Aggregation and Sparsity via ℓ_1 Penalized Least Squares

Florentina Bunea[1], Alexander B. Tsybakov[2], and Marten H. Wegkamp[1]

[1] Florida State University, Department of Statistics, Tallahassee FL 32306, USA
{bunea,wegkamp}@stat.fsu.edu*
[2] Université Paris VI, Laboratoire de Probabilités et Modèles Aléatoires, 4, Place Jussieu, B.P. 188, 75252 PARIS Cedex 05, France
tsybakov@ccr.jussieu.fr

Abstract. This paper shows that near optimal rates of aggregation and adaptation to unknown sparsity can be simultaneously achieved via ℓ_1 penalized least squares in a nonparametric regression setting. The main tool is a novel oracle inequality on the sum between the empirical squared loss of the penalized least squares estimate and a term reflecting the sparsity of the unknown regression function.

1 Introduction

In this paper we study aggregation in regression models via penalized least squares with data dependent ℓ_1 penalties. We begin by stating our framework. Let $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a sample of i.i.d. random pairs (X_i, Y_i) with

$$Y_i = f(X_i) + W_i, \quad i = 1, \ldots, n, \tag{1}$$

where $f : \mathcal{X} \to \mathbb{R}$ is an unknown regression function to be estimated, $\mathcal{X}$ is a Borel subset of $\mathbb{R}^d$, the X_i's are random elements in $\mathcal{X}$ with probability measure μ, and the regression errors W_i satisfy $\mathbb{E}(W_i|X_i) = 0$. Let $\mathcal{F}_M = \{f_1, \ldots, f_M\}$ be a collection of functions. The functions f_j can be viewed as estimators of f constructed from a training sample. Here we consider the ideal situation in which they are fixed; we concentrate on learning only. Assumptions **(A1)** and **(A2)** on the regression model (1) are supposed to be satisfied throughout the paper.

Assumption (A1). *The random variables W_i are independent, identically distributed with $\mathbb{E}(W_i|X_i) = 0$ and $\mathbb{E}[\exp(|W_i|)|X_i] \le b$, for some $b > 0$. The random variables X_i are independent, identically distributed*

* Research of Bunea and Wegkamp is supported in part by NSF grant DMS 0406049

with measure μ.

Assumption (A2). *The functions $f : \mathcal{X} \to \mathbb{R}$ and $f_j : \mathcal{X} \to \mathbb{R}$, $j = 1, \ldots, M$, with $M \geq 2$, belong to the class $\mathcal{F}_0$ of uniformly bounded functions defined by*

$$\mathcal{F}_0 \stackrel{\text{def}}{=} \left\{ g : \mathcal{X} \to \mathbb{R} \,\middle|\, \|g\|_\infty \leq L \right\}$$

where $L < \infty$ is a constant that is not necessarily known to the statistician and $\|g\|_\infty = \sup_{x \in \mathcal{X}} |g(x)|$.

Some references to aggregation of arbitrary estimators in regression models are [13], [10], [17], [18], [9], [2], [15], [16] and [7]. This paper extends the results of [4], who consider regression with fixed design and Gaussian errors W_i.

We introduce first our aggregation scheme. For any $\lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M$, define $\mathsf{f}_\lambda(x) = \sum_{j=1}^M \lambda_j f_j(x)$ and let

$$M(\lambda) = \sum_{j=1}^{M} I_{\{\lambda_j \neq 0\}} = \text{Card } J(\lambda)$$

denote the number of non-zero coordinates of λ, where $I_{\{\cdot\}}$ denotes the indicator function, and $J(\lambda) = \{j \in \{1, \ldots, M\} \, : \, \lambda_j \neq 0\}$. The value $M(\lambda)$ characterizes the *sparsity* of the vector λ: the smaller $M(\lambda)$, the "sparser" λ. Furthermore we introduce the residual sum of squares

$$\widehat{S}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \mathsf{f}_\lambda(X_i)\}^2,$$

for all $\lambda \in \mathbb{R}^M$. We aggregate the f_j's via penalized least squares. Given a penalty term $\text{pen}(\lambda)$, the penalized least squares estimator $\widehat{\lambda} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_M)$ is defined by

$$\widehat{\lambda} = \arg\min_{\lambda \in \mathbb{R}^M} \left\{ \widehat{S}(\lambda) + \text{pen}(\lambda) \right\}, \tag{2}$$

which renders the aggregated estimator

$$\widetilde{f}(x) = \mathsf{f}_{\widehat{\lambda}}(x) = \sum_{j=1}^{M} \widehat{\lambda}_j f_j(x). \tag{3}$$

Since the vector $\widehat{\lambda}$ can take any values in $\mathbb{R}^M$, the aggregate $\widetilde{f}$ is not a model selector in the traditional sense, nor is it necessarily a convex combination of the functions f_j. We consider the penalty

$$\text{pen}(\lambda) = 2\sum_{j=1}^{M} r_{n,j}|\lambda_j| \tag{4}$$

with data-dependent weights $r_{n,j} = r_n(M)\|f_j\|_n$, and

$$r_n(M) = A\sqrt{\frac{\log(Mn)}{n}} \tag{5}$$

where $A > 0$ is a suitably large constant. We write $\|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n g^2(X_i)$ for any $g : \mathcal{X} \to \mathbb{R}$. Note that our procedure is closely related to Lasso-type methods, see e.g. [14]. These methods can be reduced to (2) where now $\text{pen}(\lambda) = \sum_{j=1}^M r|\lambda_j|$ with a tuning constant $r > 0$ that is independent of j and of the data.

