AGGREGATION FOR REGRESSION LEARNING

Florentina Bunea, Alexandre B. Tsybakov, Marten H. Wegkamp

AGGREGATION FOR REGRESSION LEARNING

FLORENTINA BUNEA[†1], ALEXANDRE B. TSYBAKOV, AND MARTEN H. WEGKAMP[†]

ABSTRACT. This paper studies statistical aggregation procedures in 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 separately for each one of (MS), (C) and (L) cases. 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. We propose such a procedure which consists in mixing of the initial estimators with the weights obtained by penalized least squares. Two different penalities 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 of independent random pairs (X_i, Y_i) with

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

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 either random vectors with probability measure μ supported on $\mathcal{X}$ or fixed elements in $\mathcal{X}$, and the errors W_i are zero mean random variables, conditionally on the X_i's.

Aggregation of arbitrary estimators in regression models has recently received increasing attention: Nemirovski (2000), Juditsky and Nemirovski (2000), Yang (2000, 2001, 2004), Catoni (2004), Györfi *et al.* (2002), Wegkamp (2003), Tsybakov (2003), Birgé (2003). A

Date: October 2004.

1991 *Mathematics Subject Classification.* Primary 62G08, Secondary 62C20, 62G05, 62G20.

Key words and phrases. aggregation, minimax risk, model selection, nonparametric regression, oracle inequalities, penalized least squares, statistical learning.

[1] Corresponding author.

[†]Research partially supported by NSF grant DMS 0406049 .

motivating factor is the existence of many different methods of estimation, leading to possibly competing estimators. Local polynomial kernel smoothing methods and penalized least squares or likelihood estimators (which include B-splines and wavelet type estimators) are two classes of methods that cover the major trends in nonparametric estimation in regression. When no method is a clear winner, one may prefer to combine different estimators obtained via different methods. Furthermore, within each method one can obtain competing estimators for different values of the smoothing parameter (the bandwidth in kernel procedures and, for the other examples, the calibrating constant in the penalty term or, correspondingly, the threshold value). This is usually the case when adaptive estimation is considered. In all these situations we are faced with a large collection of concurrent estimators $\widehat{f}_1, \ldots, \widehat{f}_M$. A natural idea is then to look for a new, improved, estimator $\widetilde{f}$ constructed by combining $\widehat{f}_1, \ldots, \widehat{f}_M$ 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.

In this paper we consider a more general setup for the (MS), (C) and (L) aggregation problems, following Tsybakov (2003). Namely, we do not restrict aggregates to be of the form of model selectors, convex or linear combinations of the original estimators. Instead, we only require that aggregates should be estimators that mimic the model selection, convex or linear oracles. This allows us to construct more powerful aggregates. To give precise definitions, denote by $\|g\| = \left(\int g^2(x)\mu(dx)\right)^{1/2}$ the norm of a function g in $L_2(\mathbb{R}^d, \mu)$ and set $\mathsf{f}_\lambda = \sum_{j=1}^M \lambda_j \widehat{f}_j$ for any $\lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M$. The performance of an aggregate $\widetilde{f}$ used to estimate a function $f \in L_2(\mathbb{R}^d, \mu)$ can be judged against the following mathematical target:

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

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 M vertices of Λ^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 $\inf_{\lambda \in H^M} \mathbb{E}_f \|\mathsf{f}_\lambda - f\|^2$ 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 say that the aggregate $\widetilde{f}$ mimics the (L), (C) or (MS) 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 $\psi_{n,M}$. As shown in Tsybakov (2003), for the Gaussian regression model we have, under mild conditions

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

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:

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

since the sets over which the infima are taken are nested. Thus, the bound (1.2) for (MS) aggregation realizes the trade-off between the largest oracle risk and the smallest remainder term. The bound (1.2) for (L) aggregation realizes the trade-off between the smallest oracle risk and the largest remainder term. The bound (1.2) for (C) aggregation realizes the trade-off between an intermediate oracle risk and intermediate remainder term. If the number of estimators to be aggregated is small, $M \le \sqrt{n}$, the remainder term in the (C) bound is identical to that in the (L) bound, but the oracle risk in the (L) bound is always superior to that in the (C) bound. Thus (L) aggregation is preferable to (C) aggregation in this case, but no comparison can be made with (MS) aggregation. If the number of estimators to be aggregated is large, $M > \sqrt{n}$, the remainder term in the (L) bound becomes too large, but, in a strict sense, there is no winner among the three aggregation techniques. 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) for (MS) aggregation in Gaussian regression; Nemirovski (2000), Juditsky and Nemirovski (2000), Tsybakov (2003) for (C) aggregation with $M > \sqrt{n}$; and Tsybakov (2003) for (L) aggregation with known marginal measure μ and for (C) aggregation with $M \le \sqrt{n}$. For

more general situations there exist less precise results of the type

$$\mathbb{E}_f\|\widetilde{f} - f\|^2 \le C_0 \inf_{\lambda\in H^M} \mathbb{E}_f\|\mathsf{f}_\lambda - f\|^2 + \Delta_{n,M}, \tag{1.4}$$

where $C_0 > 1$ is a constant independent of f and n, and $\Delta_{n,M}$ is a remainder term, not necessarily having the same behavior in n and M as the optimal one $\psi_{n,M}$. A disadvantage of (1.4) over (1.2) is that, when the oracle risk $R^* = \inf_{\lambda\in H^M} \mathbb{E}_f\|\mathsf{f}_\lambda - f\|^2$ is large, the additional term $(C_0 - 1)R^*$ on the right-hand side of (1.4) may be much larger than the remainder term $\Delta_{n,M}$, thus substantially spoiling the convergence properties. This effect is less pronounced if $C_0 = 1 + \varepsilon$ for some arbitrarily small $\varepsilon > 0$ or for $\varepsilon = \varepsilon_n \to 0$ as $n \to \infty$.

Bounds of the type (1.4) in regression problems have been obtained by many authors mainly for the model selection case (when H^M is the set of vertices of the simplex Λ^M), see, for example, Kneip (1994), Barron *et al.* (1999), Lugosi and Nobel (1999), Catoni (2004), Györfi *et al.* (2002), Baraud (2000, 2002), Bartlett *et al.* (2002), Wegkamp (2003), Birgé (2003), Bunea (2004), Bunea and Wegkamp (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 Catoni (2004), Yang (2000, 2001, 2002), Györfi *et al.* (2002), Wegkamp (2003), Birgé (2003), and Tsybakov (2003). Here we make the standard assumption that $\widehat{f}_1, \dots, \widehat{f}_M$ are uniformly bounded, but otherwise they can be arbitrary.

Various convex aggregation procedures for nonparametric regression have emerged in the last decade. They include bootstrap based methods, as suggested by LeBlanc and Tibshirani (1996) and cross-validation based stacking, as in Wolpert (1992) or Breiman (1996). 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. Juditsky and Nemirovski (2000), Nemirovski (2000) propose a stochastic approximation algorithm that achieves the bound (1.2) for (C) aggregation with optimal rate $\psi_{n,M}$ in the case $M > n/\log n$. They also show that the bound is achieved by usual (non-penalized) least squares convex aggregation. Yang (2000, 2001, 2004) suggest several methods of convex aggregation, in particular ARM (adaptive regression by mixing). He proves bounds of the form (1.4) with constants C_0 that are typically much larger than 1 and with rates $\Delta_{n,M}$ that can be equal or approximately equal to the optimal rates $\psi_{n,M}$ when M is a power of n. Audibert (2004) establishes (1.2) for a PAC-Bayesian method of convex aggregation with almost optimal rates, up to a logarithmic factor.

Birgé (2003) suggests a convex aggregation method satisfying (1.4) with a constant C_0 that can be much greater than 1 and with a rate that is optimal for $M > \sqrt{n}$ and suboptimal for $M \le \sqrt{n}$. On the other hand, Koltchinskii (2004, Section 8) proves (1.2) for a convex aggregate $\widetilde{f}$ with optimal rate for $M \le \sqrt{n}$ and with almost optimal rate for $M > \sqrt{n}$.

Linear aggregation procedures have received substantially less attention. For regression models with random design, a procedure achieving the bound (1.2) with optimal rate $\psi_{n,M}$ of (L) aggregation can be found in Tsybakov (2003). For Gaussian white noise models, linear aggregation has been discussed earlier by Nemirovski (2000).

Aggregation procedures are typically based on sample splitting. The initial sample $\mathcal{D}_n$ is divided into two independent subsamples $\mathcal{D}_m^1$ and $\mathcal{D}_\ell^2$ of sizes m and ℓ, respectively, where $m \gg \ell$ and $m + \ell = n$. The first subsample $\mathcal{D}_m^1$ (called training sample) is used to construct estimators $\widehat{f}_1, \dots, \widehat{f}_M$ and the second subsample $\mathcal{D}_\ell^2$ (called learning sample) is used to aggregate them (*i.e.*, to construct $\widetilde{f}$). In this paper we do not consider sample splitting schemes but rather deal with an idealized scheme. Following Nemirovski (2000), the first subsample is fixed 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 learning. Our aim is to find estimators based on the sample $\mathcal{D}_n$ that would mimic simultaneously the linear, convex and model selection oracles with the fastest possible rates (or, equivalently, with the smallest possible remainder terms $\Delta_{n,M}$). A passage to the initial model is straightforward: it is enough to condition on the first subsample, to use the learning bounds of the type (1.2), (1.4) obtained for the idealized scheme, and then to take expectations of both sides of the inequalities over the distribution of the whole sample $\mathcal{D}_n$.

