Submitted to the Annals of Statistics

AGGREGATION FOR GAUSSIAN REGRESSION

By Florentina Bunea*, Alexander B. Tsybakov and Marten H. Wegkamp*

Florida State University and Université Paris VI

This paper studies statistical aggregation procedures in the regression setting. A motivating factor is the existence of many different methods of estimation, leading to possibly competing estimators.

We consider here three different types of aggregation: model selection (MS) aggregation, convex (C) aggregation and linear (L) aggregation. The objective of (MS) is to select the optimal single estimator from the list; that of (C) is to select the optimal convex combination of the given estimators; and that of (L) is to select the optimal linear combination of the given estimators. We are interested in evaluating the rates of convergence of the excess risks of the estimators obtained by these procedures. Our approach is motivated by recent minimax results in Nemirovski (2000) and Tsybakov (2003).

There exist competing aggregation procedures achieving optimal convergence for each of the (MS), (C) and (L) cases separately. Since the bounds in these results are not directly comparable with each other, we suggest an alternative solution. We prove that all the three optimal bounds can be nearly achieved via a single "universal" aggregation procedure. Our procedure consists in mixing the initial estimators with weights obtained by penalized least squares. Two different penalties are considered: one of them is related to hard thresholding techniques, the second one is a data dependent L_1-type penalty.

1. Introduction.

In this paper we study aggregation procedures and their performance for regression models. Let $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a sample

*Research partially supported by NSF grant DMS 0406049

AMS 2000 subject classifications: Primary 62G08; secondary 62C20, 62G05, 62G20

Keywords and phrases: aggregation, minimax risk, model selection, nonparametric regression, oracle inequalities, penalized least squares

of independent random pairs (X_i, Y_i) with

$$Y_i = f(X_i) + W_i, \quad i = 1, \dots, n, \tag{1.1}$$

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 fixed elements in $\mathcal{X}$, and the errors W_i are zero mean random variables.

Aggregation of arbitrary estimators in regression models has recently received increasing attention: Nemirovski (2000), Juditsky and Nemirovski (2000), Yang (2000, 2001, 2004), Györfi *et al.* (2002), Birgé (2003), Tsybakov (2003), Wegkamp (2003) and Catoni (2004). A motivating factor is the existence of many different methods of estimation, leading to possibly competing estimators $\widehat{f}_1, \dots, \widehat{f}_M$. A natural idea is then to look for a new, improved, estimator $\widetilde{f}$ constructed by combining $\widehat{f}_1, \dots, \widehat{f}_M$ in a suitable way. Such an estimator $\widetilde{f}$ is called *aggregate* and its construction is called aggregation.

There exist three main aggregation problems: model selection (MS) aggregation, convex (C) aggregation and linear (L) aggregation. They are discussed in detail by Nemirovski (2000). The objective of (MS) is to select the optimal (in a sense to be defined) single estimator from the list; that of (C) is to select the optimal convex combination of the given estimators; and that of (L) is to select the optimal linear combination of the given estimators.

Aggregation procedures are typically based on sample splitting. The initial sample $\mathcal{D}_n$ is divided into a training sample, used to construct estimators $\widehat{f}_1, \dots, \widehat{f}_M$, and an independent validation sample, used to learn, i.e., to construct $\widetilde{f}$. In this paper we do not consider sample splitting schemes but rather deal with an idealized scheme. We fix the training sample and thus instead of estimators $\widehat{f}_1, \dots, \widehat{f}_M$, we have fixed functions $f_1, \dots, f_M$. That is, we focus our attention on aggregation. A passage to the initial model is straightforward: it is enough to condition on the training sample, and derive the bounds of Theorems 3.1 and 4.1 below. Then, we can take expectations in both sides of these inequalities over the distribution of the whole sample $\mathcal{D}_n$.

To give precise definitions, denote by $\|g\|_n = \left\{n^{-1}\sum_{i=1}^n g^2(X_i)\right\}^{1/2}$ the empirical norm of a function g in $\mathbb{R}^d$ and set $\mathsf{f}_\lambda = \sum_{j=1}^M \lambda_j f_j$ for any $\lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M$. The performance of an aggregate $\widetilde{f}$ can be judged against the following mathematical target:

$$\mathbb{E}_f \|\widetilde{f} - f\|_n^2 \le \inf_{\lambda \in H^M} \mathbb{E}_f \|\mathsf{f}_\lambda - f\|_n^2 + \Delta_{n,M}, \tag{1.2}$$

where $\Delta_{n,M} \ge 0$ is a remainder term *independent of* f characterizing the price to pay for aggregation, and the set H^M is either the whole $\mathbb{R}^M$ (for linear aggregation), or the simplex $\Lambda^M = \left\{\lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M : \lambda_j \ge 0,\ \sum_{j=1}^M \lambda_j \le 1\right\}$ (for convex aggregation), or the set of vertices of Λ^M, except the vertex $(0, \ldots, 0) \in \mathbb{R}^M$ (for model selection aggregation). Here and later $\mathbb{E}_f$ denotes the expectation with respect to the joint distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$ under model (1.1). The random functions f_λ attaining $\inf_{\lambda \in H^M} \mathbb{E}_f \|\mathsf{f}_\lambda - f\|_n^2$ in (1.2) for the three values taken by H^M are called (L), (C) and (MS) oracles, respectively. Note that these minimizers are not estimators since they depend on the true f.

We also introduce a fourth type of aggregation: subset selection, or (S) aggregation. For (S) aggregation we fix an integer $D \le M$ and put $H^M = \Lambda^{M,D}$, where $\Lambda^{M,D}$ denotes the set of all λ such that D of the coefficients of λ are equal to 1 and the remaining $M - D$ coefficients are zero. Note that the (MS) aggregation is a special case of subset selection ((S) aggregation) for $D = 1$. The literature on subset selection techniques is very large, and dates back to Akaike (1974), Mallows (1973) and Schwarz (1978). We refer to the recent comprehensive survey by Rao and Wu (2001) for references on methods geared mainly to parametric models. For a review of techniques leading to subset selection in nonparametric settings we refer to Barron, Birgé and Massart (1999) and the references therein.

We say that the aggregate $\widetilde{f}$ mimics the (L), (C), (MS) or (S) oracle if it satisfies (1.2) for the corresponding set H^M, with the minimal possible price for aggregation $\Delta_{n,M}$. Minimal possible values $\Delta_{n,M}$ for the three problems can be defined via a minimax setting and they are called optimal rates of aggregation [Tsybakov (2003)] and further denoted by $\psi_{n,M}$. Extending the work by Tsybakov (2003) in the random design case to the fixed design case, we will show in Section 3 and 5 that under mild conditions

$$
\psi_{n,M} \asymp \begin{cases} M/n & \text{for (L) aggregation,} \\ M/n & \text{for (C) aggregation, if } M \leq \sqrt{n}, \\ \sqrt{\{\log(1+M/\sqrt{n})\}/n} & \text{for (C) aggregation, if } M > \sqrt{n}, \\ \{D\log(1+M/D)\}/n & \text{for (S) aggregation,} \\ (\log M)/n & \text{for (MS) aggregation.} \end{cases} \tag{1.3}
$$

This implies that linear aggregation has the highest price, (MS) aggregation has the lowest one, and convex aggregation occupies an intermediate place. The oracle risks on the right in (1.2) satisfy a reversed inequality:

$$
\inf_{1\leq j\leq M} \mathbb{E}_f\|f_j - f\|_n^2 \geq \inf_{\lambda\in\Lambda^M} \mathbb{E}_f\|\mathsf{f}_\lambda - f\|_n^2 \geq \inf_{\lambda\in\mathbb{R}^M} \mathbb{E}_f\|\mathsf{f}_\lambda - f\|_n^2,
$$

since the sets over which the infima are taken are nested. There is no winner among the three aggregation techniques and the question how to choose the best among them remains open.

The ideal oracle inequality (1.2) is available only for some special cases. See Catoni (2004), Bunea and Nobel (2005), and Juditsky *et al.* (2005b) for (MS) aggregation; Nemirovski (2000), Juditsky and Nemirovski (2000), Tsybakov (2003), Juditsky *et al.* (2005a) for (C) aggregation with $M > \sqrt{n}$; and Tsybakov (2003), for (L) aggregation and for (C) aggregation with $M \leq \sqrt{n}$. For more general situations there exist less precise results of the type

$$
\mathbb{E}_f\|\widetilde{f} - f\|_n^2 \leq (1+\varepsilon) \inf_{\lambda\in H^M} \mathbb{E}_f\|\mathsf{f}_\lambda - f\|_n^2 + \Delta_{n,M,\varepsilon}, \tag{1.4}
$$

where $\varepsilon > 0$ is a constant independent of f and n, and $\Delta_{n,M,\varepsilon}$ is a remainder term, not necessarily having the same behavior in n and M as the optimal one $\psi_{n,M}$.

