 200, 20)$.

TABLE 1
Means and standard deviations of performance measures over 500 *realizations for the* ES, *Lasso*, MC+ *and* SCAD *estimators. Top: estimation performance* $|\hat{\theta} - \theta^*|_2^2$. *Bottom: Prediction performance*: $|\mathbf{X}(\hat{\theta} - \theta^*)|_2^2/n$

(M, n, S)	ES	Lasso	MC+	SCAD
(200, 100, 10)	0.14	0.82	0.18	0.17
	(0.11)	(0.28)	(0.17)	(0.15)
(500, 200, 20)	0.29	1.78	0.31	0.29
	(0.16)	(0.43)	(0.14)	(0.12)
(200, 100, 10)	0.12	0.50	0.15	0.14
	(0.08)	(0.15)	(0.10)	(0.10)
(500, 200, 20)	0.25	1.02	0.27	0.26
	(0.11)	(0.22)	(0.11)	(0.10)

return the vector $D^{-1}\hat{\theta}$ as a solution to the fused problem.

We report the boxplots of the two performance measures $|\mathbf{X}(\hat{\theta} - \theta^*)|_2^2/n$ and $|\hat{\theta} - \theta^*|_2^2$ in Figure 5. It is clear that, in this example, Exponential Screening outperforms the three other estimators. Moreover, MC+ and SCAD perform particularly poorly in the case $(M, n, S) = (500, 200, 20)$. Their output on a typical example is illustrated in Figure 4. We can see that they yield an estimator that takes only two values, thus missing most of the structure of the problem. It seems that this behavior can be explained by the fact that the estimators are trapped in a local minimum close to zero.

APPENDIX

The proof of (2.2) is standard, and similar results have been formulated in the literature for various other setups. We give it here for the sake of completeness. From the definition of the empirical risk minimizer $\hat{f}^{\mathrm{ERM}}$, we have

$$\hat{R}_n(\hat{f}^{\mathrm{ERM}}) \le \hat{R}_n(f^*),$$

where f^* is any minimizer of the true risk $R(\cdot)$ over $\mathcal{H}$. Simple algebra yields

$$R(\hat{f}^{\mathrm{ERM}}) \le R(f^*) + 2\mathbb{E}\langle \hat{f}^{\mathrm{ERM}} - f^*, \mathbf{Y} - \eta\rangle,$$

where for two functions f, g from $\mathcal{X}$ to $\mathbb{R}$ we set $\langle f, g\rangle = \frac{1}{n}\sum_{i=1}^n f(x_i)g(x_i)$. Next, observe that

$$\mathbb{E}\langle \hat{f}^{\mathrm{ERM}} - f^*, \mathbf{Y} - \eta\rangle \le \mathbb{E}\max_{f\in\mathcal{H}}\langle f - f^*, \mathbf{Y} - \eta\rangle \le 2\sigma\sqrt{\frac{2\log M}{n}},$$

where we used the fact that $\|f^* - f\| \le 2$ for any $f \in \mathcal{H}$, and the inequality $\mathbb{E}[\max_{1\le i\le M} a_i^\top \xi] \le \sigma \cdot \sqrt{2\log M}$ valid for any $a_1, \dots, a_n \in \mathbb{R}^n$, $|a_i|_2 \le 1$, where $\xi = (\xi_1, \dots, \xi_n)^\top$.

We now turn to the proof of (2.4). Consider the random matrix $\mathbb{X}$ of size $n \times M$ such that its elements $\mathbb{X}_{i,j}, i = 1, \dots, n, j = 1, \dots, M$ are i.i.d. Rademacher random variables, that is, random variables taking values 1 and -1 with probability $1/2$. Moreover, assume that

$$\frac{2}{n}\log\left(1 + \frac{eM}{2}\right) < C_1, \tag{8.1}$$

for some positive constant $C_1 < 1/2$. Note that (8.1) follows from (2.3) if C_0 is chosen small enough. Theorem 5.2 in Baraniuk et al. (2008) (see also Section 5.2.1 in Rigollet and Tsybakov, 2011) entails that if (8.1) holds for C_1 small enough, then there exists a nonempty set $\mathcal{M}$ of matrices obtained as realizations of the matrix $\mathbb{X}$ that enjoy the following weak restricted isometry (WRI) property. For any $X \in \mathcal{M}$, there exists constants $\underline{\kappa} \ge \bar{\kappa} > 0$, such that for any $\lambda \in \mathbb{R}^M$ with at most 2 nonzero coordinates,

