Journal of Machine Learning Research 12 (2011) 1813-1833 Submitted 12/09; Revised 9/10; Published 6/11

# Hyper-Sparse Optimal Aggregation

**Stéphane Gaïffas** STEPHANE.GAIFFAS@UPMC.FR
*Laboratoire de Statistique Théorique et Appliquée*
*Université Pierre et Marie Curie - Paris 6*
*75005, Paris, FRANCE*

**Guillaume Lecué** GUILLAUME.LECUE@UNIV-MLV.FR
*CNRS, Laboratoire d'Analyse et Mathématiques appliquées*
*Université Paris-Est - Marne-la-vallée*
*77454, Marne-la-Valle Cedex 2, FRANCE*

**Editor:** John Shawe-Taylor

## Abstract

Given a finite set $F$ of functions and a learning sample, the aim of an aggregation procedure is to have a risk as close as possible to risk of the best function in $F$. Up to now, optimal aggregation procedures are convex combinations of every elements of $F$. In this paper, we prove that optimal aggregation procedures combining only two functions in $F$ exist. Such algorithms are of particular interest when $F$ contains many irrelevant functions that should not appear in the aggregation procedure. Since selectors are suboptimal aggregation procedures, this proves that two is the minimal number of elements of $F$ required for the construction of an optimal aggregation procedure in every situations. Then, we perform a numerical study for the problem of selection of the regularization parameters of the Lasso and the Elastic-net estimators. We compare on simulated examples our aggregation algorithms to aggregation with exponential weights, to Mallow's $C_p$ and to cross-validation selection procedures.

**Keywords:** aggregation, exact oracle inequality, empirical risk minimization, empirical process theory, sparsity, Lasso, Lars

## 1. Introduction

Let $(\Omega,\mu)$ be a probability space and $\nu$ be a probability measure on $\Omega\times\mathbb{R}$ such that $\mu$ is its marginal on $\Omega$. Assume $(X,Y)$ and $D_n := (X_i,Y_i)_{i=1}^n$ to be $n+1$ independent random variables distributed according to $\nu$, and that we are given a finite set $F=\{f_1,\dots,f_M\}$ of real-valued functions on $\Omega$, usually called a *dictionary*, or a set of *weak learners*. This set of functions is often a set of estimators computed on a *training* sample, which is independent of the sample $D_n$ (*learning* sample).

We consider the problem of prediction of $Y$ from $X$ using the functions given in $F$ and the sample $D_n$. If $f:\Omega\to\mathbb{R}$, we measure its error of prediction, or risk, by the expectation of the squared loss

$$R(f)=\mathbb{E}(f(X)-Y)^2.$$

If $\widehat{f}$ depends on $D_n$, its risk is the conditional expectation

$$R(\widehat{f})=\mathbb{E}[(\widehat{f}(X)-Y)^2|D_n].$$

The aim of the problem of aggregation is to construct a procedure $\tilde{f}$ (called an *aggregate*) using $D_n$ and $F$ with a risk which is very close to the smallest risk over $F$. Namely, one wants to prove that $\tilde{f}$

satisfies an inequality of the form

$$R(\tilde{f}) \leq \min_{f \in F} R(f) + r(F,n) \tag{1}$$

with a large probability or in expectation. Inequalities of the form (1) are called *exact oracle inequalities* and $r(F,n)$ is called the *residue*. A classical result (Juditsky et al., 2008) says that aggregates with values in $F$ cannot satisfy an inequality like (1) with a residue smaller than $((\log M)/n)^{1/2}$ for every $F$. Nevertheless, it is possible to mimic the oracle (an *oracle* is a element in $F$ achieving the minimal risk over $F$) up to the residue $(\log M)/n$ (see Juditsky et al., 2008 and Lecué and Mendelson, 2009, among others) using an aggregate $\tilde{f}$ that combines all the elements of $F$. In this case, we say that $\tilde{f}$ is an *optimal aggregation procedure*. This notion of optimality is given in Tsybakov (2003) and Lecué and Mendelson (2009), and it is the one we will refer to in this paper.

Given the set of functions $F$, a natural way to predict $Y$ is to compute the empirical risk minimization procedure (ERM), the one that minimizes the empirical risk

$$R_n(f) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2$$

over $F$. This very basic principle is at the core of aggregation procedures (for regression with squared loss). An aggregate is typically represented as a convex combination of the elements of $F$. Namely,

$$\widehat{f} := \sum_{j=1}^{M} \theta_j(D_n,F) f_j,$$

where $(\theta_j(D_n,F))_{j=1}^M$ is a vector of non-negative coordinates suming to 1. Up to now, most of the optimal aggregation procedures are based on exponential weights: aggregation with cumulated exponential weights (ACEW), see Catoni (2001), Yang (2004), Yang (2000), Juditsky et al. (2008), Juditsky et al. (2005), Audibert (2009) and aggregation with exponential weights (AEW), see Leung and Barron (2006) and Dalalyan and Tsybakov (2007), among others. The weights of the ACEW are given by

$$\theta_j^{(\mathrm{ACEW})} := \frac{1}{n}\sum_{k=1}^{n} \frac{\exp(-R_k(f_j)/T)}{\sum_{l=1}^{M}\exp(-R_k(f_l)/T)},$$

where $T$ is the so-called *temperature* parameter. The weights of the AEW are given by

$$\theta_j^{(\mathrm{AEW})} := \frac{\exp(-R_n(f_j)/T)}{\sum_{l=1}^{M}\exp(-R_n(f_j)/T)}.$$

The ACEW satisfies (1) for $r(F,n) \sim (\log M)/n$, see references above, so it is optimal in the sense of Tsybakov (2003). The AEW has been proved to be optimal in the regression model with deterministic design for large temperatures in Dalalyan and Tsybakov (2007). Altough, for small temperatures, AEW can be suboptimal both in expectation and with large probability (cf. Lecué and Mendelson, 2010).