Bounds of the type (1.4) in regression problems have been obtained by many authors mainly for the model selection case, see, for example, Kneip (1994), Barron *et al.* (1999), Lugosi and Nobel (1999), Györfi *et al.* (2002), Baraud (2000, 2002), Birgé and Massart (2001a), Birgé and Massart (2001b), Bartlett *et al.* (2002), Wegkamp (2003), Birgé (2003), Bunea (2004), Catoni (2004), and the references cited in these works. Most of the papers on model selection treat particular restricted families of estimators, such as orthogonal series estimators, spline estimators, etc. An interesting recent development due to Leung and Barron (2004) covers model selection for all estimators admitting Stein's unbiased estimation of the risk. There are relatively few results on (MS) aggregation when the estimators are allowed to be arbitrary, see Yang (2000, 2001, 2002), Györfi *et al.* (2002), Birgé (2003), Tsybakov (2003), Wegkamp (2003) and Catoni (2004).

Various convex aggregation procedures for nonparametric regression have emerged in the last decade. The literature on oracle inequalities of the type (1.2) and (1.4) for the (C) aggregation case is not nearly as large as the one on model selection. We refer to Juditsky and Nemirovski (2000), Nemirovski (2000), Yang (2000, 2001, 2004), Birgé (2003), Tsybakov (2003), Koltchinskii (2004), Audibert (2005), Juditsky *et al.* (2005a), Bunea and Nobel (2005).

Finally, linear aggregation procedures are discussed in Nemirovski (2000), Tsybakov (2003) and Bunea and Nobel (2005).

Given the existence of competing aggregation procedures achieving either optimal (MS), or (C), or (L) bounds, there is an ongoing discussion as to which procedure is the best one. Since this cannot be decided by merely comparing the optimal bounds, we suggest an alternative solution. We show that all the three optimal (MS), (C) and (L) bounds can be nearly achieved via a single aggregation procedure. We also show that this procedure leads to near optimal bounds for the newly introduced (S) aggregation, for any subset size D. Our answer will thus meet the desiderata of both model (subset) selection and model averaging. The procedures that we suggest for aggregation are based on penalized least squares. We consider two penalties that can be associated with hard thresholding and soft thresholding (L_1 or Lasso type penalty), respectively.

The paper is organized as follows. Section 2 introduces notation and assumptions used throughout the paper. In Section 3 we show that a hard threshold aggregate satisfies inequalities of the type (1.4) with the optimal remainder term $\psi_{n,M}$. We establish the oracle inequalities for all three sets H^M under consideration, hence showing that the hard threshold aggregate achieves simultaneously the (S) (and hence the (MS)), the (C) and the (L) bounds. In Section 4 we study aggregation with the L_1 penalty and we obtain (1.4) simultaneously for the (S), (C) and (L) cases, with a remainder term $\Delta_{n,M}$ that differs from the optimal $\psi_{n,M}$ only in a logarithmic factor. We give the corresponding lower bounds for (S), (C) and (L) aggregation in Section 5, complementing the results obtained for the random design case by Tsybakov (2003). All proofs are deferred to the appendices.

2. Notation and assumptions.

The following two assumptions on the regression model (1.1) are supposed to be satisfied throughout the paper.

ASSUMPTION (A1) *The random variables W_i are independent and Gaussian $N(0, \sigma^2)$.*

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*

$$\mathcal{F}_0 \overset{\text{def}}{=} \left\{ g : \mathcal{X} \to \mathbb{R} \,\middle|\, \sup_{x \in \mathcal{X}} |g(x)| \leq L \right\}$$

where $L < \infty$ is a constant that is not necessarily known to the statistician.

For any $\lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M$, define

$$\mathsf{f}_\lambda(x) = \sum_{j=1}^{M} \lambda_j f_j(x).$$

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. For each $\lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M$, let $M(\lambda)$ denote the number of non-zero coordinates of λ, that is,

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

where $I_{\{\cdot\}}$ denotes the indicator function, and $J(\lambda) = \{j \in \{1, \ldots, M\} : \ \lambda_j \neq 0\}$. Furthermore we introduce the residual sum of squares

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

and the function

$$L(\lambda) = 2 \log \left(\frac{eM}{M(\lambda) \vee 1} \right),$$

for all $\lambda \in \mathbb{R}^M$. The method that we propose is based on aggregating 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

$$\widehat{\lambda} = \arg \min_{\lambda \in \mathbb{R}^M} \left\{ \widehat{S}(\lambda) + \text{pen}(\lambda) \right\}, \tag{2.1}$$

which renders in turn the aggregated estimator

$$\widetilde{f}(x) = \mathsf{f}_{\widehat{\lambda}}(x). \tag{2.2}$$

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. Nevertheless, we will show that it mimics the (S), (C) and (L) oracles when one of the following two penalties is used:

$$\text{pen}(\lambda) = \frac{2\sigma^2}{n} \left\{ 1 + \frac{2+a}{1+a} \sqrt{L(\lambda)} + \frac{1+a}{a} L(\lambda) \right\} M(\lambda) \tag{2.3}$$

or

$$\text{pen}(\lambda) = 2\sqrt{2}\sigma \sqrt{\frac{\log M + \log n}{n}} \sum_{j=1}^{M} \|\lambda_j f_j\|_n. \tag{2.4}$$

In (2.3), $a > 0$ is a parameter to be set by the user. We refer to the penalty in (2.3) as *hard threshold penalty*. This is motivated by the well known fact that, in the sequence space model (where the functions $f_1, \dots, f_M$ are orthonormal with respect to the scalar product induced by the norm $\|\cdot\|_n$), the penalty $\text{pen}(\lambda) \sim M(\lambda)$ leads to $\widehat{\lambda}_j$'s that are hard threshold values of the Y_j's (see, for instance, Härdle *et al.* (1998), page 138). Our penalty (2.3) is not exactly of that form, but it differs from it only in a logarithmic factor.

The penalty (2.4), again in the sequence space model, leads to $\widehat{\lambda}_j$'s that are soft threshold values of Y_j's. We will call it therefore soft threshold penalty or L_1-penalty. Penalized least squares estimators with soft threshold penalty $\text{pen}(\lambda) \sim \sum_{j=1}^M |\lambda_j|$ are closely related to Lasso-type estimators [Efron *et al.* (2004), see also Antoniadis and Fan (2001), Fan and Li (2001), Fan and Peng (2004), where other related penalties are discussed].

Our results show that the hard threshold penalty (2.3) allows optimal aggregation under (A1) and (A2). The soft threshold penalty (2.4) allows near optimal aggregation under somewhat different conditions.

3. Optimal aggregation with the hard threshold penalty.

In this section we show that the penalized least squares aggregate (2.2) corresponding to the penalty term (2.3) achieves simultaneously the (MS), (L), and (C) bounds of the form (1.4) with the correct rates $\Delta_{n,M} = \psi_{n,M}$. Consequently, the smallest bound is achieved by our aggregate. The next theorem presents an oracle inequality that implies all the three bounds, as well as a bound for (S) aggregation.

THEOREM 3.1. *Assume (A1) and (A2). Let $\widetilde{f}$ be the penalized least squares aggregate defined in (2.2) with penalty (2.3). Then, for all integers $n \geq 1$ and $M \geq 2$,*

$$\begin{aligned}(3.1)\quad &\mathbb{E}_f \|\widetilde{f} - f\|_n^2 \\ &\leq (1+a) \inf_{\lambda \in \mathbb{R}^M} \left[\|\mathsf{f}_\lambda - f\|_n^2 + \frac{\sigma^2}{n} \left\{ 5 + \frac{2+3a}{a} L(\lambda) \right\} M(\lambda) \right] + \frac{\sigma^2}{n} \frac{6(1+a)^2}{a(e-1)}.\end{aligned}$$