$$\underline{\kappa}^2|\lambda|_2^2 \le \frac{|X\lambda|_2^2}{n} \le \bar{\kappa}^2|\lambda|_2^2, \tag{8.2}$$

when (8.1) is satisfied. For $X \in \mathcal{M}$, let $\phi_1, \dots, \phi_M$ be any functions on $\mathcal{X}$ satisfying

$$\phi_j(x_i) = X_{i,j}, \quad i = 1, \dots, n, j = 1, \dots, M,$$

where $X_{i,j}$ are the entries of X. Note that $\|\phi_j\| = 1$ since $X_{i,j} \in \{-1, 1\}$.

Fix $\tau > 0$ to be chosen later, and set

$$f_j = \tau(1 + \alpha)\phi_j, \quad j = 1, \dots, M,$$

where we set for brevity $\alpha = (\sigma/3)\sqrt{\frac{\log M}{\bar{\kappa}^2 n}}$. Moreover, consider the functions

$$\eta_j = \tau\alpha\phi_j, \quad j = 1, \dots, M.$$

Using (2.3) we choose τ small enough to ensure that $\|\eta_j\| \le 1$ and $\|f_j\| \le 1$ for any $j = 1, \dots, M$.

We write $R_j(\cdot)$ to denote the risk function $R(\cdot)$ when $\eta = \eta_j$ in (1.1). It is easy to check that

$$\min_{f\in\mathcal{H}} R_j(f) = R_j(f_j) = \|f_j - \eta_j\|^2. \tag{8.3}$$

As it is customary in the proof of minimax lower bounds, we reduce our estimation problem to a testing problem as follows. Let $\psi \in \{1, \dots, M\}$ be the random variable, or *test*, defined by $\psi = j$ if and only if