The main goal of this paper is to show that the aggregate $\widetilde{f}$ satisfies the following two properties.

P1. Optimality of aggregation. The loss $\|\widetilde{f} - f\|_n^2$ is simultaneously smaller, with probability close to 1, than the model selection, convex and linear oracle bounds of the form $C_0 \inf_{\lambda\in H^M} \|\mathsf{f}_\lambda - f\|_n^2 + \varDelta_{n,M}$, where $C_0 \geq 1$ and $\varDelta_{n,M} \geq 0$ is a remainder term independent of f. The set H^M is either the whole $\mathbb{R}^M$ (for linear aggregation), or the simplex $\varLambda^M$ in $\mathbb{R}^M$ (for convex aggregation), or the set of vertices of $\varLambda^M$, except the vertex $(0,\ldots,0) \in \mathbb{R}^M$ (for model selection aggregation). Optimal (minimax) values of $\varDelta_{n,M}$, called optimal rates of aggregation, are given in [15], and they have the form

$$\psi_{n,M} \asymp \begin{cases} M/n & \text{for (L) aggregation,} \\ M/n & \text{for (C) aggregation, if } M \leq \sqrt{n}, \\ \sqrt{\{\log(1+M/\sqrt{n})\}/n} & \text{for (C) aggregation, if } M > \sqrt{n}, \\ (\log M)/n & \text{for (MS) aggregation.} \end{cases} \tag{6}$$

Corollary 2 in Section 3 below shows that these optimal rates are attained by our procedure within a $\log(Mn)$ factor.

P2. Taking advantage of the sparsity. If $\lambda^* \in \mathbb{R}^M$ is such that $f = f_{\lambda^*}$ (classical linear regression) or f can be sufficiently well approximated by f_{λ^*} then, with probability close to 1, the ℓ_1 norm of $\hat{\lambda} - \lambda^*$ is bounded, up to known constants and logarithms, by $M(\lambda^*)/\sqrt{n}$. This means that the estimator $\hat{\lambda}$ of the parameter λ^* adapts to the sparsity of the problem: its rate of convergence is faster when the "oracle" vector λ^* is sparser. Note, in contrast, that for the ordinary least squares estimator the corresponding rate is $M/\sqrt{n}$, with the overall dimension M, regardless on the sparsity of λ^*.

To show **P1** and **P2** we first establish a new type of oracle inequality in Section 2. Instead of deriving oracle bounds for the deviation of $\widetilde{f}$ from f, which is usually the main object of interest in the literature, we obtain a stronger result. Namely, we prove a simultaneous oracle inequality for the sum of two deviations: that of $\widetilde{f}$ from f and that of $\hat{\lambda}$ from the "oracle" value of λ. Similar developments in a different context are given by [5] and [12]. The two properties **P1** and **P2** can be then shown as consequences of this result.

2 Main oracle inequality

In this section we state our main oracle bounds. We define the matrices $\Psi_{n,M} = \left(\frac{1}{n}\sum_{i=1}^n f_j(X_i) f_{j'}(X_i)\right)_{1 \le j,j' \le M}$ and the diagonal matrices $\mathrm{diag}(\Psi_{n,M}) = \mathrm{diag}(\|f_1\|_n^2, \ldots, \|f_M\|_n^2)$. We consider the following assumption on the class $\mathcal{F}_M$.

Assumption (A3). *For any $n \ge 1$, $M \ge 2$ there exist constants $\kappa_{n,M} > 0$ and $0 \le \pi_{n,M} < 1$ such that*

$$\mathbb{P}\left(\Psi_{n,M} - \kappa_{n,M}\,\mathrm{diag}(\Psi_{n,M})) \ge 0\right) \ge 1 - \pi_{n,M},$$

where $A \ge 0$ for a square matrix A, means that A is positive semi-definite. Assumption (A3) is trivially fulfilled with $\kappa_{n,M} \equiv 1$ if $\Psi_{n,M}$ is a diagonal matrix, with some eigenvalues possibly equal to zero. In particular, there exist degenerate matrices $\Psi_{n,M}$ satisfying Assumption (A3). Assumption (A4) below subsumes (A3) for appropriate choices of $\kappa_{n,M}$ and $\pi_{n,M}$, see the proof of Theorem 2.

Denote the inner product and the norm in $L_2(\mu)$ by $< \cdot, \cdot >$ and $\|\cdot\|$ respectively. Define $c_0 = \min\{\|f_j\| : j \in \{1, \ldots, M\} \text{ and } \|f_j\| > 0\}$.