PROOF. See appendix A. □

COROLLARY 3.2. *Under the conditions of Theorem 3.1, there exists a constant $C > 0$ such that for all integers $n \geq 1$ and $M \geq 2$ and $D \leq M$, the following upper bounds for $R_{M,n} \overset{\text{def}}{=} \mathbb{E}_f \|\widetilde{f} - f\|_n^2$ hold:*

$$
\begin{aligned}
&(3.2)\quad R_{M,n} &\leq&\quad (1+a) \inf_{1\leq j\leq M} \|f_j - f\|_n^2 + C(1+a+a^{-1})\sigma^2 \frac{\log M}{n}\\
&(3.3)\quad R_{M,n} &\leq&\quad (1+a) \inf_{\lambda\in\Lambda^{M,D}} \|\mathsf{f}_\lambda - f\|_n^2 + C(1+a+a^{-1})\sigma^2 \frac{D}{n}\log\left(\frac{M}{D}+1\right)\\
&(3.4)\quad R_{M,n} &\leq&\quad (1+a) \inf_{\lambda\in\mathbb{R}^M} \|\mathsf{f}_\lambda - f\|_n^2 + C(1+a+a^{-1})\sigma^2 \frac{M}{n}\\
&(3.5)\quad R_{M,n} &\leq&\quad (1+a) \inf_{\lambda\in\Lambda^{M}} \|\mathsf{f}_\lambda - f\|_n^2 + C(1+a+a^{-1})(L^2+\sigma^2)\psi_n^C(M),
\end{aligned}
$$

where

$$
\psi_n^C(M) = \begin{cases} M/n & \text{if } M \leq \sqrt{n},\\ \sqrt{\{\log(eM/\sqrt{n})\}/n} & \text{if } M > \sqrt{n}. \end{cases}
$$

PROOF. See appendix A. □

REMARK 1. Typically, the variance $\sigma^2 = \mathbb{E}_f W^2$ is unknown and we need to substitute an estimate in the penalty (2.3). We consider the situation described in the introduction where the functions f_j are estimators based on an independent (training) data set $\mathcal{D}'_\ell$ that consists of observations (X'_j, Y'_j) following (1.1). Let $\widehat{\sigma}^2$ be an estimate of σ^2 based on $\mathcal{D}'_\ell$ only. We write $\mathbb{E}_f^{(1)}$ ($\mathbb{E}_f^{(2)}$) for expectation with respect to $\mathcal{D}'_\ell$ ($\mathcal{D}_n$), respectively and let $\mathbb{E}_f = \mathbb{E}_f^{(1)}\mathbb{E}_f^{(2)}$ be the product expectation. Let $\widehat{f}$ be the aggregate corresponding to penalty (2.3) with σ^2 replaced by $\widehat{\sigma}^2$. Note that

$$
\mathbb{E}_f\|\widehat{f} - f\|_n^2 = \mathbb{E}_f^{(1)}\mathbb{E}_f^{(2)}\|\widehat{f} - f\|_n^2 I_{\{2\widehat{\sigma}^2\geq\sigma^2\}} + \mathbb{E}_f^{(1)}\mathbb{E}_f^{(2)}\|\widehat{f} - f\|_n^2 I_{\{2\widehat{\sigma}^2\geq\sigma^2\}}.
$$

Inspection of the proof of Theorem 3.1 shows that we may bound $\mathbb{E}_f^{(2)}\|\widehat{f}-f\|_n^2 I_{\{2\widehat{\sigma}^2\geq\sigma^2\}}$ simply by the right hand side of (3.1) with σ^2 substituted by $2\widehat{\sigma}^2$, as Theorem 3.1

holds for any penalty term larger than (2.3). Consequently we find

$$\mathbb{E}_f \|\widehat{f} - f\|_n^2 I_{\{2\widehat{\sigma}^2 \geq \sigma^2\}} \leq \frac{2\mathbb{E}_f^{(1)}\widehat{\sigma}^2}{n} \frac{6(1+a)^2}{a(e-1)}$$
$$+(1+a) \inf_{\lambda \in \mathbb{R}^M} \left[\mathbb{E}_f^{(1)} \|\mathsf{f}_\lambda - f\|_n^2 + \frac{2\mathbb{E}_f^{(1)}\widehat{\sigma}^2}{n} \left\{ 5 + \frac{2+3a}{a} L(\lambda) \right\} M(\lambda) \right].$$

Next, we observe that $\mathbb{E}_f^{(2)} \|\widehat{f} - f\|_n^2 \leq 6\sigma^2 + 2L^2$. For this, we use the reasoning leading to (A.4) in the proof of Theorem 4.1, in which we replace I_{A^c} by 1 throughout. Notice that this argument holds for any positive penalty term $\mathrm{pen}(\lambda)$ such that $\mathrm{pen}(\lambda_0) = 0$ with $\lambda_0 = (0, \ldots, 0)$, and hence it holds for the penalty term used here. Thus

$$\mathbb{E}_f \|\widehat{f} - f\|_n^2 I_{\{2\widehat{\sigma}^2 < \sigma^2\}} \leq \left(6\sigma^2 + 2L^2\right) \mathbb{P}_f^{(1)}\{2\widehat{\sigma}^2 < \sigma^2\}.$$

Combining the three displays above we see that $\widehat{f}$ achieves a bound similar to (3.1) if the estimator $\widehat{\sigma}^2$ satisfies $\mathbb{P}_f^{(1)}\{2\widehat{\sigma}^2 < \sigma^2\} \leq c_1/n$ and $\mathbb{E}_f^{(1)}\widehat{\sigma}^2 \leq c_2\sigma^2$, for some finite constants c_1, c_2. Since the sample variance based on $\mathcal{D}'_\ell$, with $\ell \geq cn$, for some positive constant c, meets both requirements, it can always play the role of $\widehat{\sigma}^2$.

4. Near optimal aggregation with a data dependent L_1 penalty.

In this section we show that the penalized least squares aggregate (2.2) using a penalty of the form (2.4) achieves simultaneously the (MS), (L), and (C) bounds of the form (1.4) with near optimal rates $\Delta_{n,M} = \bar{\psi}_{n,M}$. We require the following additional assumption.

ASSUMPTION (A3) *Define the matrices*

$$\Psi_n = \left(\frac{1}{n} \sum_{i=1}^n f_j(X_i) f_{j'}(X_i) \right)_{1 \leq j,j' \leq M}, \quad \mathrm{diag}(\Psi_n) = \mathrm{diag}(\|f_1\|_n^2, \ldots, \|f_M\|_n^2).$$

There exists $\kappa = \kappa_{n,M} > 0$ such that the matrix $\Psi_n - \kappa\,\mathrm{diag}(\Psi_n)$ is positive semi-definite for any given $n \geq 1$, $M \geq 2$.

Note that this assumption does not exclude the matrices Ψ_n whose ordered eigenvalues can be arbitrarily close to 0 as $M \to \infty$. Degenerate matrices Ψ_n are not excluded neither.

The next theorem presents an oracle inequality similar to the one of Theorem 3.1.

Theorem 4.1. *Assume (A1), (A2) and (A3). Let $\widetilde{f}$ be the penalized least squares aggregate defined by (2.2) with penalty (2.4). Then, for all $\varepsilon > 0$, and all integers $n \geq 1$, $M \geq 2$, we have,*

$$\begin{aligned} &\mathbb{E}_f \|\widetilde{f} - f\|_n^2 \\ &\leq \inf_{\lambda \in \mathbb{R}^M} \left\{ (1+\varepsilon)\|\mathsf{f}_\lambda - f\|_n^2 + \left(32 + 8\varepsilon + \frac{32}{\varepsilon}\right) \frac{\sigma^2}{\kappa} \frac{\log M + \log n}{n} M(\lambda) \right\} \\ &\quad + \frac{4L^2 + 12\sigma^2}{n\sqrt{\pi(\log M + \log n)}} + 6\sigma^2 \sqrt{\frac{n+2}{n}} \exp\left(-\frac{n}{16}\right). \end{aligned} \tag{4.1}$$

Proof. See appendix A. □

Corollary 4.2. *Let assumptions of Theorem 4.1 be satisfied. Then there exists a constant $C = C(L^2, \sigma^2, \kappa) > 0$ such that for all $\varepsilon > 0$ and for all integers $n \geq 1$, $M \geq 2$ and $D \leq M$,*

$$\begin{aligned} \mathbb{E}_f \|\widetilde{f} - f\|_n^2 &\leq (1+\varepsilon) \inf_{1 \leq j \leq M} \|f_j - f\|_n^2 + C\left(1 + \varepsilon + \varepsilon^{-1}\right) \frac{\log(M \vee n)}{n}. \\ \mathbb{E}_f \|\widetilde{f} - f\|_n^2 &\leq (1+\varepsilon) \inf_{\lambda \in \Lambda^{M,D}} \|\mathsf{f}_\lambda - f\|_n^2 + C\left(1 + \varepsilon + \varepsilon^{-1}\right) \frac{D\log(M \vee n)}{n}. \\ \mathbb{E}_f \|\widetilde{f} - f\|_n^2 &\leq (1+\varepsilon) \inf_{\lambda \in \mathbb{R}^M} \|\mathsf{f}_\lambda - f\|_n^2 + C\left(1 + \varepsilon + \varepsilon^{-1}\right) \frac{M\log(M \vee n)}{n}. \\ \mathbb{E}_f \|\widetilde{f} - f\|_n^2 &\leq (1+\varepsilon) \inf_{\lambda \in \Lambda^M} \|\mathsf{f}_\lambda - f\|_n^2 + C\left(1 + \varepsilon + \varepsilon^{-1}\right) \overline{\psi}_n^C(M), \end{aligned}$$

where

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

Proof. The argument is similar to that of the proof of Corollary 3.2. □

Remark 2. Using exactly the same reasoning as in Remark 1, we can replace σ^2 in the penalty term by twice the sample variance based on $\mathcal{D}'_\ell$.