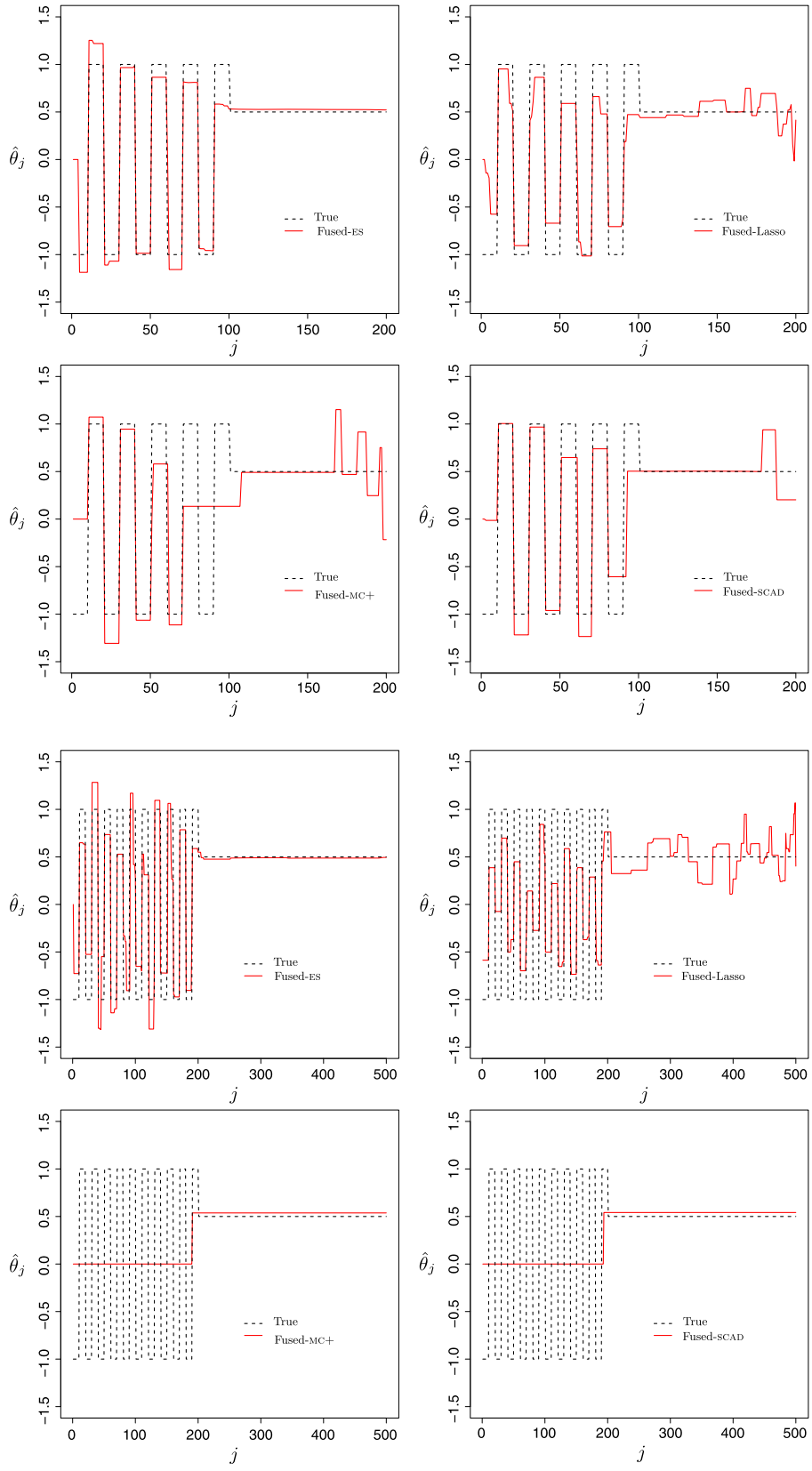

FIG. 4. *Typical realizations of the fused estimators in the cases* $(M, n, S) = (200, 100, 10)$ *(top) and* $(M, n, S) = (500, 200, 20)$ *(bottom).*