Theorem 1. *Assume (A1), (A2) and (A3). Let $\widetilde{f}$ be the penalized least squares aggregate defined by (3) with penalty (4). Then, for any $n \geq 1$, $M \geq 2$ and $a > 1$, the inequality*

$$\|\widetilde{f} - f\|_n^2 + \frac{a}{a-1}\sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j| \tag{7}$$

$$\leq \frac{a+1}{a-1}\|\mathsf{f}_\lambda - f\|_n^2 + \frac{4a^2}{\kappa_{n,M}(a-1)} r_n^2(M) M(\lambda), \qquad \forall \lambda \in \mathbb{R}^M,$$

is satisfied with probability $\geq 1 - p_{n,M}$ where

$$p_{n,M} = \pi_{n,M} + 2M\exp\left(-\frac{n r_n(M) c_0}{4L^2 b + L r_n(M) c_0/2}\right) + 2M\exp\left(-\frac{n r_n^2(M) c_0^2}{128 L^2 b}\right) + M\exp\left(-\frac{n c_0^2}{2L^2}\right).$$

Proof of Theorem 1 is given in Section 5. This theorem is general but not ready to use because the probabilities $\pi_{n,M}$ and the constants $\kappa_{n,M}$ in Assumption (A3) need to be evaluated. A natural way to do this is to deal with the expected matrices $\Psi_M = \mathbb{E}(\Psi_{n,M}) = \big(\langle f_j, f_{j'}\rangle\big)_{1\leq j,j'\leq M}$ and $\mathrm{diag}(\Psi_M) = \mathrm{diag}(\|f_1\|^2, \ldots, \|f_M\|^2)$. Consider the following analogue of Assumption (A3) stated in terms of these matrices.

Assumption (A4). *There exists $\kappa_M > 0$ such that the matrix $\Psi_M - \kappa_M \,\mathrm{diag}(\Psi_M)$ is positive semi-definite for any given $M \geq 2$.*
For discussion of this assumption, see [4] and Remark 1 below.

Theorem 2. *Assume (A1), (A2) and (A4). Let $\widetilde{f}$ be the penalized least squares aggregate defined by (3) with penalty (4). Then, for any $n \geq 1$, $M \geq 2$ and $a > 1$, the inequality*

$$\|\widetilde{f} - f\|_n^2 + \frac{a}{a-1}\sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j| \tag{8}$$

$$\leq \frac{a+1}{a-1}\|\mathsf{f}_\lambda - f\|_n^2 + \frac{16a^2}{\kappa_M(a-1)} r_n^2(M) M(\lambda), \qquad \forall \lambda \in \mathbb{R}^M,$$

is satisfied with probability $\geq 1 - p_{n,M}$ where

$$p_{n,M} = 2M\exp\left(-\frac{n r_n(M) c_0}{4L^2 b + L r_n(M) c_0/2}\right) + 2M\exp\left(-\frac{n r_n^2(M) c_0^2}{128 L^2 b}\right) + M^2\exp\left(-\frac{n}{16 L^4 M^2}\right) + 2M\exp\left(-\frac{n c_0^2}{2L^2}\right). \tag{9}$$

Remark 1. The simplest case of Theorem 2 corresponds to a positive definite matrix Ψ_M. Then Assumption (A4) is satisfied with $\kappa_M = \xi_{\min}(M)/L^2$, where $\xi_{\min}(M) > 0$ is the smallest eigenvalue of Ψ_M. Furthermore, $c_0 \geq \xi_{\min}(M)$. We can therefore replace κ_M and c_0 by $\xi_{\min}(M)/L^2$ and $\xi_{\min}(M)$, respectively, in the statement of Theorem 2.

Remark 2. Theorem 2 allows us to treat asymptotics for $n \to \infty$ and fixed, but possibly large M, and for both $n \to \infty$ and $M = M_n \to \infty$. The asymptotic considerations can suggest a choice of the tuning parameter $r_n(M)$. In fact, it is determined by two antagonistic requirements. The first one is to keep $r_n(M)$ as small as possible, in order to improve the bound (8). The second one is to take $r_n(M)$ large enough to obtain the convergence of the probability $p_{n,M}$ to 0. It is easy to see that, asymptotically, as $n \to \infty$, the choice that meets the two requirements is given by (5). Note, however, that $p_{n,M}$ contains the terms independent of $r_n(M)$, and a necessary condition for their convergence to 0 is

$$n/(M^2 \log M) \to \infty. \tag{10}$$

This condition means that Theorem 2 is only meaningful for moderately large dimensions M.

3 Optimal aggregation property

Here we state corollaries of the results of Section 2 implying the property **P1**.

Corollary 1. *Assume (A1), (A2) and (A4). Let $\widetilde{f}$ be the penalized least squares aggregate defined by (3) with penalty (4). Then, for any $n \geq 1$, $M \geq 2$ and $a > 1$, the inequality*

$$\|\widetilde{f} - f\|_n^2 \leq \inf_{\lambda \in \mathbb{R}^M} \left\{ \frac{a+1}{a-1} \|\mathsf{f}_\lambda - f\|_n^2 + \frac{16a^2}{\kappa_M (a-1)} r_n^2(M) M(\lambda) \right\}. \tag{11}$$

is satisfied with probability $\geq 1 - p_{n,M}$ where $p_{n,M}$ is given by (9).