Another interpretation of aggregation of fixed functions $f_1, \dots, f_M$ is related to parametric regression for linear models of dimension M, where M can be very large or increasing with n. In fact, assume that both X_i and $\widehat{f}_j = f_j$ are fixed (non-random), and consider the linear regression model with design matrix $(f_j(X_i))_{1 \le i \le n,\ 1 \le j \le M}$ and the empirical counterpart of the norm $\|\cdot\|$ defined by

$$\|f\|_n = \left(\frac{1}{n} \sum_{i=1}^{n} f^2(X_i) \right)^{1/2}.$$

Then, for $H^M = \Lambda^M$ or $H^M = \mathbb{R}^M$, the value $\inf_{\lambda \in H^M} \|\mathsf{f}_\lambda - f\|_n^2$ represents the best least squares approximation of an unknown function f at points X_i by the convex or linear span, respectively, of the columns of the design matrix. Consequently, estimators $\widetilde{f}$ satisfying oracle

inequalities of the form

$$\mathbb{E}_f\|\widetilde{f} - f\|_n^2 \le C_0 \inf_{\lambda\in H^M} \|\mathsf{f}_\lambda - f\|_n^2 + \Delta_{n,M} \tag{1.5}$$

mimic the best linear/convex least-squares approximation of f in a parametric regression framework, provided $C_0 \ge 1$ is close to 1. In (1.5), $\Delta_{n,M}$ can be interpreted as the price to pay for the dimension M of the regression model, and we will show that (for an appropriate choice of the aggregate $\widetilde{f}$) $\Delta_{n,M} = \psi_{n,M}$, where $\psi_{n,M}$ is the optimal rate of aggregation as defined in (1.3). For the case of linear aggregation, this can be viewed in the spirit of earlier work on linear models with growing dimension M [Yohai and Maronna (1979), Portnoy (1984)], but here we obtain non-asymptotic results and our risk is defined in terms of the regression functions and not in terms of their parameters.

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. Consequently, the smallest of the three will be achieved. Our answer will thus meet the desiderata of both model 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 soft thresholding (L_1 or Lasso type penalty) and with hard thresholding, respectively.

In Section 3.1 we show that a hard threshold aggregate satisfies inequalities of the type (1.5), with C_0 arbitrarily close to 1, and 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 (MS), (C) and (L) bounds when the empirical norm $\|\cdot\|_n$ is used to define the risk.

In Section 3.2 we study the performance of a slightly different hard threshold aggregate under the $L_2(\mathbb{R}^d, \mu)$ norm. We show that this aggregate satisfies simultaneously the oracle inequalities of the type (1.4) corresponding to the (MS) and (C) bounds, with a remainder term $\Delta_{n,M}$ that possibly differs from the optimal $\psi_{n,M}$ in a logarithmic factor, and with C_0 arbitrarily close to 1.

In Section 4 we study aggregation with the L_1 penalty and we obtain (1.5) simultaneously for the (MS), (C) and (L) cases, with C_0 arbitrarily close to 1 and with a remainder term $\Delta_{n,M}$ that differs from the optimal $\psi_{n,M}$ only in a logarithmic factor.

Finally, we study lower bounds for (MS) and (L) aggregation in the fixed design case in Section 5, complementing the results obtained for the random design case by Tsybakov (2003).

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 $f_j : \mathcal{X} \to \mathbb{R}$, $j = 1, \dots, M$, with $M \geq 2$, belong to the class $\mathcal{F}_0$ of uniformly bounded functions defined by*

$$\mathcal{F}_0 \stackrel{\text{def}}{=} \Big\{ g : \mathcal{X} \to \mathbb{R} \Big| \sup_{x \in \mathcal{X}} |g(x)| \leq L \Big\}$$

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

The normality assumption (A1) on the distribution of errors is convenient since we need certain exponential tail bounds in the proofs (see Lemma 3.10 below). For example, bounded regression can be easily incorporated in this framework using maximal inequalities due to Talagrand (1994a, b) and Panchenko (2003). More generally, subgaussian errors are allowed at the cost of increasing technicalities, see Van de Geer (2000). In order to retain a transparent presentation of both the results and proofs, we confine ourselves to the Gaussian regression framework.

For any $\lambda = (\lambda_1, \dots, \lambda_M) \in \mathbb{R}^M$, 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 (see the Introduction). Here we consider the ideal situation in which they are fixed, *i.e.*, we concentrate on learning only. The learning method that we propose is based on aggregating the f_j's via penalized least squares.

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

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

where $I\{\cdot\}$ denotes the indicator function, and $J(\lambda) = \{j \in \{1, \dots, M\} : \ \lambda_j \neq 0\}$. Introduce the residual sum of squares

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

Given a penalty term $\text{pen}(\lambda)$, the penalized least squares estimator $\widehat{\lambda} = (\widehat{\lambda}_1, \dots, \widehat{\lambda}_M)$ 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

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

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 (MS), (C) and (L) oracles when one of the following two penalties is used:

$$\text{pen}(\lambda) = K_1 \frac{M(\lambda)}{n} \log \left(1 + \frac{M}{M(\lambda) \vee 1} \right) \tag{2.2}$$

or

$$\text{pen}(\lambda) = \sum_{j=1}^{M} r_{n,j} |\lambda_j|, \tag{2.3}$$

where $K_1 > 0$ is a constant independent of M, n, and $r_{n,j}$'s are the data-dependent weights defined in (4.3).

We refer to the penalty in (2.2) as *hard threshold penalty*. This is motivated by the well known fact that, in the sequence space model (*i.e.*, when 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 thresholded values of the Y_j's (see, for instance, Härdle *et al.* (1998), page 138). Our penalty (2.2) is not exactly of that form, but it differs from it only in a logarithmic factor.

The penalty (2.3), again in the sequence space model, leads to $\widehat{\lambda}_j$'s that are soft thresholded 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 [Tibshirani (1996), Efron *et al.* (2004)]. Our results show that, with $r_{n,j}$'s defined by (4.3), the soft threshold penalty allows near optimal aggregation. The same is true for the hard threshold penalty (2.2) under somewhat different conditions.

In what follows, we denote by $C, C_1, C_2, \ldots$ finite positive constants, possibly different on different occasions.

3. Near optimal aggregation with the hard threshold penalty

3.1. The fixed design case. In this section we show that the penalized least squares estimator using a penalty of the form (2.2) achieves simultaneously the (MS), (L), and (C) bounds of the form (1.5) with the correct rates $\Delta_{n,M} = \psi_{n,M}$. Consequently, the smallest bound is achieved by our aggregate. The results of this section are established for the empirical loss $\|\widetilde{f} - f\|_n^2$. The next theorem presents an oracle inequality which implies all the three bounds.

Theorem 3.1. *Let $X_i \in \mathcal{X}$, $i = 1, \ldots, n$, be fixed. Let $\widetilde{f}$ be the penalized least squares estimate defined in (2.1) with penalty (2.2). There exist constants $C_1, C_2 > 0$ such that for all $a > 1$, for $K_1 = K_0 a\sigma^2$, with $K_0 > 0$ large enough, and for all integers $n \geq 1$ and $M \geq 2$,*

$$\begin{aligned} &\mathbb{E}_f\|\widetilde{f} - f\|_n^2 \\ &\leq \inf_{\lambda \in \mathbb{R}^M} \left\{ \frac{a+1}{a-1}\|\mathsf{f}_\lambda - f\|_n^2 + \ C_1 a\sigma^2 \frac{M(\lambda)}{n} \log\left(1 + \frac{M}{M(\lambda) \vee 1}\right)\right\} + C_2 \frac{a\sigma^2}{n}. \end{aligned} \tag{3.1}$$

This theorem is proved in Section 3.3. The following three corollaries present bounds of the form (1.5) for (MS), (L), and (C) aggregation, respectively.

Corollary 3.2 (MS). *Let the assumptions of Theorem 3.1 be satisfied. Then there exists a constant $C_3 > 0$ such that for all $\varepsilon > 0$, for $K_1 = K_1(\varepsilon, \sigma^2)$ large enough and for all integers $n \geq 1$ and $M \geq 2$,*

$$\mathbb{E}_f\|\widetilde{f} - f\|_n^2 \leq (1+\varepsilon) \inf_{1 \leq j \leq M} \|f_j - f\|_n^2 + C_3\sigma^2 \left(1 + \varepsilon^{-1}\right) \frac{\log M}{n}.$$

Proof. Since the infimum on the right of (3.1) is taken over all $\lambda \in \mathbb{R}^M$, the bound easily follows by considering only the subset consisting of the M vertices $(\lambda_1, \ldots, \lambda_M) = (1, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$ in Λ^M, and by putting $a = 1 + 2/\varepsilon$. □

COROLLARY 3.3 (L). *Let the assumptions of Theorem 3.1 be satisfied. Then there exists a constant $C_3 > 0$ such that for all $\varepsilon > 0$, for $K_1 = K_1(\varepsilon, \sigma^2)$ large enough and for all integers $n \geq 1$ and $M \geq 2$,*

$$\mathbb{E}_f \|\widetilde{f} - f\|_n^2 \leq (1+\varepsilon) \inf_{\lambda \in \mathbb{R}^M} \|\mathsf{f}_\lambda - f\|_n^2 + C_3 \sigma^2 \left(1 + \varepsilon^{-1}\right) \frac{M}{n}.$$