In these aggregates, no coefficient $\theta_j$ is equal to zero, although they can be very small, depending on the value of $R_n(f_j)$ and $T$ (this makes in particular the choice of $T$ of importance). So, even the worse elements of $F$ have an influence on the aggregate. This can be a problem when one wants to use aggregation to construct adaptive procedures. Indeed, one could imagine large dictionaries

containing many different types of estimators (kernel estimators, projection estimators, etc.) with many different parameters (smoothing parameters, groups of variables, etc.). Some of the estimators are likely to be more adapted than the others, depending on the kind of models that fits well the data, and, there may be only few of them among a large dictionary. An aggregate that combines only the most adapted estimators from the dictionary and that removes the irrelevant ones is suitable in this case. The challenge is then to find such a procedure which is still an optimal aggregate. An improvement going in this direction has been made using a preselection step in Lecué and Mendelson (2009). This preselection step allows to remove all the estimators in $F$ which performs badly on a learning subsample. In this paper, we want to go a step further: we look for an aggregation algorithm that shares the same property of optimality, but with as few non-zero coefficients $\theta_j$ as possible, hence the name *hyper-sparse aggregate*. This leads to the following question:

**Question 1** *What is the minimal number of non-zero coefficients $\theta_j$ such that an aggregation procedure $\tilde{f} = \sum_{j=1}^{M} \theta_j f_j$ is optimal?*

It turns out that the answer to Question 1 is two. Indeed, if every coefficient is zero, excepted for one, the aggregate coincides with an element of $F$, and as we mentioned before, such a procedure can only achieve the rate $((\log M)/n)^{1/2}$ (unless extra properties are satisfied by $F$ and $\nu$). In Definition 1 below (see Section 2) we construct three procedures, where two of them (see (6) and (7)) only have two non-zero coefficients $\theta_j$. We prove in Theorem 2 below that these procedures are optimal, since they achieve the rate $(\log M)/n$.

## 2. Definition of the Aggregates and Results

First, we need to introduce some notations and assumptions. Let us recall that the $\psi_1$-norm of a random variable $Z$ is given by $\|Z\|_{\psi_1} := \inf\{c > 0 : \mathbb{E}[\exp(|Z|/c))] \leq 2\}$. We say that $Z$ is sub-exponential when $\|Z\|_{\psi_1} < +\infty$. We work under the following assumptions.

**Assumption 1** *We can write*

$$Y = f_0(X) + \varepsilon,$$

*where $\varepsilon$ is such that $\mathbb{E}(\varepsilon|X) = 0$ and $\mathbb{E}(\varepsilon^2|X) \leq \sigma_\varepsilon^2$ a.s. for some constant $\sigma_\varepsilon > 0$. Moreover, we assume that one of the following points holds.*

- *(Bounded setup) There is a constant $b > 0$ such that:*

$$\max\left(\|Y\|_\infty, \sup_{f\in F} \|f(X)\|_{L_\infty}\right) \leq b. \tag{2}$$

- *(Sub-exponential setup) There is a constant $b > 0$ such that:*

$$\max\left(\|\varepsilon\|_{\psi_1}, \sup_{f\in F} \|f(X) - f_0(X)\|_{L_\infty}\right) \leq b. \tag{3}$$

Note that Assumption (3) allows for an unbounded output $Y$. The results given below differ a bit depending on the considered assumption (there is an extra $\log n$ term in the sub-exponential case). To simplify the notations, we assume from now on that we have $2n$ observations from a sample $D_{2n} = (X_i, Y_i)_{i=1}^{2n}$. Let us define our aggregation procedures.

**Definition 1 (Aggregation procedures)** *Follow the following steps:*

**(0. Initialization)** *Choose a confidence level $x > 0$. If* (2) *holds, define*

$$\phi = \phi_{n,M}(x) = b\sqrt{\frac{\log M + x}{n}}.$$

*If* (3) *holds, define*

$$\phi = \phi_{n,M}(x) = (\sigma_\varepsilon + b)\sqrt{\frac{(\log M + x)\log n}{n}}.$$

**(1. Splitting)** *Split the sample $D_{2n}$ into $D_{n,1} = (X_i, Y_i)_{i=1}^n$ and $D_{n,2} = (X_i, Y_i)_{i=n+1}^{2n}$.*

**(2. Preselection)** *Use $D_{n,1}$ to define a random subset of $F$ :*

$$\widehat{F}_1 = \Big\{ f \in F : R_{n,1}(f) \leq R_{n,1}(\widehat{f}_{n,1}) + c\max\left(\phi \|\widehat{f}_{n,1} - f\|_{n,1}, \phi^2\right) \Big\}, \tag{4}$$

*where $\|f\|_{n,1}^2 = n^{-1}\sum_{i=1}^n f(X_i)^2$, $R_{n,1}(f) = n^{-1}\sum_{i=1}^n (f(X_i) - Y_i)^2$, $\widehat{f}_{n,1} \in \operatorname{argmin}_{f\in F} R_{n,1}(f)$.*

**(3. Aggregation)** *Choose $\widehat{\mathcal{F}}$ as one of the following sets:*

$$\widehat{\mathcal{F}} = \operatorname{conv}(\widehat{F}_1) = \text{ the convex hull of } \widehat{F}_1 \tag{5}$$

$$\widehat{\mathcal{F}} = \operatorname{seg}(\widehat{F}_1) = \text{ the segments between the functions in } \widehat{F}_1 \tag{6}$$

$$\widehat{\mathcal{F}} = \operatorname{star}(\widehat{f}_{n,1}, \widehat{F}_1) = \text{ the segments between } \widehat{f}_{n,1} \text{ with the elements of } \widehat{F}_1, \tag{7}$$

*and return the ERM relative to $D_{n,2}$ :*

$$\tilde{f} \in \operatorname*{argmin}_{g\in\widehat{\mathcal{F}}} R_{n,2}(g),$$

*where $R_{n,2}(f) = n^{-1}\sum_{i=n+1}^{2n}(f(X_i) - Y_i)^2$.*

These algorithms are illustrated in Figures 1 and 2. In Figure 1 we summarize the aggregation steps in the three cases. In Figure 2 we give a simulated illustration of the preselection step, and we show the value of the weights of the AEW for a comparison. As mentioned above, the Step 3 of the algorithm returns, when $\widehat{\mathcal{F}}$ is given by (6) or (7), an aggregate which is a convex combination of only two functions in $F$, among the ones remaining after the preselection step. The preselection step was introduced in Lecué and Mendelson (2009), with the use of (5) only for the aggregation step.