REMARK 3. Inspection of the proofs shows that the constants $C = C(L^2, \sigma^2, \kappa)$ in Corollary 4.2 have the form $C = A_1 + A_2/\kappa$, where A_1 and A_2 are constants independent of κ. In general, κ may depend on n and M. However, if $\kappa > c$ for some constant $c > 0$, independent of n and M, as discussed in Remarks 3 and 4 below, the rates of aggregation given in Corollary 4.2 are near optimal, up to logarithmic factors. They are exactly optimal (*cf.* (1.3) and the lower bounds of the next section) for some configurations of n, M: for (MS)-aggregation if $n^{a'} \leq M \leq n^a$, and for (C)-aggregation if $n^{1/2} \leq M \leq n^a$, where $0 < a' < a < \infty$.

REMARK 4. If ξ_{min}, the smallest eigenvalue of the matrix Ψ_n, is positive, (A3) is satisfied for $\kappa = \xi_{min}/L^2$. In the standard parametric regression context where M is fixed and Ψ_n/n converges to a nonsingular $M \times M$ matrix, $\xi_{min} > c$ – and therefore $\kappa > c/L^2$ – for some $c > 0$, independent of M and n.

REMARK 5. Assumption (A3) is trivially satisfied with $\kappa = 1$ if Ψ_n is a diagonal matrix. An example illustrating this situation is related to the orthogonal series nonparametric regression: $M = M_n$ is allowed to converge to ∞ as $n \to \infty$ and the basis functions f_j are orthogonal with respect to the empirical norm. Another example is related to sequence space models, where the estimators $f_j = \widehat{f}_j$ are constructed from non-intersecting blocks of coefficients. Aggregating such mutually orthogonal estimators leads to adaptive estimators with good asymptotic properties [*cf.*, *e.g.*, Nemirovski (2000)].

5. Lower bounds. In this section we provide lower bounds showing that the remainder terms in the upper bounds obtained in the previous sections are optimal or near optimal. For regression with random design and the $L_2(\mathbb{R}^d, \mu)$-risks, such lower bounds for aggregation with optimal rates $\psi_{n,M}$ as given in (1.3) were established by Tsybakov (2003). The next theorem extends them to aggregation for the regression model with fixed design. Furthermore we state these bounds in a more general form, considering not only the expected squared risks, but also other loss functions, and instead of the (MS) aggregation bound, we provide the more general (S) aggregation bound.

Let $w : \mathbb{R} \to [0, \infty)$ be a *loss function, i.e.*, a monotone non-decreasing function satisfying $w(0) = 0$ and $w \not\equiv 0$.

THEOREM 5.1. *Let the integers n, M, D be such that $(2 \vee D) \leq M \leq n$, and let $X_1, \ldots, X_n$ be distinct points. Assume that H^M is either the whole $\mathbb{R}^M$ (for the (L) aggregation case), or the simplex Λ^M (for the (C) aggregation case), or the set $\Lambda^{M,D}$ (for the (S) aggregation case). Let the corresponding $\psi_{n,M}$ be given by (1.3) for (L) and (C) aggregation and, for (S) aggregation, let*

$$\psi_{n,M} = \frac{D}{n} \log\left(\frac{M}{D} + 1\right)$$

with $M \log(M/D + 1) \leq n$ and either $D = 1$ or $M \geq 6D$. Then there exist $f_1, \ldots, f_M \in \mathcal{F}_0$ such that

$$\inf_{T_n} \sup_{f \in \mathcal{F}_0} \mathbb{E}_f w\Big[\psi_{n,M}^{-1}\Big(\|T_n - f\|_n^2 - \inf_{\lambda \in H^M} \|\mathsf{f}_\lambda - f\|_n^2\Big)\Big] \geq c, \tag{5.1}$$

where $\inf_{T_n}$ denotes the infimum over all estimators and the constant $c > 0$ does not depend on n, M and D.

PROOF. See appendix A. □

Setting $w(u) = u$ in Theorem 5.1 we get the lower bounds for expected squared risks showing optimality or near optimality of the remainder terms in the oracle inequalities of Corollaries 3.2 and 4.2. The choice of $w(u) = I\{u > a\}$ with some fixed $a > 0$ leads to the lower bounds for probabilities showing near optimality of the remainder terms in the corresponding upper bounds "in probability" obtained in Bunea, Tsybakov and Wegkamp (2004).

APPENDIX A: PROOFS

A.1. Proof of Theorem 3.1. We define Λ_m as the set of elements in $\mathbb{R}^M$ with exactly m non-zero coefficients,

$$\Lambda_m = \left\{ \lambda \in \mathbb{R}^M : \; M(\lambda) = m \right\}.$$

Let $J_{m,k},\; k = 1, \ldots, \binom{M}{m}$, be all the subsets of $\{1, \ldots, M\}$ of cardinality m and we define

$$\Lambda_{m,k}(\lambda) = \{ \lambda = (\lambda_1, \ldots, \lambda_M) \in \Lambda_m : \; \lambda_j \neq 0 \Leftrightarrow j \in J_{m,k} \}.$$

Thus the collection $\left\{ \Lambda_{m,k} : \; 1 \leq k \leq \binom{M}{m} \right\}$ forms a partition of the set Λ_m. Next we observe that

$$\inf_{\lambda \in \mathbb{R}^M} \left\{ \widehat{S}(\lambda) + \mathrm{pen}(\lambda) \right\} = \inf_{0 \leq m \leq M} \inf_{1 \leq k \leq \binom{M}{m}} \inf_{\lambda \in \Lambda_{m,k}} \left\{ \widehat{S}(\lambda) + \mathrm{pen}(\lambda) \right\}$$

and the penalty

$$\mathrm{pen}(\lambda) \quad = \quad \frac{2\sigma^2}{n} \left\{ 1 + \frac{2+a}{1+a} \sqrt{L(\lambda)} + \frac{1+a}{a} L(\lambda) \right\} M(\lambda)$$

is the same for all $\lambda \in \Lambda_m$ as $M(\lambda) = m$ and $L(\lambda) = L_m \equiv 2 \ln \left(eM/(m \vee 1) \right)$ for all $\lambda \in \Lambda_m$.

We are now in the position to apply Theorem 2 in Birgé and Massart (2001b). This result implies that (setting their parameters $\theta = a/(1+a)$ and $K = 2$)

$$\begin{aligned} \mathbb{E}_f \| \widetilde{f} - f \|_n^2 \quad &\leq \quad (1+a) \inf_{0 \leq m \leq M} \inf_{1 \leq k \leq \binom{M}{m}} \left\{ \inf_{\lambda \in \Lambda_{m,k}} \| \mathsf{f}_\lambda - f \|_n^2 + \mathrm{pen}(\lambda) - \frac{m\sigma^2}{n} \right\} \\ &\quad + \frac{(1+a)^2}{a} \frac{\sigma^2}{n} \Sigma \left\{ \frac{(2+a)^2}{(1+a)^2} + 2 \right\}, \end{aligned}$$

where Σ is given by

$$\Sigma \quad = \quad \sum_{m=1}^{M} \sum_{k=1}^{\binom{M}{m}} \exp(-m L_m) = \sum_{m=1}^{M} \binom{M}{m} \exp(-m L_m).$$

Using the crude bound $\binom{M}{m} \leq (eM/m)^m$ [see, for example, Devroye *et al.* (1996), page 218], we may bound Σ by

$$\Sigma \leq \sum_{m=1}^{M} \left(\frac{eM}{m} \right)^{-m} \leq \sum_{m=1}^{M} e^{-m} \leq \frac{1}{e-1}.$$