Proof. Since $x \mapsto x \log(1 + M/x)$ is increasing for $1 \leq x \leq M$,

$$\sup_{\lambda \in \mathbb{R}^M} \frac{M(\lambda)}{n} \log\left(1 + \frac{M}{M(\lambda) \vee 1}\right) = \frac{M}{n} \log 2.$$

The result then follows from (3.1) with $a = 1 + 2/\varepsilon$. □

COROLLARY 3.4 (C). *Let the assumptions of Theorem 3.1 be satisfied. Then there exists a constant $C_3' > 0$ depending on L and σ^2 such that for all $\varepsilon > 0$, for $K_1 = K_1(\varepsilon, \sigma^2)$ large enough and for all integers $n \geq 1$ and $M \geq 2$,*

$$\mathbb{E}_f \|\widetilde{f} - f\|_n^2 \leq (1+\varepsilon) \inf_{\lambda \in \Lambda^M} \|\mathsf{f}_\lambda - f\|_n^2 + C_3' \left(1 + \varepsilon + \varepsilon^{-1}\right) \psi_n^C(M),$$

where

$$\psi_n^C(M) = \begin{cases} M/n & \text{if } M \leq \sqrt{n}, \\ \sqrt{\{\log(1 + M/\sqrt{n})\}/n} & \text{if } M > \sqrt{n}. \end{cases}$$

Proof. For $M \leq \sqrt{n}$ the result follows from Corollary 3.3. Assume now that $M > \sqrt{n}$ and let m be the integer part of

$$x_{n,M} = \sqrt{\frac{n \log 2}{\log(1 + M/\sqrt{n})}}.$$

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

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

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

$$\min_{g \in \mathcal{C}} \|g - f\|_n^2 \leq \min_{\lambda \in \Lambda^M} \|\mathsf{f}_\lambda - f\|_n^2 + \frac{L^2}{m}. \tag{3.2}$$

For completeness, we give the proof of (3.2) in the Appendix. Since $M(\lambda) \leq m \leq x_{n,M}$ for the vectors λ corresponding to $g \in \mathcal{C}$, and since $x \mapsto x \log\left(1 + \frac{M}{x}\right)$ is increasing for $1 \leq x \leq M$, we get from (3.1):

$$\mathbb{E}_f \|\widetilde{f} - f\|_n^2 \leq \inf_{g \in \mathcal{C}} \left\{ \frac{a+1}{a-1} \|g - f\|_n^2 + C_1 a \sigma^2 \, \frac{x_{n,M}}{n} \log\left(1 + \frac{M}{x_{n,M}}\right) \right\} + \frac{C_2 a \sigma^2}{n}.$$

Using this inequality, (3.2) and the fact that $m = \lfloor x_{n,M} \rfloor \geq x_{n,M}/2$ for $x_{n,M} \geq 1$, we obtain

$$\begin{aligned} \mathbb{E}_f \|\widetilde{f} - f\|_n^2 \;\leq\; & \frac{a+1}{a-1} \inf_{\lambda \in \Lambda^M} \|\mathsf{f}_\lambda - f\|_n^2 + \left(\frac{a+1}{a-1}\right) \frac{2L^2}{x_{n,M}} \\ & + C_1 a\sigma^2 \, \frac{x_{n,M}}{n} \log\left(1 + \frac{M}{x_{n,M}}\right) + \frac{C_2 a\sigma^2}{n}. \end{aligned} \tag{3.3}$$

We use this bound for all choices of $\lambda \in \Lambda^M$ with $m \geq M(\lambda) \neq 0$. 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 $1/x_{n,M} \geq 1$, and we use the trivial upper bound

$$\frac{a+1}{a-1}\|f\|_n^2 + \frac{C_2 a\sigma^2}{n} \leq \left(\frac{a+1}{a-1}L^2 + C_2 a\sigma^2\right) \left(\frac{\log(1 + M/\sqrt{n})}{n \log 2}\right)^{1/2}$$

for the right-hand side of (3.1).

To complete the proof of the Corollary, it remains to put $a = 1 + 2/\varepsilon$ and to note that

$$\log\left(1 + \frac{M}{x_{n,M}}\right) \leq 2 \log\left(1 + \frac{M}{\sqrt{n}}\right),$$

in view of the elementary inequality $\log\left(1 + (\log 2)^{-1/2} y \sqrt{\log(1+y)}\right) \leq 2\log(1+y)$, for all $y \geq 1$. □

We remark now that the aggregate considered in Theorem 3.1 satisfies also the bounds "in probability" that are similar in spirit to (3.1) and its corollaries.

THEOREM 3.5. *Let $X_i \in \mathcal{X}$, $i = 1, \ldots, n$, be fixed. Let $\widetilde{f}$ be the penalized least squares estimate defined in (2.1) with penalty (2.2). There exist constants $C_1, L_1, L_2 > 0$ such that for all $a > 1$, for $K_1 = K_0 a\sigma^2$, with $K_0 > 0$ large enough, and for all integers $n \geq 1$, $M \geq 2$ and any $\delta > 0$,*

$$\begin{aligned} \mathbb{P}\left(\|\widetilde{f} - f\|_n^2 \geq \inf_{\lambda \in \mathbb{R}^M} \left\{ \frac{a+1}{a-1}\|\mathsf{f}_\lambda - f\|_n^2 + \; C_1 a\sigma^2 \frac{M(\lambda)}{n} \log\left(1 + \frac{M}{M(\lambda) \vee 1}\right)\right\} + \delta\right) \\ \leq L_1 \exp\left(-L_2 \frac{n\delta}{a\sigma^2}\right). \end{aligned} \tag{3.4}$$

As in the case of Theorem 3.1, we can consequently obtain the analogues of Corollaries 3.2 - 3.4, by replacing the infimum in (3.4) by its particular form for the cases (MS), (L) and (C), respectively. We do not include each case, for brevity.

3.2. The random design case. In this subsection we show that an oracle inequality similar to (3.1) continues to hold if the empirical norm $\|\cdot\|_n$ is replaced by the $L_2(\mathbb{R}^d,\mu)$ norm $\|\cdot\|$. This result is more difficult to obtain and we do not achieve exactly the same bounds.

We need to restrict minimization of the penalized sum of squares to a bounded set in $\mathbb{R}^M$. Define, for any $T>0$,

$$\Lambda_{M,T}=\left\{\lambda\in\mathbb{R}^M:\ \sum_{j=1}^M|\lambda_j|\le T\right\}.$$

The penalty term needs to be chosen slightly larger than before:

$$\mathrm{pen}(\lambda)=K_1\frac{M(\lambda)}{n}\log\left(1+\frac{M\vee n}{M(\lambda)\vee 1}\right) \tag{3.5}$$

for some large $K_1>0$. We note that here K_1 is not necessarily the same as in (2.2), we just use the same notation for factors in the penalty term.

Theorem 3.6. *Assume that $X_1,\dots,X_n$ are independent random variables with common probability measure μ. Let $T<\infty$ be fixed, and set*

$$B=L^2(T+1)^2.$$

Let $\widetilde{f}=f_{\widehat{\lambda}}$ where

$$\widehat{\lambda}=\arg\min_{\lambda\in\Lambda_{M,T}}\{\widehat{S}(\lambda)+\mathrm{pen}(\lambda)\}$$

with the penalty given in (3.5). Then there exist constants $C_1,C_2>0$ such that for all $a>1$, for $K_1=K_1(a,B,\sigma^2)$ large enough, and for all integers $n\ge 1$ and $M\ge 2$,

$$\begin{aligned}&\mathbb{E}_f\|\widetilde{f}-f\|^2\\ &\le\inf_{\lambda\in\Lambda_{M,T}}\left\{\frac{a+1}{a-1}\|\mathsf{f}_\lambda-f\|^2+\ C_1a\sigma^2\frac{M(\lambda)}{n}\log\left(1+\frac{M\vee n}{M(\lambda)\vee 1}\right)\right\}+C_2\frac{a(\sigma^2+B)}{n}.\end{aligned} \tag{3.6}$$

Because of the slight increase in the penalty, the remainder term in (3.6) is somewhat larger than the one given in (3.1): we now have $M\vee n$ in place of M under the logarithm.

As corollaries, one obtains the following (MS) and (C) bounds for the estimator $\widetilde{f}$ defined in Theorem 3.6.

Corollary 3.7 (MS). *Let the assumptions of Theorem 3.6 be satisfied and $T\ge 1$. Then there exists a constant $C>0$ such that for all $\varepsilon>0$, for $K_1=K_1(\varepsilon,\sigma^2)$ large enough and for all integers $n\ge 1$ and $M\ge 2$,*

$$\mathbb{E}_f\|\widetilde{f}-f\|^2\le(1+\varepsilon)\inf_{1\le j\le M}\|f_j-f\|^2+C\sigma^2\left(1+\varepsilon^{-1}\right)\frac{\log(M\vee n)}{n}.$$

COROLLARY 3.8 (C). *Let the assumptions of Theorem 3.6 be satisfied and $T \geq 1$. Then there exists a constant $C' > 0$ depending on L and σ^2 such that for all $\varepsilon > 0$, for $K_1 = K_1(\varepsilon, \sigma^2)$ large enough and for all integers $n \geq 1$ and $M \geq 2$,*

$$\mathbb{E}_f \|\widetilde{f} - f\|^2 \leq (1+\varepsilon) \inf_{\lambda \in \Lambda^M} \|\mathsf{f}_\lambda - f\|^2 + C' \left(1 + \varepsilon + \varepsilon^{-1}\right) \widetilde{\psi}_n^C(M),$$

where

$$\widetilde{\psi}_n^C(M) = \begin{cases} (M \log n)/n & \text{if } M \leq \sqrt{n}, \\ \sqrt{\{\log(1 + (M \vee n)/\sqrt{n})\}/n} & \text{if } M > \sqrt{n}. \end{cases}$$

As compared to Corollaries 3.2 and 3.4, these results present slightly different rates of convergence: here the factor $\log M$ is replaced by $\log n$ for values $M < n$. The proofs are omitted since Corollaries 3.7 and 3.8 readily follow from the oracle inequality (3.6) and the fact that $\Lambda^M \subset \Lambda_{M,T}$ for $T \geq 1$ via an argument similar to the proofs of Corollaries 3.2 and 3.4.