From the computational point of view, the procedure (7) is the most appealing: an ERM in $\operatorname{star}(\widehat{f}_{n,1}, \widehat{F})$ can be computed in a fast and explicit way, see Algorithm 1 below. The next Theorem proves that each procedure given in Definition 1 are optimal.

**Theorem 2** *Let $x > 0$ be a confidence level, $F$ be a dictionary with cardinality $M$ and $\tilde{f}$ be one of the aggregation procedure given in Definition 1. If* (2) *holds, we have, with $\nu^{2n}$-probability at least $1 - 2e^{-x}$:*

$$R(\tilde{f}) \leq \min_{f\in F} R(f) + c_b\frac{(1+x)\log M}{n},$$

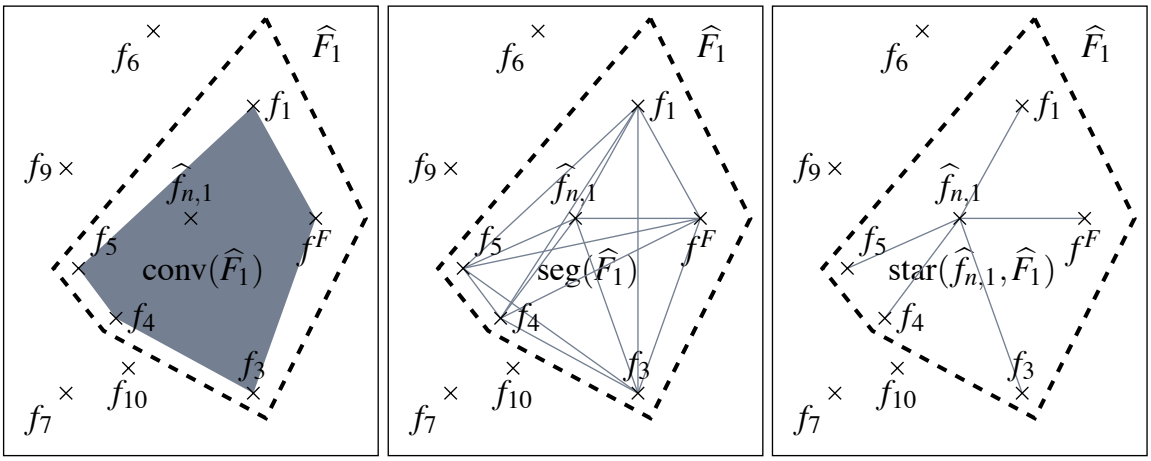


Figure 1: Aggregation algorithms: ERM over $\mathrm{conv}(\widehat{F}_1)$, $\mathrm{seg}(\widehat{F}_1)$, or $\mathrm{star}(\widehat{f}_{n,1},\widehat{F}_1)$.

*where $c_b$ is a constant depending on $b$.*

*If* (3) *holds, we have, with $\nu^{2n}$-probability at least $1-4e^{-x}$:*

$$R(\tilde{f}) \le \min_{f\in F} R(f) + c_{\sigma_\varepsilon,b}\frac{(1+x)\log M\log n}{n}.$$

**Remark 3** *Note that the definition of the set $\widehat{F}_1$, and thus $\tilde{f}$, depends on the confidence $x$ through the factor $\phi_{n,M}(x)$.*

**Remark 4** *To simplify the proofs, we don't give the explicit values of the constants. However, when* (2) *holds, one can choose $c = 4(1+9b)$ in* (4) *and $c = c_1(1+b)$ when* (3) *holds (where $c_1$ is the absolute constant appearing in Theorem 6). Of course, this is not likely to be the optimal choice.*

Now, we give details for the computation of the star-shaped aggregate, namely the aggregate $\tilde{f}$ given by Definition 1 when $\widehat{\mathcal{F}}$ is (7). Indeed, if $\lambda \in [0,1]$, we have

$$R_{n,2}(\lambda f + (1-\lambda)g) = \lambda R_{n,2}(f) + (1-\lambda)R_{n,2}(g) - \lambda(1-\lambda)\|f-g\|_{n,2}^2,$$

so the minimum of $\lambda \mapsto R_{n,2}(\lambda f + (1-\lambda)g)$ is achieved at

$$\lambda_{n,2}(f,g) = 0 \vee \frac{1}{2}\Big(\frac{R_{n,2}(g) - R_{n,2}(f)}{\|f-g\|_{n,2}^2} + 1\Big) \wedge 1,$$

where $a\vee b = \max(a,b)$, $a\wedge b = \min(a,b)$. So,

$$\min_{\lambda\in[0,1]} R_{n,2}(\lambda f + (1-\lambda)g) = R_{n,2}(\lambda_{n,2}(f,g)f + (1-\lambda_{n,2}(f,g))g),$$

which is equal to

$$\begin{aligned} &R_{n,2}(g) \quad \text{if} \quad R_{n,2}(f) - R_{n,2}(g) > \|f-g\|_{n,2}^2, \\ &R_{n,2}(f) \quad \text{if} \quad R_{n,2}(f) - R_{n,2}(g) < -\|f-g\|_{n,2}^2, \end{aligned}$$