This corollary is similar to a result in [4], but there the predictors X_i are assumed to be non-random and the oracle inequality is obtained for the expected risk. Arguing as in [4], we easily deduce from Corollary 1 the following result.

Corollary 2. *Let assumptions of Corollary 1 be satisfied and let* $r_n(M)$ *be as in (5). Then, for any* $\varepsilon > 0$*, there exists a constant* $C > 0$ *such that the inequalities*

$$\|\widetilde{f} - f\|_n^2 \le (1+\varepsilon) \inf_{1\le j\le M} \|f_j - f\|_n^2 + C\left(1+\varepsilon+\varepsilon^{-1}\right)\frac{\log(M\vee n)}{n}, \quad (12)$$

$$\|\widetilde{f} - f\|_n^2 \le (1+\varepsilon) \inf_{\lambda\in\mathbb{R}^M} \|\mathsf{f}_\lambda - f\|_n^2 + C\left(1+\varepsilon+\varepsilon^{-1}\right)\frac{M\log(M\vee n)}{n}, \quad (13)$$

$$\|\widetilde{f} - f\|_n^2 \le (1+\varepsilon) \inf_{\lambda\in\Lambda^M} \|\mathsf{f}_\lambda - f\|_n^2 + C\left(1+\varepsilon+\varepsilon^{-1}\right)\overline{\psi}_n^C(M), \quad (14)$$

are satisfied with probability $\ge 1 - p_{n,M}$*, where* $p_{n,M}$ *is given by (9) and*

$$\overline{\psi}_n^C(M) = \begin{cases} (M\log n)/n & \text{if } M \le \sqrt{n}, \\ \sqrt{(\log M)/n} & \text{if } M > \sqrt{n}. \end{cases}$$

This result shows that the optimal (M), (C) and (L) bounds given in (6) are nearly attained, up to logarithmic factors, if we choose the tuning parameter $r_n(M)$ as in (5).

4 Taking advantage of the sparsity

In this section we show that our procedure automatically adapts to the unknown sparsity of $f(x)$. We consider the following assumption to formulate our notion of sparsity.

Assumption (A5). *There exists* $\lambda^* = \lambda^*(f)$ *such that*

$$\|\mathsf{f}_{\lambda^*} - f\|_\infty^2 \le r_n^2(M)M(\lambda^*). \quad (15)$$

Assumption (A5) is obviously satisfied in the parametric framework $f \in \{\mathsf{f}_\lambda, \lambda \in \mathbb{R}^M\}$. It is also valid in many nonparametric settings. For example, if the functions f_j form a basis, and f is a smooth function that can be well approximated by the linear span of $M(\lambda^*)$ basis functions (cf., e.g., [1], [11]). The vector λ^* satisfying (15) will be called oracle. In fact, Assumption (A5) can be viewed as a definition of the oracle.

We establish inequalities in terms of $M(\lambda^*)$ not only for the pseudo-distance $\|\tilde{f} - f\|_n^2$, but also for the ℓ_1 distance $\sum_{j=1}^M |\widehat{\lambda}_j - \lambda_j^*|$, as a consequence of Theorem 2. In fact, with probability close to one (see Lemma

1 below), if $\|f_j\| \ge c_0 > 0$, $\forall j = 1, \dots, M$, we have

$$\sum_{j=1}^{M} r_{n,j} |\widehat{\lambda}_j - \lambda_j| \ge \frac{r_n(M) c_0}{2} \sum_{j=1}^{M} |\widehat{\lambda}_j - \lambda_j|. \tag{16}$$

Together with (15) and Theorem 2 this yields that, with probability close to one,

$$\sum_{j=1}^{M} |\widehat{\lambda}_j - \lambda_j^*| \le C r_n(M) M(\lambda^*), \tag{17}$$

where $C > 0$ is a constant. If we choose $r_n(M)$ as in (5), this achieves the aim described in **P2**.

Corollary 3. *Assume (A1), (A2), (A4), (A5) and* $\min_{1 \le j \le M} \|f_j\| \ge c_0 > 0$. *Let* $\widetilde{f}$ *be the penalized least squares aggregate defined by (3) with penalty (4). Then, for any* $n \ge 1$, $M \ge 2$ *we have*

$$\mathbb{P}\left(\|\widetilde{f} - f\|_n^2 \le C_1 r_n^2(M) M(\lambda^*) \right) \ge 1 - p_{n,M}^*, \tag{18}$$

$$\mathbb{P}\Big(\sum_{j=1}^{M} |\widehat{\lambda}_j - \lambda_j^*| \le C_2 r_n(M) M(\lambda^*) \Big) \ge 1 - p_{n,M}^*, \tag{19}$$

where $C_1, C_2 > 0$ *are constants depending only on* κ_M *and* c_0, $p_{n,M}^* = p_{n,M} + M \exp\{-nC_0^2/(2L^2)\}$ *and the* $p_{n,M}$ *are given in Theorem 2.*

Remark 3. Part (18) of Corollary 3 can be compared to [11] who consider the same regression model with random design and obtain inequalities similar to (18) for a more specific setting where the f_j's are the basis functions of a reproducing kernel Hilbert space, the matrix Ψ_M is close to the identity matrix and the random errors of the model are uniformly bounded. Part (19) (the sparsity property) of Corollary 3 can be compared with [6] who consider the regression model with non-random design points $X_1, \dots, X_n$ and Gaussian errors W_i and control the ℓ_2 (not ℓ_1) deviation between $\widehat{\lambda}$ and λ^*.