3.3. PROOF OF THEOREM 3.1. Let λ be a fixed, but arbitrary point in $\mathbb{R}^M$. Define for all $1 \leq m \leq M$,

$$A_m(\lambda) = \{\bar{\lambda} = \lambda' - \lambda \in \mathbb{R}^M : \ M(\lambda') = m\}.$$

Let $J_k, \ k = 1, \ldots, \binom{M}{m}$, be all the subsets of $\{1, \ldots, M\}$ of cardinality m. Define

$$A_{m,k}(\lambda) = \left\{\bar{\lambda} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_M) \in A_m(\lambda) : \ \lambda'_j \neq 0 \Leftrightarrow j \in J_k\right\}$$

where $\lambda'_j = \bar{\lambda}_j + \lambda_j$. The collection $\left\{A_{m,k}(\lambda) : \ 1 \leq k \leq \binom{M}{m}\right\}$ forms a partition of the set $A_m(\lambda)$. Furthermore, define affine subspaces of $\mathbb{R}^n$ of the form

$$B_{m,k}(\lambda) = \left\{h = (\mathsf{f}_{\bar{\lambda}}(X_1), \ldots, \mathsf{f}_{\bar{\lambda}}(X_n)) \in \mathbb{R}^n : \ \bar{\lambda} \in A_{m,k}(\lambda)\right\}$$

and let $\Pi^\lambda_{m,k} W$ denote the projection of the vector $W = (W_1, \ldots, W_n)$ onto $B_{m,k}(\lambda)$. Clearly, $\dim(B_{m,k}(\lambda)) \leq m$. Finally, we define for each $\gamma \in \mathbb{R}^M$,

$$V_n(\gamma) = \frac{1}{n} \sum_{i=1}^n W_i \frac{\mathsf{f}_\gamma(X_i)}{\|\mathsf{f}_\gamma\|_n} \ \text{ if } \ \|\mathsf{f}_\gamma\|_n \neq 0,$$

and $V_n(\gamma) \stackrel{\text{def}}{=} 0$, otherwise.

LEMMA 3.9. *For all $a > 1, b > 0$ and $\lambda \in \mathbb{R}^M$, we have*

$$\begin{aligned}\|\widetilde{f} - f\|_n^2 \quad \leq \quad & \frac{1+b}{b}\frac{a}{a-1}\|\mathsf{f}_\lambda - f\|_n^2 + \frac{a}{a-1}K_1\frac{M(\lambda)}{n}\log\left(1 + \frac{M}{M(\lambda)\vee 1}\right) \\ & + \frac{a}{a-1}\max_{1\leq m\leq M}\max_{1\leq k\leq \binom{M}{m}}\left\{(a+b)\|\Pi^\lambda_{m,k}W\|_n^2 - \frac{K_1 m}{n}\log\left(1+\frac{M}{m\vee 1}\right)\right\} \\ & + \frac{a(a+b)}{a-1}V_n^2(\lambda).\end{aligned}$$

Proof. By the definition of $\widehat{\lambda}$, for any $\lambda \in \mathbb{R}^M$,

$$\widehat{S}(\widehat{\lambda}) + \mathrm{pen}(\widehat{\lambda}) \leq \widehat{S}(\lambda) + \mathrm{pen}(\lambda).$$

Rewriting this inequality yields

$$\|\widetilde{f} - f\|_n^2 \quad \leq \quad \|\mathsf{f}_\lambda - f\|_n^2 + 2\left\langle W, \widetilde{f} - \mathsf{f}_\lambda\right\rangle_n + \mathrm{pen}(\lambda) - \mathrm{pen}(\widehat{\lambda}),$$

where $< \cdot, \cdot >_n$ denotes the scalar product associated with the norm $\|\cdot\|_n$. Since $\|\widetilde{f} - \mathsf{f}_\lambda\|_n = 0$ implies that $\left\langle W, \widetilde{f} - \mathsf{f}_\lambda\right\rangle_n = 0$, we find

$$\begin{aligned}\|\widetilde{f} - f\|_n^2 \quad & \leq \quad \|\mathsf{f}_\lambda - f\|_n^2 + 2V_n(\widehat{\lambda} - \lambda)\|\widetilde{f} - \mathsf{f}_\lambda\|_n + \mathrm{pen}(\lambda) - \mathrm{pen}(\widehat{\lambda}) \\ & \leq \quad \|\mathsf{f}_\lambda - f\|_n^2 + 2V_n(\widehat{\lambda} - \lambda)\|\widetilde{f} - f\|_n + 2V_n(\widehat{\lambda} - \lambda)\|\mathsf{f}_\lambda - f\|_n + \mathrm{pen}(\lambda) - \mathrm{pen}(\widehat{\lambda}) \\ & \leq \quad (1 + \frac{1}{b})\|\mathsf{f}_\lambda - f\|_n^2 + aV_n^2(\widehat{\lambda} - \lambda) + \frac{1}{a}\|\widetilde{f} - f\|_n^2 + bV_n^2(\widehat{\lambda} - \lambda) + \mathrm{pen}(\lambda) - \mathrm{pen}(\widehat{\lambda}),\end{aligned}$$

where $a, b > 0$ are arbitrary, and we used the inequality $2xy \leq cx^2 + y^2/c$ valid for all $x, y \in \mathbb{R}$ and $c > 0$. Consequently, for any $a > 1, b > 0$, we find

$$\begin{aligned}\|\widetilde{f} - f\|_n^2 \quad \leq \quad & \frac{1+b}{b}\frac{a}{a-1}\|\mathsf{f}_\lambda - f\|_n^2 + \frac{a}{a-1}\mathrm{pen}(\lambda) \\ & + \frac{a}{a-1}(a+b)V_n^2(\widehat{\lambda} - \lambda) - \frac{a}{a-1}\mathrm{pen}(\widehat{\lambda}).\end{aligned}$$

Next, since $\mathbb{R}^M = \bigcup_{m=0}^M \bigcup_{k=1}^{\binom{M}{m}} A_{m,k}(\lambda)$, we find that

$$\begin{aligned}& (a+b)V_n^2(\widehat{\lambda} - \lambda) - \mathrm{pen}(\widehat{\lambda}) \\ & = (a+b)V_n^2(\widehat{\lambda} - \lambda) - \mathrm{pen}(\widehat{\lambda} - \lambda + \lambda) \\ & \leq \max_{0\leq m\leq M}\max_{1\leq k\leq \binom{M}{m}}\max_{\bar{\lambda}\in A_{m,k}(\lambda)}\left\{(a+b)V_n^2(\bar{\lambda}) - \mathrm{pen}(\bar{\lambda} + \lambda)\right\}.\end{aligned}$$

It remains to bound the term on the right in view of the last two displays. The case $m = 0$ is degenerate as $A_0(\lambda) = A_{0,1}(\lambda) = \{-\lambda\}$. Note that for $\bar{\lambda} = -\lambda$,

$$(a+b)V_n^2(\bar{\lambda}) - \mathrm{pen}(\bar{\lambda} + \lambda) = (a+b)V_n^2(\lambda),$$

since $\mathrm{pen}(0) = 0$ and $\mathsf{f}_{-\lambda} = -\mathsf{f}_\lambda$. For each $m \geq 1$, we have

$$\begin{aligned}
&\max_{1\leq k\leq \binom{M}{m}} \max_{\bar\lambda\in A_{m,k}(\lambda)} \left\{(a+b)V_n^2(\bar\lambda) - \mathrm{pen}(\bar\lambda+\lambda)\right\} \\
&\leq \max_{1\leq k\leq \binom{M}{m}} \max_{\bar\lambda\in A_{m,k}(\lambda)} \left\{(a+b)\|\Pi^\lambda_{m,k}W\|_n^2 - \mathrm{pen}(\bar\lambda+\lambda)\right\} \\
&\quad \text{by the orthogonality of } W-\Pi^\lambda_{m,k}W \text{ and } (\mathsf{f}_{\bar\lambda}(X_1),\dots,\mathsf{f}_{\bar\lambda}(X_n)) \text{ for all } \bar\lambda\in A_{m,k}(\lambda) \\
&= \max_{1\leq k\leq \binom{M}{m}} \left\{(a+b)\|\Pi^\lambda_{m,k}W\|_n^2 - \frac{K_1}{n} m\log\left(1+\frac{M}{m\vee 1}\right)\right\} \\
&\quad \text{in view of (3.5) and since } M(\bar\lambda+\lambda)=m \text{ for all } \bar\lambda\in A_{m,k}(\lambda).
\end{aligned}$$

This concludes the proof of the lemma. □

From now on, we take $a = b > 1$. Since, by Assumption (A1), the errors W_i are normal $N(0,\sigma^2)$, the standardized statistic $n\sigma^{-2}\|\Pi^\lambda_{m,k}W\|_n^2$ has a χ^2 distribution with m degrees of freedom for all $1 \leq k \leq \binom{M}{m}$. The following tail bound for such a statistic will be useful.