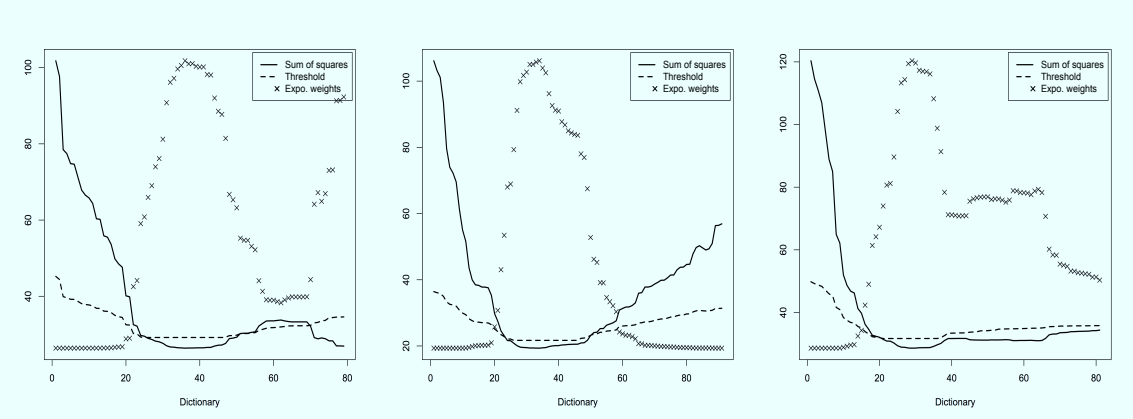


Figure 2: Empirical risk $R_{n,1}(f)$, value of the threshold $R_{n,1}(\widehat{f}_{n,1}) + 2\max(\phi\|\widehat{f}_{n,1} - f\|_{n,1}, \phi^2)$ and weights of the AEW (rescaled) for $f \in F$, where $F$ is a dictionary obtained using LARS, see Section 3 below. Only the elements of $F$ with an empirical risk smaller than the threshold are kept from the dictionary for the construction of the aggregates of Definition (1). The first and third examples correspond to a case where an aggregate with preselection step improves upon AEW, while in the second example, both procedures behaves similarly.

and to

$$\frac{R_{n,2}(f) + R_{n,2}(g)}{2} - \frac{(R_{n,2}(f) - R_{n,2}(g))^2}{4\|f-g\|_{n,2}^2} - \frac{\|f-g\|_{n,2}^2}{4}$$

if $|R_{n,2}(f) - R_{n,2}(g)| \leq \|f-g\|_{n,2}^2$. This leads to the next Algorithm 1 for the computation of $\tilde{f}$.

## 3. Simulation Study

In machine learning, the choice of the tuning parameters in a procedure based on penalization is a main issue. If the procedure is able to perform variable selection (such as the Lasso, see Tibshirani, 1996), then the tuning parameters determines which variables are selected. In many cases, including the Lasso, this choice is commonly done using a Mallow's $C_p$ heuristic (see Efron et al., 2004) or using the $V$-fold or the leave-one-out cross validations. Since aggregation procedures are known (see references above) to outperform selectors in terms of prediction error, it is tempting to use aggregation for the choice of the tuning parameters. Unfortunately, as we mentioned before, most aggregation procedures provide non-zero weights to many non relevant element in a dictionary: this is a problem for variable selection. Indeed, if we use, for instance, the AEW on a dictionary consisting of the full path of Lasso estimators (provided by the Lars algorithm, see Efron et al., 2004), then the resulting aggregate is likely to select all the variables since the Lasso with a small regularization parameter is very close (and equal if it is zero) to ordinary least-squares (which does not perform any variables selection). So, in this context, the hyper-sparse aggregate of Section 2 is of particular interest. In this section, we compare the prediction error and the accuracy of variable selection of our star-shaped aggregation algorithm to Mallow's $C_p$ heuristic, leave-one-out cross-validation and 10-fold cross-validation. In Section 3.2 we consider a dictionary consisting of the

**Algorithm 1**: Computation of the star-shaped aggregate.

**Input**: dictionary $F$, data $(X_i,Y_i)_{i=1}^{2n}$, and a confidence level $x>0$
**Output**: star-shaped aggregate $\tilde{f}$
Split $D_{2n}$ into two samples $D_{n,1}$ and $D_{n,2}$
**foreach** $j\in\{1,\dots,M\}$ **do**
Compute $R_{n,1}(f_j)$ and $R_{n,2}(f_j)$, and use this loop to find $\widehat{f}_{n,1}\in \operatorname{argmin}_{f\in F} R_{n,1}(f)$
**end**
**foreach** $j\in\{1,\dots,M\}$ **do**
Compute $\|f_j-\widehat{f}_{n,1}\|_{n,1}$ and $\|f_j-\widehat{f}_{n,1}\|_{n,2}$
**end**
Construct the set of preselected elements

$$\widehat{F}_1=\left\{f\in F: R_{n,1}(f)\le R_{n,1}(\widehat{f}_{n,1})+c\max\left(\phi\|\widehat{f}_{n,1}-f\|_{n,1},\phi^2\right)\right\},$$

where $\phi$ is given in Definition 1.
**foreach** $f\in\widehat{F}_1$ **do**
compute

$$R_{n,2}(\lambda_{n,2}(\widehat{f}_{n,1},f)\widehat{f}_{n,1}+(1-\lambda_{n,2}(\widehat{f}_{n,1},f))f)$$

and keep the element $f_{\widehat{j}}\in\widehat{F}_1$ that minimizes this quantity
**end**
**return**

$$\tilde{f}=\lambda_{n,2}(\widehat{f}_{n,1},f_{\widehat{j}})\widehat{f}_{n,1}+(1-\lambda_{n,2}(\widehat{f}_{n,1},f_{\widehat{j}}))f_{\widehat{j}},$$

entire sequence of Lasso estimators and a dictionary consisting of entire sequences of the elastic-net estimators (see Zou and Hastie, 2005) corresponding to several ridge penalization parameters, so this dictionary contains the Lasso, the elastic-net, the ridge and the ordinary least-squares estimators.