For all $\lambda \in \Lambda_m$, we have

$$\begin{aligned} n\mathrm{pen}(\lambda) - m\sigma^2 &= \sigma^2 m \left(1 + 2\frac{2+a}{1+a}\sqrt{L_m} + 2\frac{1+a}{a}L_m\right) \\ &\le \sigma^2 m \left\{1 + \left(\frac{2+a}{1+a}\right)^2 + \left(1 + 2\frac{1+a}{a}\right)L_m\right\} \\ &\le \sigma^2 m \left(5 + \frac{2+3a}{a}L_m\right). \end{aligned}$$

Consequently we find

$$\begin{aligned} \mathbb{E}_f\|\widetilde{f} - f\|_n^2 &\le (1+a)\inf_{0\le m\le M}\inf_{1\le k\le\binom{M}{m}}\left\{\inf_{\lambda\in\Lambda_{m,k}}\|f - \mathsf{f}_\lambda\|_n^2 + \frac{\sigma^2 m}{n}\left(5 + \frac{2+3a}{a}L_m\right)\right\} \\ &\quad + \frac{(1+a)^2}{a}(4+2)\frac{\sigma^2}{n}\frac{1}{e-1} \\ &= (1+a)\inf_{\lambda\in\mathbb{R}^M}\left[\|f - \mathsf{f}_\lambda\|_n^2 + \frac{\sigma^2 M(\lambda)}{n}\left\{5 + \frac{2+3a}{a}L(\lambda)\right\}\right] + \frac{6(1+a)^2}{a(e-1)}\frac{\sigma^2}{n}, \end{aligned}$$

which proves the result. □

A.2. Proof of Corollary 3.2.

Proof of (3.2) and (3.3). Since the infimum on the right of (3.1) is taken over all $\lambda \in \mathbb{R}^M$, the (S) bound (3.3) easily follows by considering only the subset consisting of the $\binom{M}{D}$ vectors $\lambda \in \Lambda^{M,D}$ for which $M(\lambda) = D$ and $L(\lambda) = 2\log(eM/D) \le 6\log(M/D+1)$. The (MS) bound (3.2) is a special case of (3.3) for $D = 1$. □

Proof of (3.4). Since $x \mapsto x\log(eM/x)$ is increasing for $1 \le x \le M$,

$$\sup_{\lambda\in\mathbb{R}^M}\frac{M(\lambda)}{n}L(\lambda) = \frac{2M}{n}.$$

The result then follows from (3.1). □

Proof of (3.5). For $M \le \sqrt{n}$ the result follows from (3.4). Assume now that $M > \sqrt{n}$ and let m be the integer part of

$$x_{n,M} = \sqrt{n}\bigg/\sqrt{\log(eM/\sqrt{n})}.$$

Clearly, $0 \le m \le x_{n,M} \le M$. First, consider the case $m \ge 1$. Denote by $\mathcal{C}$ the set of functions h of the form

$$h(x) = \frac{1}{m}\sum_{j=1}^{M} k_j f_j(x), \;\; k_j \in \{0, 1, \dots, m\}, \;\; \sum_{j=1}^{m} k_j \le m.$$

The following approximation result can be obtained by the "Maurey argument" (see, for example, Barron (1993), Lemma 1, or Nemirovski (2000), pages 192, 193):

$$\min_{g\in\mathcal{C}} \|g - f\|_n^2 \le \min_{\lambda\in\Lambda^M} \|\mathsf{f}_\lambda - f\|_n^2 + \frac{L^2}{m}. \tag{A.1}$$

For completeness, we give the proof of (A.1) in the Appendix B. Since $M(\lambda) \le m \le x_{n,M}$ for the vectors λ corresponding to $g \in \mathcal{C}$, and since $x \mapsto x\log\left(\frac{eM}{x}\right)$ is increasing for $1 \le x \le M$, we get from (3.1)

$$\mathbb{E}_f\|\widetilde{f} - f\|_n^2 \le \inf_{g\in\mathcal{C}}\left\{C_1\|g - f\|_n^2 + C_2\frac{x_{n,M}}{n}\log\left(\frac{eM}{x_{n,M}}\right)\right\} + \frac{C_3}{n}$$

for some constants C_1, C_2 and C_3 depending on a and σ^2. Using this inequality, (A.1) and the fact that $m = \lfloor x_{n,M}\rfloor \ge x_{n,M}/2$ for $x_{n,M} \ge 1$, we obtain

$$\mathbb{E}_f\|\widetilde{f} - f\|_n^2 \quad\le\quad C_1 \inf_{\lambda\in\Lambda^M}\|\mathsf{f}_\lambda - f\|_n^2 + C_1\frac{2L^2}{x_{n,M}} + C_2\frac{x_{n,M}}{n}\log\left(\frac{eM}{x_{n,M}}\right) + \frac{C_3}{n}.$$

We use this bound for all choices of $\lambda \in \Lambda^M$ with $m \ge M(\lambda) \ne 0$. For $m = 0$, we only need to consider the singular case $\lambda = 0$ as $M(\lambda) = 0$ if and only if $\lambda = 0$. Note that for $m = 0$, we have the trivial upper bound $C_1\|f\|_n^2 + C_3 n^{-1}$ for the right-hand side of (3.1) and clearly $n^{-1} \le \psi_n^C(M)$. To complete the proof of the Corollary, we note that

$$\log\left(\frac{eM}{x_{n,M}}\right) = \log\left(\frac{eM}{\sqrt{n}}\sqrt{\log\left(\frac{eM}{\sqrt{n}}\right)}\right) \le 3\log\left(\frac{eM}{\sqrt{n}}\right),$$

in view of the elementary inequality $\log(y\sqrt{\log(y)}) \le 3\log(y)$, for all $y \ge 0$. □

A.3. Proof of Theorem 4.1. We begin as in Loubes and Van de Geer (2002). First we define

$$r_n = 2\sqrt{2}\sigma\sqrt{\frac{\log M + \log n}{n}}$$

and $r_{n,j} = r_n \|f_j\|_n$. By definition, $\widetilde{f} = \mathsf{f}_{\widehat{\lambda}}$ satisfies

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

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

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

We define the random variables

$$V_j = \frac{1}{n} \sum_{i=1}^{n} f_j(X_i) W_i, \quad 1 \le j \le M,$$

and the event

$$A = \bigcap_{j=1}^{M} \{2|V_j| \le r_{n,j}\}.$$

The normality assumption (A1) on W_i implies that $\sqrt{n}\, V_j \sim N\left(0, \sigma^2 \|f_j\|_n^2\right)$, $1 \le j \le M$. Applying the union bound followed by the standard tail bound for the $N(0,1)$ distribution, yields

$$\begin{aligned} \mathbb{P}(A^c) &\le \sum_{j=1}^{M} \mathbb{P}\{\sqrt{n}|V_j| > \sqrt{n} r_{n,j}/2\} \le \sum_{j=1}^{M} \frac{4}{\sqrt{2\pi}} \frac{\sigma \|f_j\|_n}{\sqrt{n} r_{n,j}} \exp\left(- \frac{n r_{n,j}^2}{8\sigma^2 \|f_j\|_n^2} \right) \\ &= \frac{1}{n\sqrt{\pi(\log M + \log n)}}. \end{aligned} \tag{A.2}$$

Then, on the set A, we find

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

and therefore, still on the set A,

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

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, on the set A,

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

by the triangle inequality and the fact that $\lambda_j = 0$ for $j \notin J(\lambda)$. By assumption (A3), we have

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

Combining this with the Cauchy-Schwarz and triangle inequalities, respectively, we find further that, on the set A,

$$\begin{aligned}\text{(A.3)}\quad \|\widetilde{f}-f\|_n^2 &\le \|\mathsf{f}_\lambda - f\|_n^2 + 2\sum_{j\in J(\lambda)} r_{n,j}|\widehat{\lambda}_j-\lambda_j|\\ &\le \|\mathsf{f}_\lambda - f\|_n^2 + 2r_n\sqrt{M(\lambda)/\kappa}\left(\|\widetilde{f}-f\|_n + \|\mathsf{f}_\lambda - f\|_n\right).\end{aligned}$$

Inequality (A.3) is of the simple form $v^2 \le c^2 + vb + cb$ with $v = \|\widetilde{f}-f\|_n$, $b = 2r_n\sqrt{M(\lambda)/\kappa}$ and $c = \|\mathsf{f}_\lambda - f\|_n$. 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 $v^2 \le v^2/(2\alpha) + \alpha\, b^2 + (2\alpha+1)/(2\alpha)\, c^2$, whence $v^2 \le a/(a-1)\{b^2(a/2) + c^2(a+1)/a\}$ for $a = 2\alpha > 1$. Recalling that (A.3) is valid on the set A, we now get that

$$\mathbb{E}_f\left[\|\widetilde{f}-f\|_n^2 I_A\right] \le \inf_{\lambda\in\mathbb{R}^M}\left\{\frac{a+1}{a-1}\|\mathsf{f}_\lambda - f\|_n^2 + \frac{2a^2}{\kappa(a-1)}r_n^2 M(\lambda)\right\}, \quad \forall\ a > 1.$$