LEMMA 3.10. *Let Z_d denote a random variable having the χ^2 distribution with d degrees of freedom. Then for all $x > 0$,*

$$\mathbb{P}\{Z_d - d \geq x\sqrt{2d}\} \leq \exp\left(-\frac{x^2}{2(1+x\sqrt{2/d})}\right). \tag{3.7}$$

Proof. See Cavalier *et al.* (2002), equation (27) at page 857. □

LEMMA 3.11. *There exists $C > 0$ such that, for any integer $n \geq 1$ and any $a > 1$, $K_1 = K_0 a\sigma^2$ with $K_0 > 0$ large enough,*

$$\mathbb{E}_f \max_{1\leq m\leq M} \max_{1\leq k\leq \binom{M}{m}} \left\{2a\|\Pi^\lambda_{m,k}W\|_n^2 - \frac{K_1}{n} m\log\left(1+\frac{M}{m\vee 1}\right)\right\} \leq C\frac{a\sigma^2}{n}, \tag{3.8}$$

$$\mathbb{E}_f V_n^2(\lambda) \leq \frac{\sigma^2}{n}. \tag{3.9}$$

Proof. Inequality (3.9) is trivial and we will prove only (3.8). For any $\delta > 0$ we have

$$
\begin{aligned}
p_\delta &\stackrel{\mathrm{def}}{=} \mathbb{P}\left[\max_{1\le m\le M}\max_{1\le k\le \binom{M}{m}}\left\{2a\|\Pi^\lambda_{m,k}W\|_n^2 - \frac{K_1}{n}m\log\left(1+\frac{M}{m\vee 1}\right)\right\}\ge\delta\right]\\
&\le \sum_{m=1}^{M}\sum_{k=1}^{\binom{M}{m}}\mathbb{P}\left[2a\|\Pi^\lambda_{m,k}W\|_n^2 - \frac{K_1}{n}m\log\left(1+\frac{M}{m\vee 1}\right)\ge\delta\right]\\
&= \sum_{m=1}^{M}\sum_{k=1}^{\binom{M}{m}}\mathbb{P}\left[Z_m \ge \frac{K_1}{2a\sigma^2}m\log\left(1+\frac{M}{m}\right)+\frac{n\delta}{2a\sigma^2}\right]\\
&= \sum_{m=1}^{M}\binom{M}{m}\mathbb{P}\left[\frac{Z_m-m}{\sqrt{2m}} \ge \frac{K_1}{2a\sigma^2}\frac{\sqrt{m}}{\sqrt{2}}\log\left(1+\frac{M}{m}\right)-\frac{\sqrt{m}}{\sqrt{2}}+\frac{n\delta}{2a\sigma^2\sqrt{2m}}\right]\\
&\le \sum_{m=1}^{M}\binom{M}{m}\exp\left(-C_0\left\{\frac{mK_1}{a\sigma^2}\log\left(1+\frac{M}{m}\right)+\frac{n\delta}{a\sigma^2}\right\}\right)
\end{aligned}
$$

by Lemma 3.10 for $K_1 = K_0 a\sigma^2$ with $K_0 > 0$ large enough and some universal constant $C_0 > 0$. Using the crude bound $\binom{M}{m} \le (eM/m)^m$ [see, for example, Devroye *et al.* (1996), page 218], the inequality $1+\log x \le 2\log(1+x)$, $\forall\, x \ge 1$, and taking K_0 such that $C_0K_0 > 4$ we get

$$
\begin{aligned}
\sum_{m=1}^{M}\binom{M}{m}\exp\left(-C_0\frac{mK_1}{a\sigma^2}\log\left(1+\frac{M}{m}\right)\right) &\le \sum_{m=1}^{M}\exp\left(-m\log\left(1+\frac{M}{m}\right)\right)\\
&\le \sum_{m=1}^{\infty}\exp(-m\log 2) < \infty.
\end{aligned}
$$

These inequalities finally yield the bound on the tail probabilities

$$
p_\delta \le C_3\exp\left(-C_4\frac{n\delta}{a\sigma^2}\right) \tag{3.10}
$$

for some constants $C_3, C_4 > 0$, which easily implies the bound (3.8) on the expected value. □

Proof of Theorem 3.1. Theorem 3.1 follows directly from Lemmas 3.9 and 3.11. □

Proof of Theorem 3.5. First notice that, by Lemma 3.9, for $a = b > 1$ there exists $C_1 > 0$ such that

$$
\begin{aligned}
&\mathbb{P}\left(\|\widetilde{f}-f\|_n^2 \ge \inf_{\lambda\in\mathbb{R}^M}\left\{\frac{a+1}{a-1}\|\mathsf{f}_\lambda - f\|_n^2 + C_1a\sigma^2\frac{M(\lambda)}{n}\log\left(1+\frac{M}{M(\lambda)\vee 1}\right)\right\}+\delta\right)\\
&\le \mathbb{P}\left(\frac{a}{a-1}\max_{1\le m\le M}\max_{1\le k\le\binom{M}{m}}\left\{2a\|\Pi^\lambda_{m,k}W\|_n^2 - \frac{K_1m}{n}\log\left(1+\frac{M}{m\vee 1}\right)\right\}\ge\delta/2\right)\\
&\qquad +\mathbb{P}\left(\frac{2a^2}{a-1}V_n^2(\lambda)\ge\delta/2\right).
\end{aligned}
$$

Next, the rescaled variable $n\sigma^{-2}V_n^2(\lambda)$ has a χ^2 distribution with 1 degree of freedom. Combining the exponential bound for tail probabilities of χ^2 random variables (Lemma 3.10) and the exponential bound (3.10) completes the proof. □

3.4. Proof of Theorem 3.6. By the same reasoning as in the proof of Theorem 3.1,

$$\begin{aligned}
\|\widetilde{f}-f\|^2 &= (1+a)\|\widetilde{f}-f\|_n^2 + \left\{\|\widetilde{f}-f\|^2-(1+a)\|\widetilde{f}-f\|_n^2\right\} \\
&\le (1+a)\left\{\|\mathsf{f}_\lambda-f\|_n^2 + 2\left\langle W,\widetilde{f}-\mathsf{f}_\lambda\right\rangle_n + \mathrm{pen}(\lambda)-\mathrm{pen}(\widehat{\lambda})\right\} \\
&\quad + \left\{\|\widetilde{f}-f\|^2-(1+a)\|\widetilde{f}-f\|_n^2\right\} \\
&= (1+a)\left\{\|\mathsf{f}_\lambda-f\|_n^2 + 2\left\langle W,\widetilde{f}-\mathsf{f}_\lambda\right\rangle_n + \mathrm{pen}(\lambda)-\frac{\mathrm{pen}(\widehat{\lambda})}{2}\right\} \\
&\quad + \left\{\|\widetilde{f}-f\|^2-(1+a)\|\widetilde{f}-f\|_n^2-\frac{1+a}{2}\mathrm{pen}(\widehat{\lambda})\right\}.
\end{aligned}$$

The first term on the right, provided $K_1>0$ is chosen large enough, can be handled in exactly the same way as in the proof of Theorem 3.1. It remains to study the second term on the right.

Considering separately the cases $M(\lambda)=0$ and $1\le M(\lambda)\le M$ we obtain

$$\begin{aligned}
&\|\widetilde{f}-f\|^2-(1+a)\|\widetilde{f}-f\|_n^2-\frac{1+a}{2}\mathrm{pen}(\widehat{\lambda}) \\
&\le \max\left\{U_0,\ \max_{1\le m\le M}\sup_{\lambda:M(\lambda)=m}\left[U_\lambda-\frac{1+a}{2}\mathrm{pen}(\lambda)\right]\right\}
\end{aligned}$$

where $U_\lambda=\|\mathsf{f}_\lambda-f\|^2-(1+a)\|\mathsf{f}_\lambda-f\|_n^2$. For each $1\le m\le M$, let the sets $A_{m,k}(0)$, $1\le k\le\binom{M}{m}$, form a partitioning of the set $A_m(0)=\{\lambda\in\mathbb{R}^M:\ M(\lambda)=m\}$. Deduce that, for any $\delta>0$,

$$\begin{aligned}
&\mathbb{P}\left\{\|\widetilde{f}-f\|^2-(1+a)\|\widetilde{f}-f\|_n^2-\frac{1+a}{2}\mathrm{pen}(\widehat{\lambda})\ge\delta\right\} \\
&\le \mathbb{P}\left\{U_0\ge\delta/2\right\}+\sum_{m=1}^{M}\mathbb{P}\left\{\sup_{\lambda:M(\lambda)=m}U_\lambda\ge D(\delta)\right\} \\
&\le \mathbb{P}\left\{U_0\ge\delta/2\right\}+\sum_{m=1}^{M}\sum_{k=1}^{\binom{M}{m}}\mathbb{P}\left\{\sup_{\lambda\in A_{m,k}(0)}U_\lambda\ge D(\delta)\right\}
\end{aligned}\tag{3.11}$$

where

$$D(\delta)=\frac{(1+a)K_1}{2n}m\log\left(1+\frac{n\vee M}{m\vee 1}\right)+\frac{\delta}{2}.$$

The following result establishes a bound on the shatter coefficient of the class of subgraphs of the functions $(\mathsf{f}_\lambda - f)^2$ that will be subsequently used to control the behavior of the empirical process on the right-hand side of (3.11).