**Remark 5** *Note that since an aggregation algorithm is "generic", in the sense that it can be applied to any dictionary, one could consider larger dictionaries, containing many instances of different type of estimators, for several choices of the tuning parameters, like the Adaptive Lasso (see Zou, 2006) among many other instances of the Lasso. We believe that the conclusion of the numerical study proposed here would be the same as for a much larger dictionary. Indeed, let us recall that here, the focus is on the comparison of selection and aggregation procedures for the choice of tuning parameters, and not on the comparison of the procedures inside the dictionary themselves.*

### 3.1 Examples of Models

We simulate $n$ independent copies of the linear regression model

$$Y=\beta^\top X+\varepsilon,$$

where $\beta\in\mathbb{R}^p$. Several settings are considered, see Models 1-6 below, including sparse and non-sparse vectors $\beta$ and several signal-to-noise ratios. Models 1-4 are from Tibshirani (1996).

**Model 1 (A few effects).** We set $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$, so $p = 8$, and we let $n$ to be 20 and 60. The vector $X = (X^1, \dots, X^d)$ is a centered normal vector with covariance matrix $\mathrm{Cov}(X^i, X^j) = \rho^{|i-j|}$, with $\rho = 1/2$. The noise $\varepsilon_i$ is $N(0, \sigma^2)$ with $\sigma$ equal to 1 or 3.

**Model 2 (Every effects).** This example is the same as Model 1, but with $\beta = (2, 2, 2, 2, 2, 2, 2, 2)$.

**Model 3 (A single effect).** This example is the same as Model 1, but with $\beta = (5, 0, 0, 0, 0, 0, 0, 0)$.

**Model 4 (A larger model).** We set $\beta = (0^{10}, 2^{10}, 0^{10}, 2^{10})$, where $x^y$ stands for the vector of dimension $y$ with each coordinate equal to $x$, so $p = 40$. We let $n$ to be 100 and 200. We consider covariates $X_i^j = Z_{i,j} + Z_i$ where $Z_{i,j}$ and $Z_i$ are independent $N(0,1)$ variables. This induces pairwise correlation equal to 0.5 among the covariates. The noise $\varepsilon_i$ is $N(0, \sigma^2)$ with $\sigma$ equal to 15 or 7.

**Model 5 (Sparse vector in high dimension).** We set $\beta = (2.5^5, 1.5^5, 0.5^5, 0^{185})$, so $p = 200$. We let $n$ to be 50 and 100. The first 15 covariates $(X^1, \dots, X^{15})$ and the remaining 185 covariates $(X^{16}, \dots, X^{200})$ are independent. Each of these are Gaussian vectors with the same covariance matrix as in Model 1 with $\rho = 0.5$. The noise is $N(0, \sigma^2)$ with $\sigma$ equal to 3 and 1.5.

**Model 6 (Sparse vector in high dimension, stronger correlation).** This example is the same as Model 5, but with $\rho = 0.95$.

### 3.2 Procedures

We consider a dictionary consisting of the entire sequence of Lasso estimators and a dictionary with several sequences of elastic-net estimators, corresponding to ridge parameters in the set of values $\{0, 0.01, 0.1, 1, 5, 10, 20, 50, 100\}$ (these dictionaries are computed with the `lars` and `enet` routines from `R`).[1] For each dictionary, we compute the prediction errors $|X(\widehat{\beta} - \beta)|_2$ (where $X$ is the matrix with rows $X_1^\top, \dots, X_n^\top$ and $|\cdot|_2$ is the $\ell_2^n$-norm of 200 replications (this makes the results stable enough), where $\widehat{\beta}$ is one of the following:

- $\widehat{\beta}^{(\mathrm{Oracle})}$ = the element of the dictionary with smallest prediction error
- $\widehat{\beta}^{(C_p)}$ = the Lasso estimator selected by Mallows-$C_p$ heuristic
- $\widehat{\beta}^{(10-\mathrm{Fold})}$ = the element of the dictionary selected by 10-fold cross-validation
- $\widehat{\beta}^{(\mathrm{Loo})}$ = the element of the dictionary selected by leave-one-out cross-validation
- $\widehat{\beta}^{(\mathrm{AEW})}$ = The aggregate with exponential weights applied to the dictionary, with temperature parameter equal to $4\sigma^2$, see for instance Dalalyan and Tsybakov (2007)
- $\widehat{\beta}^{(\mathrm{Star})}$ = the star-shaped aggregate applied to the dictionary.

For the AEW and the star-shaped aggregate, the splits are chosen at random with size $[n/2]$ for training and $n - [n/2]$ for learning. For both aggregates we use jackknife: we compute the mean of 100 aggregates obtained with several splits chosen at random. This makes the final aggregates less

1. R can be found at `www.r-project.org`.

dependent on the split. As a matter of fact, we observed in our numerical studies that Star-shaped aggregation with the preselection step and without it (see Definition 1) provides close estimators. So, in order to improve the computational burden, the numerical results of the Star-shaped aggregate reproduced here are the ones obtained without the preselection step.

We need to explain how variable selection is performed based on $J$ star-shaped aggregates coming from $J$ random splits (here we take $J = 100$). A Star-shaped aggregate $\widehat{f}^{(j)}$, corresponding to a split $j$, can be written as

$$\widehat{f}^{(j)} = \widehat{\lambda}^{(j)} \widehat{f}^{(j)}_{\mathrm{ERM}} + (1 - \widehat{\lambda}^{(j)}) \widehat{f}^{(j)}_{\mathrm{other}},$$

where $\widehat{f}^{(j)}_{\mathrm{ERM}}$ is the ERM in $F$ corresponding to the split $j$ and $\widehat{f}^{(j)}_{\mathrm{other}}$ is the other vertex of the segment where the empirical risk is minimized (recall that the aggregate minimizes the empirical risk over the set of segments $\mathrm{star}(\widehat{f}^{(j)}_{\mathrm{ERM}}, F)$). For each split $j$, we estimate the significance of each covariate using

$$\widehat{\pi}^{(j)} = \widehat{\lambda}^{(j)} \mathbf{1}_{\widehat{\beta}^{(j)}_{\mathrm{ERM}} \neq 0} + (1 - \widehat{\lambda}^{(j)}) \mathbf{1}_{\widehat{\beta}^{(j)}_{\mathrm{other}} \neq 0},$$