Remark 4. Consider the particular case of linear parametric regression models where $f = \mathsf{f}_{\lambda^*}$. Assume for simplicity that the matrix Ψ_M is non-degenerate. Then all the components of the ordinary least squares estimate λ^{OLS} converge to the corresponding components of λ^* in probability with the rate $1/\sqrt{n}$. Thus we have

$$\sum_{j=1}^{M} |\lambda_j^{OLS} - \lambda_j^*| = O_p(M/\sqrt{n}), \tag{20}$$

as $n \to \infty$. Assume that $M(\lambda^*) \ll M$. If we knew exactly the set of non-zero coordinates $J(\lambda^*)$ of the oracle λ^*, we would perform the ordinary least squares on that set to obtain (20) with the rate $O_p(M(\lambda^*)/\sqrt{n})$. However, neither $J(\lambda^*)$, nor $M(\lambda^*)$ are known. If $r_n(M)$ is chosen as in (5) our estimator $\widehat{\lambda}$ achieves the same rate, up to logarithms without prior knowledge of $J(\lambda^*)$.

5 Proofs of the theorems

Proof of Theorem 1. By definition, $\widetilde{f} = \mathsf{f}_{\widehat{\lambda}}$ satisfies

$$\widehat{S}(\widehat{\lambda}) + \sum_{j=1}^{M} 2r_{n,j}|\widehat{\lambda}_j| \le \widehat{S}(\lambda) + \sum_{j=1}^{M} 2r_{n,j}|\lambda_j|$$

for all $\lambda \in \mathbb{R}^M$, which we may rewrite as

$$\|\widetilde{f} - f\|_n^2 + \sum_{j=1}^{M} 2r_{n,j}|\widehat{\lambda}_j| \le \|\mathsf{f}_\lambda - f\|_n^2 + \sum_{j=1}^{M} 2r_{n,j}|\lambda_j| + \frac{2}{n}\sum_{i=1}^{n} W_i(\widetilde{f} - \mathsf{f}_\lambda)(X_i).$$

We define the random variables $V_j = \frac{1}{n}\sum_{i=1}^{n} f_j(X_i)W_i$, $1 \le j \le M$ and the event $E_1 = \bigcap_{j=1}^{M} \{2|V_j| \le r_{n,j}\}$. If E_1 holds we have

$$\frac{2}{n}\sum_{i=1}^{n} W_i(\widetilde{f} - \mathsf{f}_\lambda)(X_i) = 2\sum_{j=1}^{M} V_j(\widehat{\lambda}_j - \lambda_j) \le \sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j|$$

and therefore, still on E_1,

$$\|\widetilde{f} - f\|_n^2 \le \|\mathsf{f}_\lambda - f\|_n^2 + \sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j| + \sum_{j=1}^{M} 2r_{n,j}|\lambda_j| - \sum_{j=1}^{M} 2r_{n,j}|\widehat{\lambda}_j|.$$

Adding the term $\sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j|$ to both sides of this inequality yields further, on E_1,

$$\begin{aligned}
&\|\widetilde{f} - f\|_n^2 + \sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j| \\
&\leq \|\mathsf{f}_\lambda - f\|_n^2 + 2\sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j| + \sum_{j=1}^{M} 2r_{n,j}|\lambda_j| - \sum_{j=1}^{M} 2r_{n,j}|\widehat{\lambda}_j| \\
&= \|\mathsf{f}_\lambda - f\|_n^2 + \left(\sum_{j=1}^{M} 2r_{n,j}|\widehat{\lambda}_j - \lambda_j| - \sum_{j \notin J(\lambda)} 2r_{n,j}|\widehat{\lambda}_j| \right) \\
&\quad + \left(- \sum_{j \in J(\lambda)} 2r_{n,j}|\widehat{\lambda}_j| + \sum_{j \in J(\lambda)} 2r_{n,j}|\lambda_j| \right).
\end{aligned}$$