It remains to bound $\mathbb{E}_f\|\widetilde{f}-f\|_n^2 I_{A^c}$. Writing $\|W\|_n^2 = n^{-1}\sum_{i=1}^n W_i^2$ and using the inequality $(x+y)^2 \le 2x^2+2y^2$, we find that

$$\begin{aligned}\mathbb{E}_f\|\widetilde{f}-f\|_n^2 I_{A^c} &\le 2\mathbb{E}_f\widehat{S}(\widetilde{f})I_{A^c} + 2\mathbb{E}_f\widehat{S}(f)I_{A^c}\\ &= 2\mathbb{E}_f\widehat{S}(\widetilde{f})I_{A^c} + 2\mathbb{E}_f\|W\|_n^2 I_{A^c}.\end{aligned}$$

Next, since $\mathrm{pen}(\widetilde{\lambda}) \ge 0$ and by the definition of $\widetilde{f}$, for $\lambda_0 = (0,\ldots,0)' \in \mathbb{R}^M$,

$$\begin{aligned}\mathbb{E}_f\widehat{S}(\widetilde{f})I_{A^c} &\le \mathbb{E}_f\left\{\widehat{S}(\widetilde{f}) + \mathrm{pen}(\tilde{\lambda})\right\}I_{A^c}\\ &\le \mathbb{E}_f\left\{\widehat{S}(f_{\lambda_0}) + \mathrm{pen}(\lambda_0)\right\}I_{A^c}\\ &= \mathbb{E}_f\widehat{S}(f_{\lambda_0})I_{A^c}\\ &\le 2\mathbb{E}_f\|f\|_n^2 I_{A^c} + 2\mathbb{E}_f\|W\|_n^2 I_{A^c}\\ &\le 2L^2\mathbb{P}(A^c) + 2\mathbb{E}_f\|W\|_n^2 I_{A^c},\end{aligned}$$

whence

$$\mathbb{E}_f\|\widetilde{f}-f\|_n^2 I_{A^c} \le 4L^2\mathbb{P}(A^c) + 6\mathbb{E}_f\|W\|_n^2 I_{A^c}. \tag{A.4}$$

In order to bound the last term on the right, we introduce the event

$$B = \left\{\frac{1}{n}\sum_{i=1}^n W_i^2 \le 2\sigma^2\right\}.$$

This set has probability larger than $1-\exp(-n/8)$ since

$$\begin{aligned}\mathbb{P}\{B^c\} &= \mathbb{P}\left\{\chi_n^2 - n > \sqrt{2n}\sqrt{n/2}\right\}\\ &\le \exp\left(-\frac{n/2}{2+2\sqrt{n/2}\sqrt{2/n}}\right)\\ &= \exp(-n/8)\end{aligned}$$

by Lemma B.2 below. Observe further that

$$\mathbb{E}_f\|W\|_n^2 I_{A^c} \le 2\sigma^2\mathbb{P}\{A^c\} + \mathbb{E}_f\|W\|_n^2 I_{B^c}$$

and by the Cauchy-Schwarz inequality we find

$$\begin{aligned}\mathbb{E}_f\|W\|_n^2 I_{B^c} &\le \left(\mathbb{E}_f\|W\|_n^4\right)^{1/2}\exp(-n/16)\\ &= \sqrt{\left(\frac{3\sigma^4}{n} + \frac{n-1}{n}\sigma^4\right)}\exp(-n/16).\end{aligned}$$

Collecting all these bounds, and using the bound (A.2) on $\mathbb{P}\{A^c\}$, we obtain

$$\begin{aligned}\mathbb{E}_f\|\widetilde{f}-f\|_n^2 I_{A^c} &\le 4L^2\mathbb{P}(A^c)+6\mathbb{E}_f\|W\|_n^2 I_{A^c}\\ &\le \frac{4L^2+12\sigma^2}{n\sqrt{\pi(\log M+\log n)}}+6\sigma^2\sqrt{\frac{n+2}{n}}\exp(-n/16).\end{aligned}$$

The proof of the theorem is complete by taking $\varepsilon = 2/(a-1)$. □

A.4. Proof of Theorem 5.1.

We proceed similarly to Tsybakov (2003). The proof is based on the following easy corollary of the Fano lemma [which can be obtained, for example, by combining Theorems 2.2 and 2.5 in Tsybakov (2004)].

LEMMA A.1. *Let w be a loss function, $A > 0$ be such that $w(A) > 0$, and let $\mathcal{C}$ be a set of functions on $\mathcal{X}$ of cardinality $N = \mathrm{card}(\mathcal{C}) \ge 2$ such that*

$$\|f-g\|_n^2 \ge 4s^2 > 0, \quad \forall\ f,g\in\mathcal{C}, \quad f\ne g,$$

and the Kullback divergences $K(\mathbb{P}_f,\mathbb{P}_g)$ between the measures $\mathbb{P}_f$ and $\mathbb{P}_g$ satisfy

$$K(\mathbb{P}_f,\mathbb{P}_g)\le(1/16)\log N, \quad \forall\ f,g\in\mathcal{C}.$$

Then for $\psi = s^2/A$ we have

$$\inf_{T_n}\sup_{f\in\mathcal{C}}\mathbb{E}_f w\Big[\psi^{-1}\|T_n-f\|_n^2\Big]\ge c_1 w(A),$$

where $\inf_{T_n}$ denotes the infimum over all estimators and $c_1 > 0$ is a constant.

The (S) aggregation case. Pick M disjoint subsets $S_1,\dots,S_M$ of $\{X_1,\dots,X_n\}$, each S_j of cardinality $\log(M/D+1)$ (w.l.o.g. we assume that $\log(M/D+1)$ is an integer) and define the functions

$$f_j(x)=\gamma I_{\{x\in S_j\}}, \quad j=1,\dots,M,$$

where $\gamma \le L$ is a positive constant to be chosen. Consider the set of functions $\mathcal{V} = \{\mathsf{f}_\lambda : \lambda\in\Lambda^{M,D}\}$. Clearly, $\mathcal{V}\subset\mathcal{F}_0$. Thus, it suffices to prove the (S) lower bound of the theorem where the supremum over $f \in \mathcal{F}_0$ is replaced by that over $f \in \mathcal{V}$. But for $f \in \mathcal{V}$ we have $\min_{\lambda \in \Lambda^{M,D}} \|\mathsf{f}_\lambda - f\|_n^2 = 0$, and therefore to finish the proof for the (S) case, it suffices to bound from below the quantity $\inf_{T_n} \sup_{f \in \mathcal{V}} \mathbb{E}_f w(\psi_{n,M}^{-1} \|T_n - f\|_n^2)$ where $\psi_{n,M} = D \log(M/D+1)/n$. This will be done by applying Lemma A.1. In fact, note that for every two functions f_λ and $\mathsf{f}_{\bar\lambda}$ in $\mathcal{V}$ we have

$$\|\mathsf{f}_\lambda - \mathsf{f}_{\bar\lambda}\|_n^2 = \frac{\gamma^2 \log(M/D+1)}{n} \rho(\lambda, \bar\lambda) \tag{A.5}$$

and $\rho(\lambda,\bar\lambda) \le D$, where $\rho(\lambda,\bar\lambda) \overset{\text{def}}{=} \sum_{j=1}^M I_{\{\lambda_j \ne \bar\lambda_j\}}$ is the Hamming distance between $\lambda = (\lambda_1, \dots, \lambda_M) \in \Lambda^{M,D}$ and $\bar\lambda = (\bar\lambda_1, \dots, \bar\lambda_M) \in \Lambda^{M,D}$. Lemma 4 in Birgé and Massart (2001a) (see also Gilbert (1952)) asserts that if $M \ge 6D$ there exists a subset $\Lambda' \subset \Lambda^{M,D}$ such that, for some constant $\tilde c > 0$ independent of M and D,

$$\log \operatorname{card}(\Lambda') \ge \tilde c D \log\left(\frac{M}{D} + 1\right) \tag{A.6}$$

and

$$\rho(\lambda, \bar\lambda) \ge \tilde c D, \quad \forall\ \lambda, \bar\lambda \in \Lambda', \quad \lambda \ne \bar\lambda. \tag{A.7}$$