where $\mathbf{1}_{v \neq 0} = (\mathbf{1}_{v_1 \neq 0}, \ldots, \mathbf{1}_{v_d \neq 0})$. The vector $\widehat{\pi}^{(j)}$ does a simple average of the contributions of the supports of $\widehat{\beta}^{(j)}_{\mathrm{ERM}}$ and $\widehat{\beta}^{(j)}_{\mathrm{other}}$, weighted by $\widehat{\lambda}^{(j)}$. To take into consideration each split, we simply compute the mean of the significances of each split:

$$\widehat{\pi} = \frac{1}{J} \sum_{j=1}^{J} \widehat{\pi}^{(j)}.$$

The vector $\widehat{\pi}$ contains the final significances of each covariate. This procedure is close in spirit to the stability selection procedure described in Meinshausen and Bühlmann (2010), since each aggregate is related to a subsample. Finally, the selected covariates are the one in

$$\widehat{S} = \Big\{ k \in \{1, \ldots, p\} : \widehat{\pi}_k \geq \widehat{t} \Big\},$$

where $\widehat{t}$ is a random threshold given by

$$\widehat{t} = \frac{1}{2} \Big( 1 + \frac{\widehat{q}^2}{p^2 \beta} \Big),$$

where $\widehat{q} = \min(\widehat{s}, \sqrt{0.7p})$, $\beta = p/10$ and $\widehat{s} = \frac{1}{J} \sum_{j=1}^{J} \sum_{k=1}^{p} \widehat{\pi}^{(j)}_k$ is the average sparsity (number of non-zero coefficients) for each splits. This choice of threshold follows the arguments from Meinshausen and Bühlmann (2010), together with some empirical tuning.

For each of the Models 1-6, the boxplots of the 200 prediction errors are given in Figures 3 and 4. Note that in a high dimensional setting ($p > n$), we don't reproduce the $C_p$'s prediction errors, since in this case the `lars` package does not give it correctly. For the elastic-net dictionary, the boxplot of the predictions errors are given for Models 1-4 in Figure 5. The results concerning variables selection for the Lars and the Elastic-Net dictionaries are given in Tables 1 and 2. In these tables we reproduce the number of selected variables by each procedure, and the number of noise variables (the selected variables which are not active in the true model).

### 3.3 Conclusion

In most cases, the Star-Shaped aggregate improves upon the AEW and the considered selection procedures both in terms of prediction error and variable selection. The proposed variable selection algorithm based on star-shaped aggregation and stability selection tends to select smaller models than the $C_p$ and cross-validation methods (see Table 1, Models 1-4) leading to less noise variables. In particular, in high-dimensional cases ($p > n$), it is much more stable regarding the sample size and noise level, and provides better results most of the time (see Table 1, Models 5-6). In terms of prediction error, the Star-Shaped always improve the AEW, and is better than the $C_p$ and cross-validations in most cases. We can say that, roughly, the $C_p$ and the cross-validations are better than the Star-Shaped aggregate only for non-sparse vectors (since these selection procedures tend to select larger models), in particular when $n$ is small and $\sigma$ is large. We can conclude by saying that, in the worst cases, the Star-shaped algorithm has prediction and selection performances which are comparable to cross-validations and $C_p$ heuristic, but, on the other hand, it can improve them a lot (in particular for sparse vectors). One can think of the Star-Shaped aggregation algorithm as an alternative to cross-validation and $C_p$.

## Acknowledgments

This work is supported by French Agence Nationale de la Recherce (ANR) ANR Grant "PROGNOSTIC" ANR-09-JCJC-0101-01.

## Appendix A. Proofs

We will use the following notations. If $f^F \in \operatorname{argmin}_{f\in F} R(f)$, we will consider the excess loss

$$\mathcal{L}_f = \mathcal{L}_F(f)(X,Y) := (Y - f(X))^2 - (Y - f^F(X))^2,$$

and use the notations

$$P\mathcal{L}_f := \mathbb{E}\mathcal{L}_f(X,Y), \quad P_n\mathcal{L}_f := \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_f(X_i,Y_i).$$

### A.1 Proof of Theorem 2

Let us prove the result in the $\psi_1$ case, the other case is similar. Fix $x > 0$ and let $\widehat{\mathcal{F}}$ be either (5), (6) or (7). Set $d := \operatorname{diam}(\widehat{F}_1, L_2(\mu))$. Consider the second half of the sample $D_{n,2} = (X_i,Y_i)_{i=n+1}^{2n}$. By Corollary 8 (see Appendix A.2 below), with probability at least $1 - 4\exp(-x)$ (relative to $D_{n,2}$), we have for every $f \in \widehat{\mathcal{F}}$

$$\Big|\frac{1}{n}\sum_{i=1+n}^{2n} \mathcal{L}_{\widehat{\mathcal{F}}}(f)(X_i,Y_i) - \mathbb{E}\big(\mathcal{L}_{\widehat{\mathcal{F}}}(f)(X,Y)|D_{n,1}\big)\Big| \leq c(\sigma_\varepsilon + b)\max(d\phi, b\phi^2),$$

where $\mathcal{L}_{\widehat{\mathcal{F}}}(f)(X,Y) := (f(X)-Y)^2 - (f^{\widehat{\mathcal{F}}}(X)-Y)^2$ is the excess loss function relative to $\widehat{\mathcal{F}}$, $f^{\widehat{\mathcal{F}}} \in \operatorname{argmin}_{f\in\widehat{\mathcal{F}}} R(f)$ and where $\phi = \sqrt{((\log M + x)\log n)/n}$. By definition of $\tilde{f}$, we have