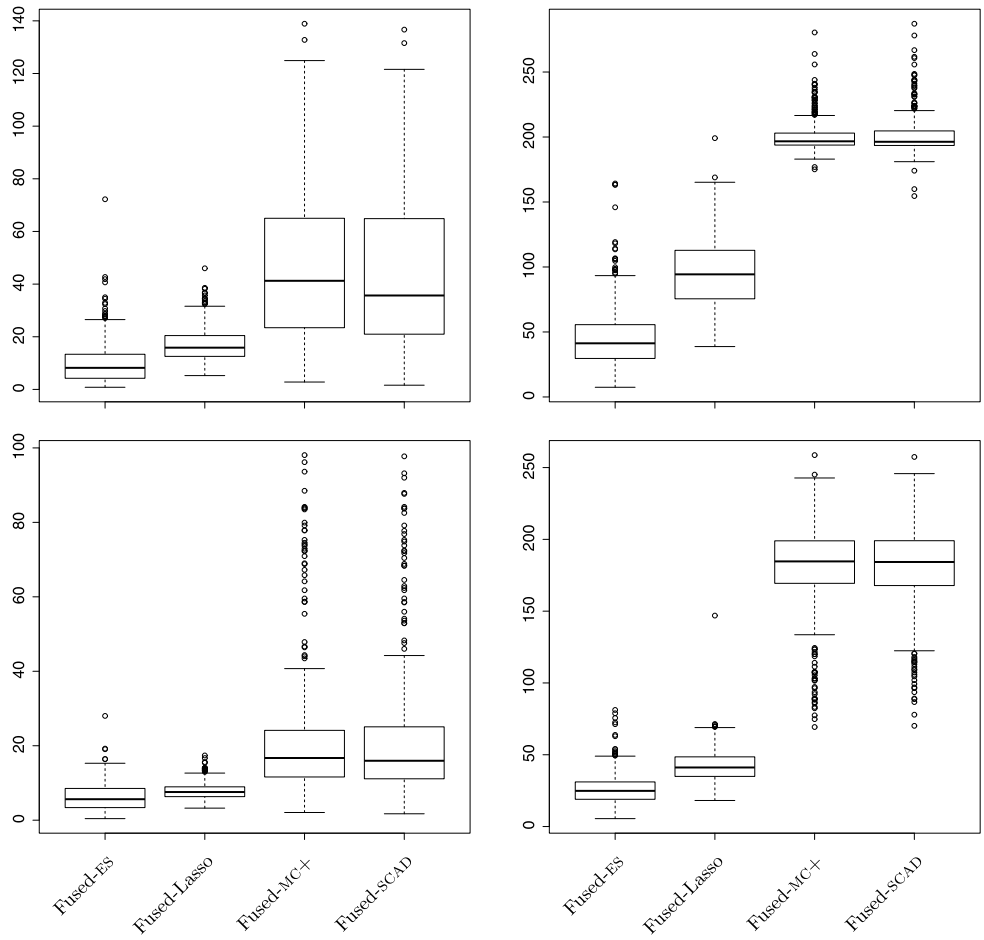

FIG. 5. *Boxplots of performance measure over* 500 *realizations for the Fused-*ES, *Fused-Lasso*, *Fused-*MC+ *and Fused-*SCAD *estimators. Top: estimation performance* $|\hat{\theta} - \theta^*|_2^2$. *Bottom: Prediction performance*: $|\mathbf{X}(\hat{\theta} - \theta^*)|_2^2/n$. *Left:* $(M, n, S) = (200, 100, 10)$. *Right:* $(M, n, S) = (500, 200, 20)$.

$\hat{S}_n = f_j$. Then, $\psi \neq j$ implies that there exists $k \neq j$ such that $\hat{S}_n = f_k$, so that

$$\begin{aligned}
&\|\hat{S}_n - \eta_j\|^2 - \|f_j - \eta_j\|^2 \\
&\quad = \|f_k - f_j\|^2 + 2\langle f_k - f_j, f_j - \eta_j\rangle \\
&\quad = \tau^2(1+\alpha)^2\|\phi_j - \phi_k\|^2 \\
&\qquad + 2\tau^2(1+\alpha)(\langle\phi_j, \phi_k\rangle - 1) \\
&\quad \geq \tau^2\alpha\|\phi_j - \phi_k\|^2.
\end{aligned}$$

From (8.2), we find that $\|\phi_j - \phi_k\|^2 \geq 2\underline{\kappa}^2$ so that

$$\|\hat{S}_n - \eta_j\|^2 - \|f_j - \eta_j\|^2 \geq \frac{2\tau^2\underline{\kappa}^2\sigma}{3\bar{\kappa}}\sqrt{\frac{\log M}{n}} \triangleq \nu_{n,M}.$$

Therefore, we conclude that $\psi \neq j$ implies that

$$R_j(\hat{S}_n) - \min_{f\in\mathcal{H}} R_j(f) \geq \nu_{n,M}.$$

Hence,

$$\begin{aligned}
&\max_{1\leq j\leq M} P_j\Big\{R_j(\hat{S}_n) - \min_{f\in\mathcal{H}} R_j(f) \geq \nu_{n,M}\Big\} \\
&\qquad \geq \inf_{\psi}\max_{1\leq j\leq M} P_j(\psi \neq j),
\end{aligned} \tag{8.4}$$

where the infimum is taken over all tests taking values in $\{1, \ldots, M\}$, and P_j denotes the joint distribution of $Y_1, \ldots, Y_n$ that are independent Gaussian random variables with mean $\eta_j(x_i)$, respectively. It follows from Tsybakov [(2009), Proposition 2.3 and Theorem 2.5] that if for any $1 \leq j, k \leq M$, the Kullback–Leibler divergence between P_j and P_k satisfies

$$\mathcal{K}(P_j, P_k) < \frac{\log M}{8}, \tag{8.5}$$

then there exists a constant $C > 0$ such that

$$\inf_{\psi}\max_{1\leq j\leq M} P_j(\psi \neq j) \geq C. \tag{8.6}$$

To check (8.5), observe that, choosing $\tau \le 1$ and applying (8.2), we get

$$\mathcal{K}(P_j, P_k) = \frac{n}{2\sigma^2}\|\eta_j - \eta_k\|^2 = \frac{\tau^2 \log M}{18\bar{\kappa}^2}\|\phi_j - \phi_k\|^2 < \frac{\log M}{8}.$$

Therefore, in view of (8.4) and (8.6), we find, using the Markov inequality, that for any selector $\hat{S}_n$,

$$\max_{1\le j\le M} E_j\Big[R_j(\hat{S}_n) - \min_{f\in\mathcal{H}} R_j(f)\Big] \ge C v_{n,M} = C_* \sigma \sqrt{\frac{\log M}{n}},$$

where E_j denotes the expectation with respect to P_j.