Consider a set of functions $\mathcal{C} = \{\mathsf{f}_\lambda : \lambda \in \Lambda'\} \subset \mathcal{V}$. From (A.5) and (A.7), for any two functions f_λ and $\mathsf{f}_{\bar\lambda}$ in $\mathcal{C}$ we have

$$\|\mathsf{f}_\lambda - \mathsf{f}_{\bar\lambda}\|_n^2 \ge \frac{\tilde c \gamma^2 D \log(M/D+1)}{n} \overset{\text{def}}{=} 4s^2. \tag{A.8}$$

Since W_j's are $N(0,\sigma^2)$ random variables, the Kullback divergence $K(\mathbb{P}_{\mathsf{f}_\lambda}, \mathbb{P}_{\mathsf{f}_{\bar\lambda}})$ between $\mathbb{P}_{\mathsf{f}_\lambda}$ and $\mathbb{P}_{\mathsf{f}_{\bar\lambda}}$ satisfies

$$K(\mathbb{P}_{\mathsf{f}_\lambda}, \mathbb{P}_{\mathsf{f}_{\bar\lambda}}) = \frac{n}{2\sigma^2} \|\mathsf{f}_\lambda - \mathsf{f}_{\bar\lambda}\|_n^2, \quad j = 1, \dots, M. \tag{A.9}$$

In view of (A.5) and (A.9), one can choose γ small enough to have

$$K(\mathbb{P}_{\mathsf{f}_\lambda}, \mathbb{P}_{\mathsf{f}_{\bar\lambda}}) \le \frac{1}{16\tilde c} D \log\left(\frac{M}{D}+1\right) \le \frac{1}{16} \log \operatorname{card}(\Lambda') = \frac{1}{16} \log \operatorname{card}(\mathcal{C})$$

for all $\lambda, \bar\lambda \in \Lambda'$. Now, to get the lower bound for the (S) case, it remains to use this inequality together with (A.8), and to apply Lemma A.1. Note that the above argument holds through under the assumption that $M \geq 6D$ which is needed to assure (A.6). In the remaining case where $D = 1, M < 6D$ we have $\psi_{n,M} \leq (\log 7)/n$, and we define the set $\mathcal{C} = \{\mathsf{f}_{\lambda'}, \mathsf{f}_{\lambda''}\}$ with $\lambda' = (1, 0, \ldots, 0) \in \Lambda^M$ and $\lambda'' = (0, \ldots, 0, 1) \in \Lambda^M$. Then $\|\mathsf{f}_{\lambda'} - \mathsf{f}_{\lambda''}\|_n^2 = 2\gamma^2 \log(M+1)/n \geq 2\gamma^2(\log 3)/n$, and the result easily follows from (A.9) and Lemma A.1.

The (C) aggregation case. Consider the orthonormal trigonometric basis in $L_2[0,1]$ defined by $\phi_1(x) \equiv 1$, $\phi_{2k}(x) = \sqrt{2}\cos(2\pi k x)$, $\phi_{2k+1}(x) = \sqrt{2}\sin(2\pi k x), k = 1, 2, \ldots$, for $x \in [0,1]$. Set

$$f_j(x) = \gamma \sum_{k=1}^{n} \phi_j(k/n) I_{\{x = X_k\}}, \quad j = 1, \ldots, M, \tag{A.10}$$

where $\gamma \leq L/\sqrt{2}$ is a positive constant to be chosen. The system of functions $\{\phi_j\}_{j=1,\ldots,M}$ is orthonormal w.r.t. the discrete measure that assigns the mass $1/n$ to each of the points $k/n, k = 1, \ldots, n$:

$$\frac{1}{n}\sum_{k=1}^{n} \phi_j(k/n)\phi_l(k/n) = \delta_{jl}, \quad j, l = 1, \ldots, n,$$

where δ_{jl} is the Kronecker delta (see, e.g., Tsybakov (2004), p.45, Lemma 1.9). Hence

$$< f_j, f_l >_n = \gamma^2 \delta_{jl}, \quad j, l = 1, \ldots, M, \tag{A.11}$$

where $< \cdot, \cdot >_n$ stands for the scalar product induced by $\|\cdot\|_n$.

Assume first that $M > \sqrt{n}$ (i.e., we are in the "sparse" case). Define an integer

$$m = \left\lceil c_2 \Big[n/\log\Big(\frac{M}{\sqrt{n}} + 1\Big)\Big]^{1/2} \right\rceil$$

for a constant $c_2 > 0$ chosen in such a way that $M \geq 6m$. Consider the finite set $\mathcal{C} \subset \Lambda^M$ composed of such convex combinations of $f_1, \ldots, f_M$ that m of the coefficients λ_j are equal to $1/m$ and the remaining $M - m$ coefficients are zero. In view of (A.11), for every pair of functions $g_1, g_2 \in \mathcal{C}$ we have

$$\|g_1 - g_2\|_n^2 \leq 2\gamma^2/m. \tag{A.12}$$

To finish the proof for $M > \sqrt{n}$ it suffices now to apply line by line the argument after the formula (10) in Tsybakov (2003) replacing there $\|\cdot\|$ by $\|\cdot\|_n$. Similarly, the proof for $M \le \sqrt{n}$ is analogous to that given in Tsybakov (2003), with the only difference that the functions f_j should be chosen as in (A.10) and $\|\cdot\|$ should be replaced by $\|\cdot\|_n$.

The (L) aggregation case. Let $H^M = \mathbb{R}^M$ and $\psi_{n,M} = M/n$. Define the functions $f_j(x) = \gamma I_{\{x=X_j\}}, \quad j = 1,\dots,M$, with $0 < \gamma \le L$ and introduce a finite set of their linear combinations

$$\mathcal{U} = \Big\{ g = \sum_{j=1}^{M} \omega_j f_j : \omega \in \Omega \Big\}, \tag{A.13}$$

where Ω is the set of all vectors $\omega \in \mathbb{R}^M$ with binary coordinates $\omega_j \in \{0,1\}$. Since the supports of f_j's are disjoint, the functions $g \in \mathcal{U}$ are uniformly bounded by γ, thus $\mathcal{U} \subset \mathcal{F}_0$. Clearly, $\min_{\lambda\in\mathbb{R}^M} \|\mathsf{f}_\lambda - f\|_n^2 = 0$ for any $f \in \mathcal{U}$. Therefore, similarly to the (MS) case, it is sufficient to bound from below the quantity $\sup_{f\in\mathcal{U}} \mathbb{E}_f w(\psi_{n,M}^{-1}\|T_n - f\|_n^2)$ where $\psi_{n,M} = M/n$, uniformly over all estimators T_n.

Note that for any $g_1 = \sum_{j=1}^M \omega_j f_j \in \mathcal{U}$ and $g_2 = \sum_{j=1}^M \omega'_j f_j \in \mathcal{U}$ we have

$$\|g_1 - g_2\|_n^2 = \frac{\gamma^2}{n}\sum_{j=1}^{M}(\omega_j - \omega'_j)^2 \le \gamma^2 M/n. \tag{A.14}$$

Let first $M \ge 8$. Then it follows from the Varshamov-Gilbert bound (see Gilbert (1952) or Tsybakov (2004), Chapter 2) that there exists a subset $\mathcal{C}'$ of $\mathcal{U}$ such that $\mathrm{card}(\mathcal{U}_0) \ge 2^{M/8}$ and

$$\|g_1 - g_2\|_n^2 \ge C_1\gamma^2 M/n. \tag{A.15}$$

for any $g_1, g_2 \in \mathcal{C}'$. Using (A.9) and (A.14) we get, for any $g_1, g_2 \in \mathcal{C}'$,

$$K(\mathbb{P}_{g_1}, \mathbb{P}_{g_2}) \le C_2\gamma^2 M \le C_3\gamma^2 \log(\mathrm{card}(\mathcal{C}')),$$

and by choosing γ small enough, we can finish the proof in the same way as in the (MS) case. If $2 \le M \le 8$, we have $\psi_{n,M} \le 8/n$, and the proof is easily obtained by choosing $f_1 \equiv 0$ and $f_2 \equiv \gamma n^{-1/2}$ and applying Lemma A.1 to the set $\mathcal{C}' = \{f_1, f_2\}$. □

APPENDIX B: TECHNICAL LEMMAS

LEMMA B.1. *Let $f, f_1, \dots, f_M \in \mathcal{F}_0$ and $1 \le m \le M$. Let $\mathcal{C}$ be the finite set of functions defined in the proof of 3.5. Then (A.1) holds and*

$$\min_{g\in\mathcal{C}} \|g - f\|_n^2 \le \min_{\lambda\in\Lambda^M} \|\mathsf{f}_\lambda - f\|_n^2 + \frac{L^2}{m}. \tag{B.1}$$