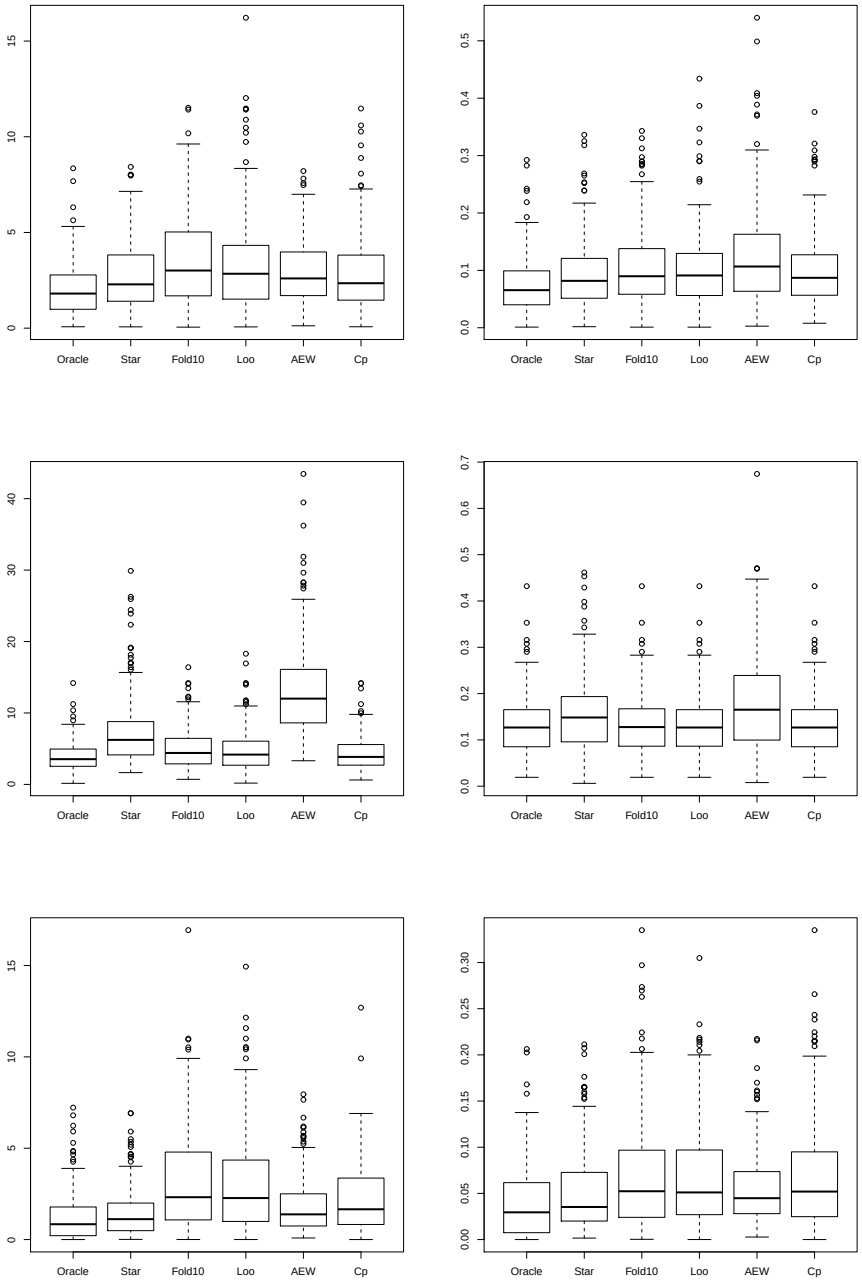


Figure 3: First line: prediction errors for Model 1, with $n = 20$, $\sigma = 3$ (left) and $n = 60$, $\sigma = 1$ (right) ; Second line : prediction errors for Model 2, with $n = 20$, $\sigma = 3$ (left) and $n = 60$, $\sigma = 1$ (right) ; thrid line: prediction errors for Model 3, with $n = 20$, $\sigma = 3$ (left) and $n = 60$, $\sigma = 1$ (right)