Recall that $J(\lambda)$ denotes the set of indices of the non-zero elements of λ, and $M(\lambda) = \text{Card } J(\lambda)$. Rewriting the right-hand side of the previous display, we find that, on E_1,

$$\|\widetilde{f} - f\|_n^2 + \sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j| \ \leq \|\mathsf{f}_\lambda - f\|_n^2 + 4 \sum_{j \in J(\lambda)} r_{n,j}|\widehat{\lambda}_j - \lambda_j| \quad (21)$$

by the triangle inequality and the fact that $\lambda_j = 0$ for $j \notin J(\lambda)$. Define the random event $E_0 = \{\Psi_{n,M} - \kappa_{n,M}\,\text{diag}(\Psi_{n,M}) \geq 0\}$. On $E_0 \cap E_1$ we have

$$\begin{aligned}
\sum_{j \in J(\lambda)} r_{n,j}^2|\widehat{\lambda}_j - \lambda_j|^2 &\leq r_n^2 \sum_{j=1}^{M} \|f_j\|_n^2|\widehat{\lambda}_j - \lambda_j|^2 \\
&= r_n^2(\widehat{\lambda} - \lambda)'\text{diag}(\Psi_{n,M})(\widehat{\lambda} - \lambda) \\
&\leq r_n^2\kappa^{-1}(\widehat{\lambda} - \lambda)'\Psi_{n,M}(\widehat{\lambda} - \lambda) \\
&= r_n^2\kappa^{-1}\|\widetilde{f} - \mathsf{f}_\lambda\|_n^2,
\end{aligned} \quad (22)$$

where, for brevity, $r_n = r_n(M)$, $\kappa = \kappa_{n,M}$. Combining (21) and (22) with the Cauchy-Schwarz and triangle inequalities, respectively, we find further that, on $E_0 \cap E_1$,

$$\begin{aligned}
&\|\widetilde{f} - f\|_n^2 + \sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j| \\
&\le \|\mathsf{f}_\lambda - f\|_n^2 + 4 \sum_{j \in J(\lambda)} r_{n,j}|\widehat{\lambda}_j - \lambda_j| \\
&\le \|\mathsf{f}_\lambda - f\|_n^2 + 4\sqrt{M(\lambda)} \sqrt{\sum_{j \in J(\lambda)} r_{n,j}^2 |\widehat{\lambda}_j - \lambda_j|^2} \\
&\le \|\mathsf{f}_\lambda - f\|_n^2 + 4 r_n \sqrt{M(\lambda)/\kappa} \left(\|\widetilde{f} - f\|_n + \|\mathsf{f}_\lambda - f\|_n \right).
\end{aligned}$$

The preceding inequality is of the simple form $v^2 + d \le c^2 + vb + cb$ with $v = \|\widetilde{f} - f\|_n$, $b = 4r_n\sqrt{M(\lambda)/\kappa}$, $c = \|\mathsf{f}_\lambda - f\|_n$ and $d = \sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j|$. After applying the inequality $2xy \le x^2/\alpha + \alpha y^2$ $(x, y \in \mathbb{R},\ \alpha > 0)$ twice, to $2bc$ and $2bv$, respectively, we easily find $v^2 + d \le v^2/(2\alpha) + \alpha\, b^2 + (2\alpha + 1)/(2\alpha)\, c^2$, whence $v^2 + d\{a/(a-1)\} \le a/(a-1)\{b^2(a/2) + c^2(a+1)/a\}$ for $a = 2\alpha > 1$. On the random event $E_0 \cap E_1$, we now get that

$$\|\widetilde{f} - f\|_n^2 + \frac{a}{a-1} \sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j| \le \frac{a+1}{a-1}\|\mathsf{f}_\lambda - f\|_n^2 + \frac{4a^2}{\kappa(a-1)} r_n^2 M(\lambda),$$

for all $a > 1$. Using Lemma 2 proved below and the fact that $\mathbb{P}\{E_0\} \ge 1 - \pi_{n,M}$ we get Theorem 1. ■

Proof of Theorem 2. Let $\mathcal{F} = \mathrm{span}(f_1, \ldots, f_M)$ be the linear space spanned by $f_1, \ldots, f_M$. Define the events $E_{0,*} = \{\Psi_{n,M} - (\kappa_M/4)\,\mathrm{diag}(\Psi_{n,M}) \ge 0\}$ and

$$E_2 = \bigcap_{j=1}^{M} \left\{ \|f_j\|_n^2 \le 2\|f_j\|^2 \right\}, \qquad E_3 = \left\{ \sup_{f \in \mathcal{F}\setminus\{0\}} \frac{\|f\|^2}{\|f\|_n^2} \le 2 \right\}.$$

Clearly, on E_2 we have $\mathrm{diag}(\Psi_{n,M}) \le 2\,\mathrm{diag}(\Psi_M)$ and on E_3 we have the matrix inequality $\Psi_{n,M} \ge \Psi_M/2$. Therefore, using Assumption (A4), we get that the complement $E_{0,*}^C$ of $E_{0,*}$ satisfies $E_{0,*}^C \cap E_2 \cap E_3 = \emptyset$, which yields

$$\mathbb{P}\{E_{0,*}^C\} \le \mathbb{P}\{E_2^C\} + \mathbb{P}\{E_3^C\}.$$

Thus, Assumption (A3) holds with $\kappa_{n,M} \equiv \kappa_M/4$ any $\pi_{n,M} \ge \mathbb{P}\{E_2^C\} + \mathbb{P}\{E_3^C\}$. Taking the particular value of $\pi_{n,M}$ as a sum of the upper bounds on $\mathbb{P}\{E_2^C\}$ and $\mathbb{P}\{E_3^C\}$ from Lemma 1 and from Lemma 3 (where we set