PROOF. Let f^* be a minimizer of $\|\mathsf{f}_\lambda - f\|_n^2$ over $\lambda \in \Lambda^M$. Clearly, f^* is of the form

$$f^* = \sum_{j=1}^M p_j f_j \quad \text{with } p_j \ge 0 \quad \text{and} \quad \sum_{j=1}^M p_j \le 1.$$

Define a probability distribution on $j = 0, 1, \dots, M$ by

$$\pi_j = \begin{cases} p_j & \text{if } j \ne 0, \\ 1 - \sum_{j=1}^M p_j & \text{if } j = 0. \end{cases}$$

Consider m i.i.d. random integers $j_1, \dots, j_m$ where each j_k is distributed according to $\{\pi_j\}$ on $\{0, 1, \dots, M\}$. Introduce the random function

$$\bar{f}_m = \frac{1}{m}\sum_{k=1}^m g_{j_k}$$

where

$$g_j = \begin{cases} f_j & \text{if } j \ne 0, \\ 0 & \text{if } j = 0. \end{cases}$$

For every $x \in \mathcal{X}$ the random variables $g_{j_1}(x), \dots, g_{j_m}(x)$ are i.i.d. with $\mathbb{E}(g_{j_k}(x)) = f^*(x)$. Thus,

$$\begin{aligned} \mathbb{E}(\bar{f}_m(x) - f^*(x))^2 &= \mathbb{E}\left(\left[\frac{1}{m}\sum_{k=1}^m \{g_{j_k}(x) - \mathbb{E}(g_{j_k}(x))\}\right]^2\right) \\ &\le \frac{1}{m}\mathbb{E}(g_{j_1}^2(x)) \le \frac{L^2}{m}\ . \end{aligned}$$

Hence for every $x \in \mathcal{X}$ and every $f \in \mathcal{F}_0$ we get

$$\begin{aligned} \mathbb{E}(\bar{f}_m(x) - f(x))^2 &= \mathbb{E}(\bar{f}_m(x) - f^*(x))^2 + (f^*(x) - f(x))^2 \\ &\leq \frac{L^2}{m} + (f^*(x) - f(x))^2. \end{aligned} \tag{B.2}$$

Integrating (B.2) over the empirical probability measure that puts mass $1/n$ at each X_i, and recalling the definition of f^* we obtain

$$\mathbb{E}\|\bar{f}_m - f\|_n^2 \leq \min_{\lambda \in \Lambda^M} \|f_\lambda - f\|_n^2 + \frac{L^2}{m}. \tag{B.3}$$

Finally, note that the random function $\bar{f}_m$ takes its values in $\mathcal{C}$, which implies that

$$\mathbb{E}\|\bar{f}_m - f\|_n^2 \geq \min_{g \in \mathcal{C}} \|g - f\|_n^2.$$

This and (B.3) prove (B.1). □

LEMMA B.2. *Let Z_d denote a random variable having the χ^2 distribution with d degrees of freedom. Then for all $x > 0$,*

$$\mathbb{P}\{Z_d - d \geq x\sqrt{2d}\} \leq \exp\left(-\frac{x^2}{2(1 + x\sqrt{2/d})}\right). \tag{B.4}$$

PROOF. See Cavalier *et al.* (2002, page 857). □

REFERENCES

[1] AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions in Automatic Control*, AC-19, 716-723.

[2] ANTONIADIS, A. and FAN, J. (2001). Regularized wavelet approximations (with discussion). *Journal of American Statistical Association*, 96, 939-967.

[3] AUDIBERT, J.-Y. (2004) Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 40: 685-736.

[4] BARAUD, Y. (2000). Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117: 467 – 493.

[5] BARAUD, Y. (2002). Model selection for regression on a random design. *ESAIM Probability & Statistics*, 7: 127 – 146.

[6] BARRON, A.R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39: 930 – 945.

[7] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113: 301 – 413.

[8] Bartlett, P.L., Boucheron, S. and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning* 48: 85 – 113.

[9] Birgé, L. (2003). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Prépublication n.862, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7. `http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2003`.

[10] Birgé, L. and Massart, P. (2001a). Gaussian model selection. *Journal of the European Mathematical Society*, **3**(3), 203-268.

[11] Birgé, L. and Massart, P. (2001b). A generalized C_p criterion for Gaussian model selection. Prépublication 647, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7. `http://www.proba.jussieu.fr/mathdoc/preprints/ index.html#2001`.

[12] Bunea, F. (2004). Consistent covariate selection and postmodel selection inference in semiparametric regression. *Annals of Statistics*, 32: 898-927.

[13] Bunea, F. and Nobel, A.B. (2005). Sequential Procedures for Aggregating Arbitrary Estimators of a Conditional Mean. Preprint Florida State University `www.stat.fsu.edu/∼flori`

[14] Bunea, F., Tsybakov, A. and Wegkamp, M.H. (2004). Aggregation for regression learning. Available at arXiv:math.ST/0410214, 8 Oct. 2004. Prépublication n.948, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7. `http://www.proba.jussieu.fr/mathdoc/preprints/ index.html#2004`.

[15] Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization.* Ecole d'Eté de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics **1851**, Springer, New York.

[16] Cavalier L., Golubev G.K., Picard D. and Tsybakov A.B. (2002) Oracle inequalities for inverse problems. *Annals of Statistics*, 30: 843 – 874.

[17] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

[18] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32: 407 – 499.

[19] Fan, J. and Li, R.Z. (2001). Variable selection via penalized likelihood. *Journal of American Statistical Association*, 1348-1360.

[20] Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Annals of Statistics*, 32, 928-961.

[21] Gilbert, E.N. (1952) A comparison of signalling alphabets. *Bell System Tech.J.*, 31:504-522.

[22] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York.

[23] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, Approximation and Statistical Applications.* Lecture Notes in Statistics, vol. 129. Springer, New York.

[24] Juditsky, A., Nazin, A., Tsybakov, A. and Vayatis, N. (2005a) Recursive aggregation of estimators via the Mirror Descent Algorithm with averaging. *Problems of Information Transmission*, 41 (4). www.proba.jussieu.fr/pageperso/vayatis/publication.html

[25] Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Annals of Statistics*, 28:681–712.

[26] Juditsky, A., Rigollet, Ph. and Tsybakov, A. (2005b) Learning by mirror averaging. Prépublication du Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7. http://www.proba.jussieu.fr/mathdoc/preprints.

[27] Kneip, A. (1994). Ordered linear smoothers. *Annals of Statistics*, 22: 835-866.

[28] Koltchinskii, V. (2004). Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, to appear.

[29] Leung, G. and Barron, A.R. (2004) Information theory and mixing least-squares regressions. Manuscript.

[30] Loubes, J.-M. and van de Geer, S.A. (2002). Adaptive estimation in regression, using soft thresholding type penalties. *Statistica Neerlandica* 56: 453 – 478.

[31] Lugosi, G. and Nobel, A. (1999). Adaptive model selection using empirical complexities. *Annals of Statistics*, 27: 1830 – 1864.

[32] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15, 661-675.

[33] Nemirovski, A. (2000). Topics in non-parametric statistics. In P. Bernard, editor, *Ecole d'Eté de Probabilités de Saint-Flour 1998*, volume XXVIII of *Lecture Notes in Mathematics*. Springer, New York.

[34] Rao, C. R. and Wu, Y. (2001). On model selection. *IMS Lecture notes-Monograph Series, P. Lahiri editor*, 38, 1 – 65.

[35] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

[36] Tsybakov, A.B. (2003). Optimal rates of aggregation. In *Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines. Lecture Notes in Artificial Intelligence*, v. 2777, p.303–313. Springer-Verlag, Heidelberg.

[37] Tsybakov, A. B. (2004). *Introduction à l'estimation non–paramétrique.* Springer, Berlin.

[38] Wegkamp, M.H. (2003). Model selection in nonparametric regression. *Annals of Statistics*, 31: 252 – 273.

[39] Yang, Y. (2000). Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74: 135 – 161.

[40] Yang, Y. (2001). Adaptive regression by mixing. *Journal of American Statistical Association*,

96: 574 – 588.

[41] Yang, Y. (2004). Aggregating regression procedures for a better performance. *Bernoulli*, 10: 25 – 47.

Department of Statistics
Florida State University
Tallahassee, Florida
E-mail: {bunea,wegkamp}@stat.fsu.edu

Laboratoire de Probabilités et Modèles Aléatoires
Université Paris VI
France
e-mail: tsybakov@ccr.jussieu.fr