ACKNOWLEDGMENTS

The first author is supported in part by the NSF DMS-09-06424, DMS-10-53987.

REFERENCES

ALQUIER, P. and LOUNICI, K. (2011). PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.* **5** 127–145. MR2786484

BARANIUK, R., DAVENPORT, M., DEVORE, R. and WAKIN, M. (2008). A simple proof of the restricted isometry property for random matrices. *Constr. Approx.* **28** 253–263. MR2453366

BARTLETT, P. L., BOUCHERON, S. and LUGOSI, G. (2002). Model selection and error estimation. *Mach. Learn.* **48** 85–113.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** 203–268. MR1848946

BREHENY, P. and HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5** 232–253. MR2810396

BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

CATONI, O. (1999). Universal aggregation rules with exact bias bounds. Technical report, Laboratoire de Probabilités et Modeles Aléatoires, Preprint 510.

CATONI, O. (2004). *Statistical Learning Theory and Stochastic Optimization*. *Lecture Notes in Math.* **1851**. Springer, Berlin. MR2163920

CATONI, O. (2007). *Pac-Bayesian Supervised Classification*: *The Thermodynamics of Statistical Learning*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **56**. IMS, Beachwood, OH. MR2483528

DALALYAN, A. S. and SALMON, J. (2011). Sharp oracle inequalities for aggregation of affine estimators. Available at ArXiv:1104.3969.

DALALYAN, A. S. and TSYBAKOV, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning Theory*. *Lecture Notes in Computer Science* **4539** 97–111. Springer, Berlin. MR2397581

DALALYAN, A. and TSYBAKOV, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.* **72** 39–61.

DALALYAN, A. and TSYBAKOV, A. B. (2012). Mirror averaging with sparsity priors. *Bernoulli* **18** 914–944.

DALALYAN, A. S. and TSYBAKOV, A. B. (2012). Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.* **78** 1423–1443. MR2926142

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GERCHINOVITZ, S. (2011). Prediction of individual sequences and prediction in the statistical framework: Some links around sparse regression and aggregation techniques. Ph.D. thesis, Univ. Paris Sud—Paris XI.

GIRAUD, C. (2007). Mixing least-squares estimators when the variance is unknown. Available at arXiv:0711.0372.

GIRAUD, C. (2008). Mixing least-squares estimators when the variance is unknown. *Bernoulli* **14** 1089–1107. MR2543587

HUANG, J. and ZHANG, T. (2010). The benefit of group sparsity. *Ann. Statist.* **38** 1978–2004. MR2676881

JUDITSKY, A., RIGOLLET, P. and TSYBAKOV, A. B. (2008). Learning by mirror averaging. *Ann. Statist.* **36** 2183–2206. MR2458184

JUDITSKY, A. B., NAZIN, A. V., TSYBAKOV, A. B. and VAYATIS, N. (2005). Recursive aggregation of estimators by the mirror descent method with averaging. *Probl. Inf. Transm.* **41** 368–384. MR2198228

KNEIP, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22** 835–866. MR1292543

KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869

LECUÉ, G. (2007). Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.* **35** 1698–1721. MR2351102

LECUÉ, G. (2012). Empirical risk minimization is optimal for the Convex aggregation problem. *Bernoulli*. To appear.

LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52** 3396–3410. MR2242356

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

LUGOSI, G. and WEGKAMP, M. (2004). Complexity regularization via localized random penalties. *Ann. Statist.* **32** 1679–1697. MR2089138

NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math.* **1738** 85–277. Springer, Berlin. MR1775640

RIGOLLET, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *Ann. Statist.* **40** 639–665.

RIGOLLET, P. and TSYBAKOV, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** 260–280. MR2356821

RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. MR2816337

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York. MR2080278

RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D*: *Nonlinear Phenomena* **60** 259–268.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 91–108. MR2136641

TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In *COLT* (B. Schölkopf and M. K. Warmuth, eds.). *Lecture Notes in Computer Science* **2777** 303–313. Springer, Berlin.

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

WEGKAMP, M. (2003). Model selection in nonparametric regression. *Ann. Statist.* **31** 252–273. MR1962506

YANG, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* **10** 25–47. MR2044592

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701