$q = M$, $g_i = f_i$) and applying Theorem 1 we get the result. ∎

Proof of Corollary 3. Let λ^* be a vector satisfying Assumption (A5). As in the proof of Theorem 2, we obtain that, on $E_1 \cap E_2 \cap E_3$,

$$\|\widetilde{f}-f\|_n^2+\frac{a}{a-1}\sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j-\lambda_j^*| \le \left\{\frac{a+1}{a-1}\|\mathsf{f}_{\lambda^*}-f\|_n^2+\frac{32a^2}{\kappa(a-1)}r_n^2 M(\lambda^*)\right\}$$

for all $a > 1$. We now note that, in view of Assumption (A5),

$$\|\mathsf{f}_{\lambda^*}-f\|_n^2 \le \|\mathsf{f}_{\lambda^*}-f\|_\infty^2 \le r_n^2 M(\lambda^*).$$

This yields (18). To obtain (19) we apply the bound (16), valid on the event E_4 defined in Lemma 1 below, and therefore we include into $p_{n,M}^*$ the term $M\exp\left(-nc_0^2/(2L^2)\right)$ to account for $\mathbb{P}\{E_4^C\}$. ∎

6 Technical Lemmas

Lemma 1. *Let Assumptions (A1) and (A2) hold. Then for the events*

$$E_2 = \{\|f_j\|_n^2 \le 2\|f_j\|^2,\ \forall\ 1 \le j \le M\}$$
$$E_4 = \{\|f_j\| \le 2\|f_j\|_n,\ \forall\ 1 \le j \le M\}$$

we have

$$\max(\mathbb{P}\{E_2^C\}, \mathbb{P}\{E_4^C\}) \le M\exp\left(-nc_0^2/(2L^2)\right). \tag{23}$$

Proof. Since $\|f_j\| = 0 \Longrightarrow \|f_j\|_n = 0\ \mu -$ a.s., it suffices to consider only the cases with $\|f_j\| > 0$. Inequality (23) then easily follows from the union bound and Hoeffding's inequality. ∎

Lemma 2. *Let Assumptions (A1) and (A2) hold. Then*

$$\mathbb{P}\{E_1^C\} \le 2M\exp\left(-\frac{nr_n(M)c_0}{4L^2b+Lr_n(M)c_0/2}\right)+2M\exp\left(-\frac{nr_n^2(M)c_0^2}{128L^2b}\right)$$
$$+M\exp\left(-\frac{nc_0^2}{2L^2}\right). \tag{24}$$

Proof. We use the following version of Bernstein's inequality (see, e.g., [3]): *Let $Z_1, \dots, Z_n$ be independent random variables such that*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|Z_i|^m \le \frac{m!}{2} w^2 d^{m-2},$$

for some positive constants w and d and for all $m \ge 2$. Then, for any $\varepsilon > 0$ we have

$$\mathbb{P}\left\{\sum_{i=1}^{n}(Z_i - \mathbb{E}Z_i) \ge n\varepsilon\right\} \le \exp\left(-\frac{n\varepsilon^2}{2(w^2 + d\varepsilon)}\right). \tag{25}$$

Here we apply this inequality to the variables $Z_{i,j} = f_j(X_i)W_i$, for each $j \in \{1, \dots, M\}$, conditioning on $X_1, \dots, X_n$. Note that $\mathbb{E}(Z_{i,j}|X_i) = 0$ by Assumption (A1) and $\|f_j\|_\infty \le L$ by Assumption (A2) for all j. Next, using Assumption (A1) we have

$$\mathbb{E}(|W_1|^m|X_1) = m!\mathbb{E}\left(\frac{|W_1|^m}{m!}\bigg|X_1\right) \le m!\mathbb{E}\left(\exp(|W_1|)|X_1\right) \le bm!.$$

Hence

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(|Z_{i,j}|^m|X_i) \le L^m\mathbb{E}(|W_1|^m|X_1) \le bm!L^m \le \frac{m!}{2}L^{m-2}(L\sqrt{2b})^2.$$

Consider the conditional probability $\mathbb{P}\{E_1^C|X_1, \dots, X_n\}$ for $(X_1, \dots, X_n) \in E_4$. Since $\|f_j\| = 0 \implies V_j = 0$ $\mu -$ a.s., it suffices to consider only the cases with $\|f_j\| > 0$. Using (25) we find that, on E_4,

$$\begin{aligned}\mathbb{P}\{E_1^C|X_1, \dots, X_n\} &\le \sum_{j:\|f_j\|>0} \mathbb{P}\left\{|V_j| \ge \frac{c_0 r_n}{4}\bigg|X_1, \dots, X_n\right\} \\ &\le 2M\exp\left(-\frac{nr_nc_0}{4L^2b + Lr_nc_0/2}\right) + 2M\exp\left(-\frac{nr_n^2c_0^2}{128L^2b}\right)\end{aligned}$$

where the last inequality holds since

$$\exp(-x/(2\alpha)) + \exp(-x/(2\beta)) \ge \exp(-x/(\alpha+\beta))$$

for $x, \alpha, \beta > 0$. Multiplying the last display by the indicator of E_4, taking expectations and using the bound on $\mathbb{P}\{E_4^C\}$ in Lemma 1, we get the result. ■