LEMMA 3.12. *Let $\mathbb{S}(n,m,k)$ be the shatter coefficient of the collection of sets*

$$\left\{(x,\beta):\ (\mathsf{f}_\lambda - f)^2(x) \geq \beta,\ \beta \geq 0,\ x \in \mathcal{X}\right\}, \qquad \lambda \in A_{m,k}(0).$$

Then, for any $1 \leq m \leq M$, $1 \leq k \leq \binom{M}{m}$, we have

$$\log \mathbb{S}(2n,m,k) \leq Cm\left\{1 + \log\left(1 + \frac{n}{m}\right)\right\}$$

where $C > 0$ is an absolute constant.

Proof. Note that

$$\begin{aligned} &\left\{(x,\beta):\ (\mathsf{f}_\lambda - f)^2(x) \geq \beta,\ \beta \geq 0\right\} \\ = \ &\left\{(x,\beta):\ \mathsf{f}_\lambda(x) - f(x) \leq -\sqrt{\beta},\ \beta \geq 0\right\} \cup \left\{(x,\beta):\ \mathsf{f}_\lambda(x) - f(x) \geq \sqrt{\beta},\ \beta \geq 0\right\} \end{aligned}$$

and recall that the VC-dimension of the collection of sets $\left\{(x,\beta):\ \mathsf{f}_\lambda(x) - f(x) \geq \sqrt{\beta},\ \beta \geq 0\right\}$, $\lambda \in A_{m,k}(0)$, is less than $m+1$, cf. Theorem 13.9 of Devroye, Györfi and Lugosi (1996) or van de Geer (2000), page 40. Similarly, the VC-dimension of $\left\{(x,\beta):\ \mathsf{f}_\lambda(x) - f(x) \leq -\sqrt{\beta},\ \beta \geq 0\right\}$, $\lambda \in A_{m,k}(0)$, is less than $m+1$. Apply Lemma 15, page 18, in Pollard (1984) to deduce that the collection of sets $\left\{(x,\beta):\ (\mathsf{f}_\lambda - f)^2(x) \geq \beta,\ \beta \geq 0\right\}$, $\lambda \in A_{m,k}(0)$, has VC-dimension V_k less than $m+1$. The shatter coefficient $\mathbb{S}(2n,m,k)$ is related to the VC-dimension of the latter class by the inequality

$$\log \mathbb{S}(2n,m,k) \leq V_k\left\{1 + \log\left(1 + \frac{2n}{V_k}\right)\right\},$$

see, for example, Theorem 4.3 on page 145 of Vapnik (1998). To conclude the proof, use the fact that the right-hand side is an increasing function of V_k . □

Now, using the inequality $D(\delta) + a\|\mathsf{f}_\lambda - f\|^2 \geq 2\sqrt{aD(\delta)}\|\mathsf{f}_\lambda - f\|$ and Theorem 5.3* on page 198 of Vapnik (1998) we get

$$\begin{aligned} &\mathbb{P}\left\{\sup_{\lambda \in A_{m,k}(0)} U_\lambda \geq D(\delta)\right\} \\ &= \mathbb{P}\left\{\exists \lambda \in A_{m,k}(0):\ \|\mathsf{f}_\lambda - f\| \neq 0 \text{ and } (1+a)\left[\|\mathsf{f}_\lambda - f\|^2 - \|\mathsf{f}_\lambda - f\|_n^2\right] \geq D(\delta) + a\|\mathsf{f}_\lambda - f\|^2\right\} \\ &\leq \mathbb{P}\left\{\sup_{\lambda \in A_{m,k}(0):\, \|\mathsf{f}_\lambda - f\| \neq 0} \frac{\|\mathsf{f}_\lambda - f\|^2 - \|\mathsf{f}_\lambda - f\|_n^2}{\|\mathsf{f}_\lambda - f\|} \geq \frac{2\sqrt{aD(\delta)}}{1+a}\right\} \\ &\leq 4\mathbb{S}(2n,m,k)\exp\left\{-\frac{anD(\delta)}{(1+a)^2 B}\right\}. \end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{k=1}^{\binom{M}{m}}\mathbb{P}\left\{\sup_{\lambda\in A_{m,k}(0)} U_\lambda \geq D(\delta)\right\}\\
&\leq 4\sum_{m=1}^{M}\sum_{k=1}^{\binom{M}{m}}\mathbb{S}(2n,m,k)\exp\left\{-\frac{anD(\delta)}{(1+a)^2B}\right\}\\
&\leq 4\sum_{m=1}^{M}\binom{M}{m}\exp\left\{Cm\left[1+\log\left(\frac{n}{m}\right)\right]\right\}\exp\left\{-\frac{aK_1m}{2(1+a)B}\log\left(1+\frac{n\vee M}{m\vee 1}\right)-\frac{an\delta}{2(1+a)^2B}\right\}\\
&\text{by Lemma 3.12}\\
&\leq C_5\exp\left(-C_6\frac{n\delta}{aB}\right),\quad \forall\ a>1,
\end{aligned}$$

for $K_1 = K_1(a,B)$ large enough, and some universal constants $C_5, C_6 > 0$, where we have used the same crude bound for $\binom{M}{m}$ as in the proof of Lemma 3.11. Furthermore,

$$\begin{aligned}
\mathbb{P}\left\{U_0 \geq \delta/2\right\} &\leq \mathbb{P}\left\{\|f\|^2 - \|f\|_n^2 \geq \frac{\sqrt{2a\delta}}{1+a}\|f\|\right\}\\
&\leq \exp\left\{-\frac{an\delta}{(1+a)^2B}\right\} \leq \exp\left\{-\frac{n\delta}{4aB}\right\},\quad \forall\ a>1,
\end{aligned}$$

where the last but one inequality follows, e.g., from Proposition 2.6 in Wegkamp (2003). The exponential bounds in the last two displays and (3.11) easily imply

$$\mathbb{E}_f\left\{\|\widetilde{f}-f\|^2 - (1+a)\|\widetilde{f}-f\|_n^2 - \frac{1+a}{2}\mathrm{pen}(\widehat{\lambda})\right\} \leq C_7\frac{Ba}{n}$$

for some constant $C_7 > 0$. This concludes the proof of Theorem 3.6. □

4. Near optimal aggregation with a data dependent L_1 penalty

We consider here only the fixed design regression. In addition to Assumptions (A1) and (A2), throughout this section we suppose the following.

Assumption (A3) *The matrix*

$$\Psi_n = \left(\frac{1}{n}\sum_{i=1}^{n} f_j(X_i)f_{j'}(X_i)\right)_{1\leq j,j'\leq M}$$

is positive definite for any given $n \geq 1$.

Let $\xi_{\min}$ be the smallest eigenvalue of the matrix Ψ_n. Note that under our assumptions

$$0 < \xi_{\min} \le \|f_j\|_n^2 \le L^2, \quad j = 1, \dots, M. \tag{4.1}$$

We propose the aggregation procedure defined by the following choice of weights:

$$\widehat{\lambda} = \operatorname*{arg\,min}_{\lambda \in \Lambda_{M,T,2}} \left\{ \widehat{S}(\lambda) + \mathrm{pen}(\lambda) \right\} \tag{4.2}$$

where

$$\Lambda_{M,T,2} = \left\{ \lambda \in \mathbb{R}^M : \sum_{j=1}^M \lambda_j^2 \le T^2 \right\},$$

for $T > 0$ large enough, and the penalty term is given by

$$\mathrm{pen}(\lambda) = \sum_{j=1}^M r_{n,j} |\lambda_j| \quad \text{with} \quad r_{n,j} = 2\sqrt{2}\sigma \|f_j\|_n \sqrt{\frac{2\log M + \log n}{n}}. \tag{4.3}$$

THEOREM 4.1. *Let $X_i \in \mathcal{X}$, $i = 1, \dots, n$, be fixed. Let $\widehat{\lambda}$ be the penalized least squares estimate defined by (4.2) with penalty (4.3). Set $\widetilde{f} = \mathsf{f}_{\widehat{\lambda}}$. Let $T > 0$ be such that $T^2 \xi_{\min} > 2L^2$. Then, for all $a > 1$, and all integers $n \ge 1$, $M \ge 2$, we have,*

$$\begin{aligned} \mathbb{E}_f \|\widetilde{f} - f\|_n^2 \quad \le \quad & \inf_{\lambda \in \mathbb{R}^M} \left\{ \frac{a+1}{a-1} \|\mathsf{f}_\lambda - f\|_n^2 + \frac{16a^2}{a-1} \left(\frac{\sigma^2 L^2}{\xi_{\min}} \right) \frac{2\log M + \log n}{n} M(\lambda) \right\} \\ & + \frac{(T + M^{-1/2})^2 L^2}{n\sqrt{\pi(2\log M + \log n)}}. \end{aligned} \tag{4.4}$$

COROLLARY 4.2 (MS). *Let assumptions of Theorem 4.1 be satisfied and $T \le (\log(M \vee n))^{1/4}$. Then there exists a constant $C = C(T, L, \sigma^2, \xi_{\min}) > 0$ such that for all $\varepsilon > 0$ and for all integers $n \ge 1$ and $M \ge 2$,*