Lemma 3. *Let $\mathcal{F} = span(g_1, \ldots, g_q)$ be the linear space spanned by some functions $g_1, \ldots, g_q$ such that $g_i \in \mathcal{F}_0$. Then*

$$\mathbb{P}\left\{ \sup_{f \in \mathcal{F}\setminus\{0\}} \frac{\|f\|^2}{\|f\|_n^2} > 2 \right\} \leq q^2 \exp\left(-\frac{n}{16L^4 q^2} \right).$$

Proof. Let $\phi_1, \ldots, \phi_N$ be an orthonormal basis of $\mathcal{F}$ in $L_2(\mu)$ with $N \leq q$. For any symmetric $N \times N$ matrix A, we define

$$\bar{\rho}(A) = \sup \sum_{j=1}^{N} \sum_{j'=1}^{N} |\lambda_j||\lambda_{j'}||A_{j,j'}|,$$

where the supremum is taken over sequences $\{\lambda_j\}_{j=1}^N$ with $\sum_j \lambda_j^2 = 1$. By Lemma 5.2 in Baraud (2002), we find that

$$\mathbb{P}\left\{ \sup_{f \in \mathcal{F}\setminus\{0\}} \frac{\|f\|^2}{\|f\|_n^2} > 2 \right\} \leq q^2 \exp(-n/16C)$$

where $C = \max\left(\bar{\rho}^2(A), \bar{\rho}(A')\right)$, and A, A' are $N \times N$ matrices with entries $\sqrt{< \phi_j^2, \phi_{j'}^2 >}$ and $\|\phi_j \phi_{j'}\|_\infty$, respectively. Clearly,

$$\bar{\rho}(A) \leq L^2 \sup_{j,j'} \sum_{j=1}^{N} \sum_{j'=1}^{N} |\lambda_j||\lambda_{j'}| = L^2 \sup_j \left(\sum_{j=1}^{N} |\lambda_j| \right)^2 \leq L^2 q$$

where we used the Cauchy-Schwarz inequality. Similarly, $\bar{\rho}(A') \leq L^2 q$. ■

References

1. Baraud, Y.: Model selection for regression on a random design. ESAIM Probability & Statistics. **7** (2002) 127–146.
2. Birgé, L.: Model selection for Gaussian regression with random design. Prépublication n. 783, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 - Paris 7 (2002). `http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2002`.
3. Birgé, L., Massart, P.: Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. Bernouilli **4** (1998) 329 – 375.
4. Bunea, F., Tsybakov, A., Wegkamp, M.H.: Aggregation for Gaussian regression. Preprint (2005). `http://www.stat.fsu.edu/~wegkamp`.
5. Bunea, F., Wegkamp, M.: Two stage model selection procedures in partially linear regression. The Canadian Journal of Statistics **22** (2004) 1–14.
6. Candes, E., Tao,T.: The Dantzig selector: statistical estimation when p is much larger than n. Preprint (2005).

7. Catoni, O.: *Statistical Learning Theory and Stochastic Optimization*. Ecole d'Eté de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics, Springer, N.Y. (2004).
8. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81** (1994) 425–455.
9. Györfi, L., Kohler, M., Krzyżak, A., Walk, H.: *A Distribution-Free Theory of Nonparametric Regression*. Springer, N.Y.(2002).
10. Juditsky, A., Nemirovski, A.: Functional aggregation for nonparametric estimation. Annals of Statistics **28** (2000) 681–712.
11. Kerkyacahrian, G., Picard, D.: Tresholding in learning theory. Prépublication n.1017, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 - Paris 7. `http://www.proba.jussieu.fr/mathdoc/preprints/ index.html#2005`.
12. Koltchinskii, V.: Model selection and aggregation in sparse classification problems. Oberwolfach Reports: Meeting on Statistical and Probabilistic Methods of Model Selection, October 2005 (to appear).
13. Nemirovski, A.: *Topics in Non-parametric Statistics*. Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998, Lecture Notes in Mathematics, v. 1738, Springer, N.Y. (2000).
14. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B. **58** (1996) 267–288.
15. Tsybakov, A.B.: Optimal rates of aggregation. Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines. Lecture Notes in Artificial Intelligence **2777** 303–313. Springer-Verlag, Heidelberg (2003).
16. Wegkamp, M.H.: Model selection in nonparametric regression. Annals of Statistics **31** (2003) .
17. Yang, Y.: Combining different procedures for adaptive regression. J.of Multivariate Analysis **74** (2000) 135–161.
18. Yang, Y.: Aggregating regression procedures for a better performance (2001). Manuscript.