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

Proof. Using assumptions on T and (4.1), we trivially get $T > \sqrt{2L^2/\xi_{\min}} \ge M^{-1/2}$. This implies that the last summand in (4.4) is $O(1/n)$. The rest of the proof is analogous to that of Corollary 3.2. □

COROLLARY 4.3 (C). *Let assumptions of Theorem 4.1 be satisfied and $T \le (\log(M \vee n))^{1/4}$. Then there exists a constant $C = C(T, L, \sigma^2, \xi_{\min}) > 0$ such that for all $\varepsilon > 0$ and for all integers $n \ge 1$ and $M \ge 2$,*

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

where

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

Proof. We bound the last summand in (4.4) as in the previous proof and we use then the argument similar to that of the proof of Corollary 3.4. □

COROLLARY 4.4 (L). *Let assumptions of Theorem 4.1 be satisfied and* $T \le (\log(M \vee n))^{1/4}$. *Then there exists a constant* $C = C(T, L, \sigma^2, \xi_{\min}) > 0$ *such that for all* $\varepsilon > 0$ *and for all integers* $n \ge 1$ *and* $M \ge 2$,

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

Proof. We bound the last summand in (4.4) as in the proof Corollary 4.2 and we use that $M(\lambda) \le M$. □

Proof of Theorem 4.1. We begin as in Loubes and Van de Geer (2002). 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 $\lambda \in \Lambda_{M,T,2}$, which we may rewrite as

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

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

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

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

$$\begin{aligned} \mathbb{P}(A^c) &\le \sum_{j=1}^M \mathbb{P}\{\sqrt{n}|V_j| > \sqrt{n} r_{n,j}/2\} \le \sum_{j=1}^M \frac{4}{\sqrt{2\pi}} \frac{\sigma \|f_j\|_n}{\sqrt{n} r_{n,j}} \exp\left(-\frac{n r_{n,j}^2}{8\sigma^2 \|f_j\|_n^2}\right) \\ &= \frac{1}{Mn\sqrt{\pi(2\log M + \log n)}}. \end{aligned} \tag{4.5}$$

Then, on the set A, we find

$$2\left\langle W, \widetilde{f} - f \right\rangle_n = 2\sum_{j=1}^{M} V_j(\widehat{\lambda}_j - \lambda_j) \le \sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j - \lambda_j|$$

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,

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

by the triangle inequality and the fact that $\lambda_j = 0$ for $j \notin J(\lambda)$. Since $\xi_{\min} > 0$, we have

$$\xi_{\min}^{-1}\|\widetilde{f} - \mathsf{f}_\lambda\|_n^2 \ge \sum_{j \in J(\lambda)} |\widehat{\lambda}_j - \lambda_j|^2.$$

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

$$\begin{aligned}
\|\widetilde{f} - f\|_n^2 \;&\le\; \|\mathsf{f}_\lambda - f\|_n^2 + 2\sum_{j \in J(\lambda)} r_{n,j}|\widehat{\lambda}_j - \lambda_j| \\
&\le\; \|\mathsf{f}_\lambda - f\|_n^2 + 2\sqrt{\xi_{\min}^{-1}}\sqrt{\sum_{j \in J(\lambda)} r_{n,j}^2}\left(\|\widetilde{f} - f\|_n + \|\mathsf{f}_\lambda - f\|_n\right) \\
&\le\; \|\mathsf{f}_\lambda - f\|_n^2 + 2\sqrt{\xi_{\min}^{-1}}\, r_n \sqrt{M(\lambda)}\left(\|\widetilde{f} - f\|_n + \|\mathsf{f}_\lambda - f\|_n\right),
\end{aligned} \tag{4.6}$$

where

$$r_n \stackrel{\text{def}}{=} 2\sqrt{2}\, L\sigma\sqrt{\frac{2\log M + \log n}{n}}.$$

Inequality (4.6) is of the simple form $v^2 \le c^2 + vb + cb$ with $v = \|\widetilde{f} - f\|_n$, $b = 2r_n\sqrt{M(\lambda)/\xi_{\min}}$ 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 (4.6) is valid on the set

A, we now get that

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

Consequently, since by the Cauchy-Schwarz inequality,

$$\|\widetilde{f}-f\|_\infty \le L(\sum_{j=1}^M|\lambda_j|+1)\le(\sqrt{M}T+1)L,$$

we find

$$\begin{aligned}\mathbb{E}_f\|\widetilde{f}-f\|_n^2 &\le \mathbb{E}_f\left[\|\widetilde{f}-f\|_n^2 I_A\right]+(\sqrt{M}T+1)^2L^2\mathbb{P}(A^c)\\ &\le \inf_{\lambda\in\Lambda_{M,T,2}}\left\{\frac{a+1}{a-1}\|\mathsf{f}_\lambda-f\|_n^2+\frac{2a^2r_n^2}{(a-1)\xi_{\min}}M(\lambda)\right\}\\ &\qquad+\frac{(T+M^{-1/2})^2L^2}{n\sqrt{\pi(2\log M+\log n)}}.\end{aligned}\tag{4.7}$$

It remains to show that (4.7) remains valid with the set $\Lambda_{M,T,2}$ replaced by the entire $\mathbb{R}^M$. For this, observe that $\lambda\notin\Lambda_{M,T,2}$ implies $\sum_{j=1}^M\lambda_j^2>T^2$, and thus $\|\mathsf{f}_\lambda\|_n^2\ge\xi_{\min}\sum_{j=1}^M\lambda_j^2>\xi_{\min}T^2$. Therefore, for $\lambda\notin\Lambda_{M,T,2}$, we have

$$\|\mathsf{f}_\lambda-f\|_n\ge\|\mathsf{f}_\lambda\|_n-\|f\|_n>\sqrt{\xi_{\min}}T-L>L$$

by our choice of T. On the other hand, for $\lambda=0\in\Lambda_{M,T,2}$, we have

$$\|\mathsf{f}_\lambda-f\|_n=\|f\|_n\le L$$

and $\mathrm{pen}(0)=0$. Thus, the value of the whole expression under the infimum in (4.7) for $\lambda=0$ is strictly smaller than the value of this expression for any $\lambda\notin\Lambda_{M,T,2}$, which proves the result. □

As in Section 3.1, we present now a statement in probability that complements the results of this section.

THEOREM 4.5. *Let $X_i\in\mathcal{X}$, $i=1,\dots,n$, be fixed. Let $\widehat{\lambda}$ be the penalized least squares estimate defined by (4.2) with $\Lambda_{M,T,2}$ replaced by $\mathbb{R}^M$ and with penalty (4.3). Set $\widetilde{f}=\mathsf{f}_{\widehat{\lambda}}$. Then, for all $a>1$, and all integers $n\ge1$, $M\ge2$, we have,*

$$\begin{aligned}&\mathbb{P}\left(\|\widetilde{f}-f\|_n^2\ge\inf_{\lambda\in\mathbb{R}^M}\left\{\frac{a+1}{a-1}\|\mathsf{f}_\lambda-f\|_n^2+\frac{16a^2}{a-1}\left(\frac{\sigma^2L^2}{\xi_{\min}}\right)\frac{2\log M+\log n}{n}M(\lambda)\right\}\right)\\ &\qquad\le\frac{1}{Mn\sqrt{\pi(2\log M+\log n)}}.\end{aligned}\tag{4.8}$$

Proof. This result follows directly from the proof of Theorem 4.1. Note first that now (4.6) is valid for all $\lambda\in\mathbb{R}^M$ and not only for $\lambda\in\Lambda_{M,T,2}$. Using (4.6) and the argument after it we

find that the left hand side in (4.8) can be bounded by $\mathbb{P}(A^c)$. The result follows by invoking (4.5).

□

REMARKS.

1. The method presented in this section is not strictly an L_1-penalized one. Indeed, it implements two penalties: the data dependent L_1-penalty $\sum_{j=1}^M r_{n,j}|\lambda_j|$, and the L_2-penalty $\sum_{j=1}^M \lambda_j^2$ that appears implicitly via the choice of the set $\Lambda_{M,T,2}$. The resulting minimization problem can be solved in practice using standard convex programming software. The L_2 part of the penalty is less influential, since it should typically be applied with $T \to \infty$ as M (respectively n) grows, which means that the restriction to $\Lambda_{M,T,2}$ becomes asymptotically negligible. Moreover, the restriction is not always needed. For example, the bound in probability (Theorem 4.5) is obtained for $\widehat{\lambda}$ that minimizes the L_1-penalized least squares over the entire $\mathbb{R}^M$.

2. Assumption (A3) is mild, and it is also made by Efron *et al.* (2004) in the context of LARS. In practice, this assumption can always be checked. A stronger assumption is that $\xi_{\min} > c$ for some constant $c > 0$, independent of n and M if one or both of these parameters are allowed to grow (which is typically the more interesting case). There are at least two important examples where such a stronger assumption holds. The first example is standard in the parametric regression context: M is fixed and $\Psi_n/n \to \Psi$ where Ψ is a nonsingular $M \times M$ matrix. The second one is related to nonparametric regression: $M = M_n$ is allowed to go to ∞ as $n \to \infty$ and the functions f_j are orthogonal with respect to the empirical norm. This corresponds, for instance, 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 may lead to adaptive estimators with good asymptotic properties [*cf., e.g.*, Nemirovski (2000)]. Local image smoothing provides us an application where the condition $\xi_{\min} > c$ is naturally satisfied. For example, Katkovnik *et al.* (2002, 2004) suggest different methods of aggregation of local image estimators obtained from non-intersecting sectors around a given pixel (these estimators are mutually orthogonal with respect to the empirical norm).

3. Inspection of the proofs shows that the constants $C = C(T, L, \sigma^2, \xi_{\min})$ in Corollaries 4.2, 4.3, 4.4 have the form $C = A_1 + A_2\xi_{\min}^{-1}$, where A_1 and A_2 are constants independent of $\xi_{\min}$. In general, $\xi_{\min}$ may depend on n and M. However, if $\xi_{\min} > c$ for some constant $c > 0$, independent of n and M, as previously discussed, the rates of aggregation given in Corollaries 4.2, 4.3, 4.4 are near optimal, up to logarithmic factors. They are even exactly optimal (*cf.* (1.3) and the lower bounds of the next section) for some configurations of n, M: for (MS)-aggregation if $n^{a'} \le M \le n^a$, and for (C)-aggregation if $n^{1/2} \le M \le n^a$, where $0 < a' < a < \infty$.

4. From the bound in Theorem 4.1, we see that T is allowed to grow with n and M (as fast as $T \asymp (\log(M \vee n))^{1/4}$ is possible). Moreover, the proof of Theorem 4.1 reveals that by taking a larger constant than $2\sqrt{2}$ in (4.3), even faster rates are allowed, for example, T can grow as a power of n. This may be needed to guarantee the condition $T^2 > 2L^2/\xi_{\min}$ for n large enough, because the value L is typically not known and $\xi_{\min}$ may depend on n and M. However, the condition $T^2 > 2L^2/\xi_{\min}$ is only needed to cover the linear aggregation. For (MS) and (C) aggregation, Corollaries 4.2, 4.3 can be obtained directly from (4.7), and thus it suffices to take any $T \ge 1$, since $\Lambda^M \subset \Lambda_{M,1,2}$, or to replace $\Lambda_{M,T,2}$ by Λ^M in the definition of $\widehat{\lambda}$.

5. Lower bounds

For regression with random design and the $L_2(\mathbb{R}^d, \mu)$-risks, lower bounds for aggregation and optimal rates $\psi_{n,M}$ as given in (1.3) were established by Tsybakov (2003). In this section we extend the lower bounds of Tsybakov (2003) for (MS) and (L) aggregation to regression with fixed design. Further, we state these bounds in a more general form, considering not only the expected squared risks, but also other loss functions. This generalization allows one to treat optimality of the upper bounds "in probability" obtained in the previous sections (Theorems 3.5, 4.5). It shows that the remainder terms in these bounds are optimal or near optimal for the (MS) and (L) aggregation.

In this section we suppose that $X_1, \dots, X_n$ are fixed and that $M \le n$. 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 $X_i \in \mathcal{X}$, $i = 1, \dots, n$, be fixed and $2 \le M \le n$. Assume that H^M is either the whole $\mathbb{R}^M$ (the (L) aggregation case) or the set of vertices of Λ^M (the (MS) aggregation case). Let the corresponding $\psi_{n,M}$ be given by (1.3) and let $M \log M \le n$ for the case of (MS)*

aggregation. Then there exist $f_1, \dots, f_M \in \mathcal{F}_0$ *such that, for any loss function* $w(\cdot)$,

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

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

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, 3.3, 4.2, 4.4. 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 (see Theorems 3.5, 4.5).

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

LEMMA 5.2. *Let* w *be a loss function,* $A > 0$ *be such that* $w(A) > 0$, *and let* $\mathcal{C}$ *be a finite set of functions on* $\mathcal{X}$ *such that* $N = \operatorname{card}(\mathcal{C}) \geq 2$,

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

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

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

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

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

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

The (MS) aggregation case. Let H^M be the set of vertices of Λ^M, $M \log M \leq n$, and $\psi_{n,M} = (\log M)/n$. Pick M disjoint subsets $S_1, \dots, S_M$ of $\{X_1, \dots, X_n\}$, each S_j of cardinality $\log M$ (w.l.o.g. we assume that $\log M$ is an integer) and define the functions

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

where $\gamma \leq L$ is a positive constant to be chosen. Clearly, $\{f_1, \dots, f_M\} \subset \mathcal{F}_0$. Thus, it suffices to prove the lower bound of the theorem where the supremum over $f \in \mathcal{F}_0$ is replaced by that over $f \in \{f_1, \dots, f_M\}$. But for such f we have $\min_{1 \leq j \leq M} \|f_j - f\|_n^2 = 0$, and to finish the proof for the (MS) case, it is sufficient to bound from below the quantity

$\sup_{f\in\{f_1,\dots,f_M\}} \mathbb{E}_f w(\psi_{n,M}^{-1}\|T_n - f\|_n^2)$, where $\psi_{n,M} = (\log M)/n$, uniformly over all estimators T_n. This is done by applying Lemma 5.2. In fact, note that, for $j \neq k$,

$$\|f_j - f_k\|_n^2 = \frac{2\gamma^2 \log M}{n} \stackrel{\text{def}}{=} 4s^2. \tag{5.2}$$

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

$$K(\mathbb{P}_{f_j}, \mathbb{P}_{f_k}) = \frac{n}{2\sigma^2}\|f_j - f_k\|_n^2, \quad j = 1,\dots,M. \tag{5.3}$$

In view of (5.2) and (5.3), one can choose γ small enough to have $K(\mathbb{P}_{f_j}, \mathbb{P}_{f_k}) \le (1/16)\log M$ for $j,k = 1,\dots,M$. Now, to get the lower bound for the (MS) case, it remains to use this inequality, identity (5.2) and Lemma 5.2.

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

$$\mathcal{U} = \Big\{ g = \sum_{j=1}^M \omega_j f_j : \omega \in \Omega \Big\}, \tag{5.4}$$

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

$$\|g_1 - g_2\|_n^2 = \frac{\gamma^2}{n}\sum_{j=1}^M (\omega_j - \omega_j')^2 \le \gamma^2 M/n. \tag{5.5}$$

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

$$\|g_1 - g_2\|_n^2 \ge C_1\gamma^2 M/n. \tag{5.6}$$

for any $g_1, g_2 \in \mathcal{U}_0$. Using (5.3) and (5.5) we get, for any $g_1, g_2 \in \mathcal{U}_0$,

$$K(\mathbb{P}_{g_1}, \mathbb{P}_{g_2}) \le C_2\gamma^2 M \le C_3\gamma^2 \log(\mathrm{card}(\mathcal{U}_0)),$$

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 5.2 to the set $\mathcal{U}_0 = \{f_1, f_2\}$. □

Appendix A

Lemma A.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 Corollary 3.4. Then (3.2) holds and*

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

Proof. Let f^* be the minimizer of $\|\mathsf{f}_\lambda - f\|^2$ 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

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

Integrating (A.2) over $\mu(dx)$ and recalling the definition of f^* we obtain

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

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

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

This and (A.3) prove (A.1). The proof of (3.2) is analogous, with the only difference that (A.2) is integrated over the empirical measure rather than over $\mu(dx)$. □

References

1. Audibert, J.-Y. (2004) Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré (B), Probability and Statistics*, 40: 685 – 736.

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

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

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

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

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

7. 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 (available at "http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2003").

8. Breiman, L. (1996). Stacked regression. *Machine Learning*, 24: 49 – 64.

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

10. Bunea, F. and Wegkamp, M.H. (2004). Two-stage model selection procedures in partially linear regression. *Canadian Journal of Statistics* 32: 105-118.

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

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

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

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

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

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

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

18. Katkovnik, V., Egiazarian, K. and Astola, J. (2002) Adaptive Window Size Image De-noising Based on Intersection of Confidence Intervals (ICI) Rule. *Journal of Mathematical Imaging and Vision*, 16:223-235.

19. Katkovnik, V., Foi, A., Egiazarian, K. and Astola, J. (2004) To be published in *Proceedings of 12 th European Signal Processing Conference, EUSIPCO-2004*, September 6-10, 2004, Vienna

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

21. Koltchinskii, V. (2004). Local Rademacher complexities and oracle inequalities in risk minimization. *Manuscript.*

22. LeBlanc, M. and Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association*, 91: 1641 – 1650.

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

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

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

26. 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.

27. Panchenko, D. (2003) Symmetrization approach to concentration inequalities in empirical processes. *Annals of Probability*, 31, 2068 – 2081.

28. Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

29. Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Annals of Statistics*, 12: 1298 – 1309.

30. Talagrand, M. (1996a) A new look at independence. *Annals of Probability*, 24: 1 – 34.

31. Talagrand, M. (1996b) New concentration inequalities in product spaces. *Invent. Math.*, 126: 505 – 563.

32. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*. 58: 267 – 288.

33. 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.

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

35. van de Geer, S. (2000). *Empirical Processes in M-Estimation*, Cambridge Univ. Press.

36. Vapnik, V. N. (1998). *Statistical Learning Theory.* Wiley, New York.

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

38. Wolpert, D. (1992). Stacked generalization. *Neural Networks*, 5: 241 – 259.

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.

42. Yohai, V.J. and Maronna, R.A. (1979). Asymptotic behavior of M-estimators for the linear model. *Annals of Statistics*, 7: 258 – 268.

Florentina Bunea, Department of Statistics, Florida State University, Tallahassee, Florida.
E-mail address: bunea@stat.fsu.edu

Alexandre B. Tsybakov, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI, France.
E-mail address: tsybakov@ccr.jussieu.fr

Marten H. Wegkamp, Department of Statistics, Florida State University, Tallahassee, Florida.
E-mail address: wegkamp@stat.fsu.edu