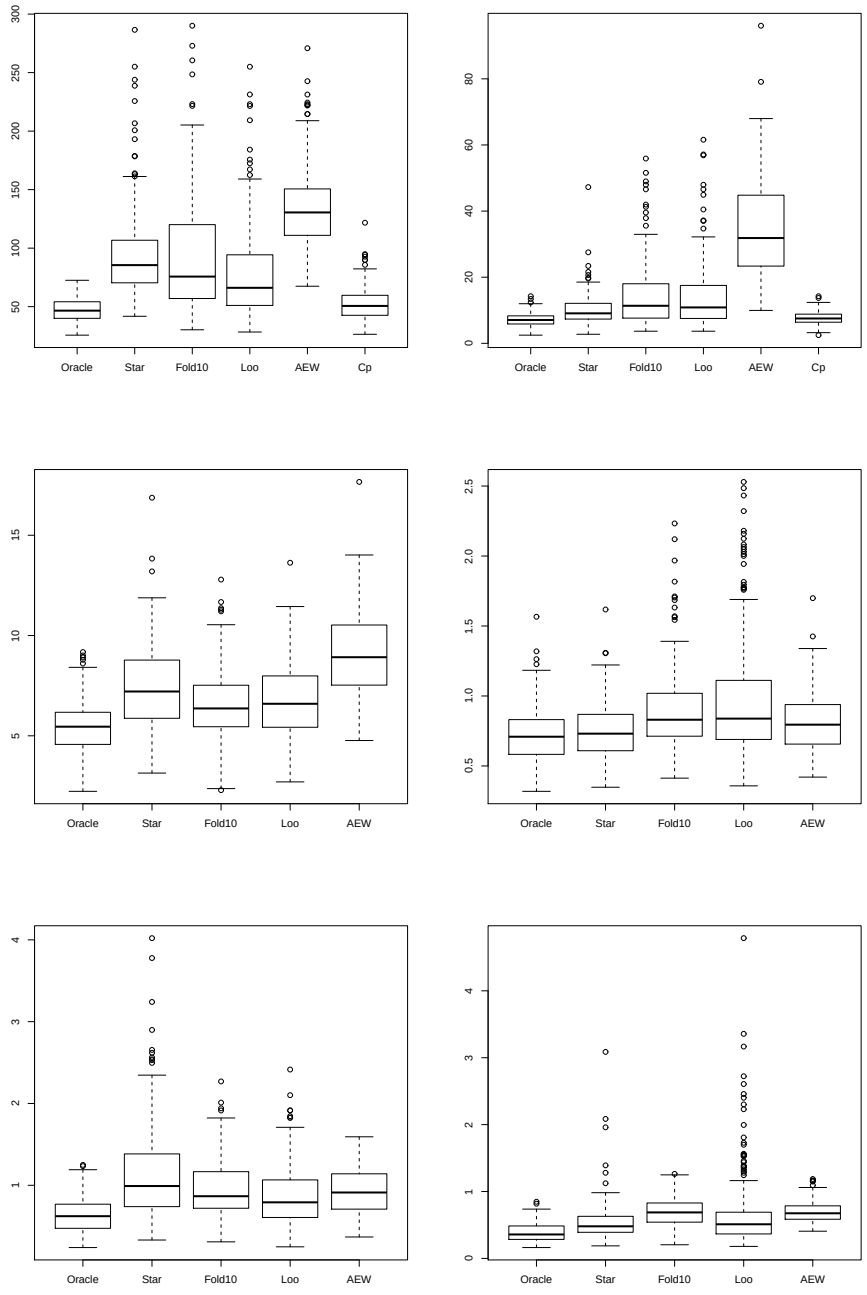


Figure 4: First line: prediction errors for Model 4, with $n = 100$, $\sigma = 15$ (left) and $n = 200$, $\sigma = 7$ (right) ; Second line: Prediction errors for Model 5, with $n = 50$, $\sigma = 3$ (left) and $n = 100$, $\sigma = 1.5$ (right) ; Third line: Prediction errors for Model 6, with $n = 50$, $\sigma = 1.5$ (left) and $n = 100$, $\sigma = 1.5$ (right)

| | **Model 1** | | | | **Model 2** | | | |
|---|---|---|---|---|---|---|---|---|
| | $n=20, \sigma=3$ | | $n=60, \sigma=1$ | | $n=20, \sigma=3$ | | $n=60, \sigma=1$ | |
| | Selected | Noise | Selected | Noise | Selected | Noise | Selected | Noise |
| Truth | 3 | 0 | 3 | 0 | 8 | 0 | 8 | 0 |
| 10-fold | 3.870 | 1.410 | 5.260 | 2.260 | 7.190 | 0 | 8 | 0 |
| Loo | 3.965 | 1.465 | 5.055 | 2.055 | 7.235 | 0 | 8 | 0 |
| Cp | 4.165 | 1.645 | 4.710 | 1.710 | 7.085 | 0 | 8 | 0 |
| Star | 2.860 | 0.675 | 4.355 | 1.355 | 6.250 | 0 | 8 | 0 |

| | **Model 3** | | | | **Model 4** | | | |
|---|---|---|---|---|---|---|---|---|
| | $n=20, \sigma=3$ | | $n=60, \sigma=1$ | | $n=100, \sigma=15$ | | $n=200, \sigma=7$ | |
| | Selected | Noise | Selected | Noise | Selected | Noise | Selected | Noise |
| Truth | 1 | 0 | 1 | 0 | 20 | 0 | 20 | 0 |
| 10-fold | 2.365 | 1.365 | 2.980 | 1.980 | 21.610 | 6.955 | 28.405 | 8.415 |
| Loo | 2.440 | 1.440 | 2.645 | 1.645 | 22.295 | 7.305 | 28.480 | 8.495 |
| Cp | 2.965 | 1.965 | 2.650 | 1.650 | 23.860 | 8.175 | 29.715 | 9.720 |
| Star | 1.655 | 0.655 | 1.855 | 0.855 | 18.065 | 4.910 | 27.850 | 7.855 |

| | **Model 5** | | | | **Model 6** | | | |
|---|---|---|---|---|---|---|---|---|
| | $n=100, \sigma=1.5$ | | $n=200, \sigma=0.5$ | | $n=100, \sigma=1.5$ | | $n=200, \sigma=0.5$ | |
| | Selected | Noise | Selected | Noise | Selected | Noise | Selected | Noise |
| Truth | 15 | 0 | 15 | 0 | 15 | 0 | 15 | 0 |
| 10-fold | 47.375 | 32.550 | 14.035 | 0 | 39.150 | 25.830 | 7.560 | 0 |
| Loo | 44.030 | 29.215 | 10.455 | 0 | 24.370 | 10.990 | 2.425 | 0 |
| Star | 15.690 | 1.245 | 17.780 | 2.780 | 13.175 | 0.055 | 15.145 | 0.150 |

Table 1: Accuracy of variable prediction in Models 1 to 6 (Lars dictionary)

$\frac{1}{n}\sum_{i=n+1}^{2n} \mathcal{L}_{\widehat{\mathcal{F}}}(\tilde{f})(X_i,Y_i) \leq 0$, so, on this event (relative to $D_{n,2}$)

$$\begin{aligned} R(\tilde{f}) &\leq R(f^{\widehat{\mathcal{F}}}) + \mathbb{E}\big(\mathcal{L}_{\widehat{\mathcal{F}}}(\tilde{f})|D_{n,1}\big) - \frac{1}{n}\sum_{i=n+1}^{2n} \mathcal{L}_{\widehat{\mathcal{F}}}(\tilde{f})(X_i,Y_i) \\ &\leq R(f^{\widehat{\mathcal{F}}}) + c(\sigma_\varepsilon + b)\max(d\phi, b\phi^2) \\ &= R(f^F) + \Big(c(\sigma_\varepsilon + b)\max(d\phi, b\phi^2) - \big(R(f^F) - R(f^{\widehat{\mathcal{F}}})\big)\Big) \\ &=: R(f^F) + \beta, \end{aligned}$$

and it remains to show that

$$\beta \leq c_{b,\sigma_\varepsilon} \frac{(1+x)\log M \log n}{n}.$$

When $\widehat{\mathcal{F}}$ is given by (5) or (6), the geometrical configuration is the same as in Lecué and Mendelson (2009), so we skip the proof.

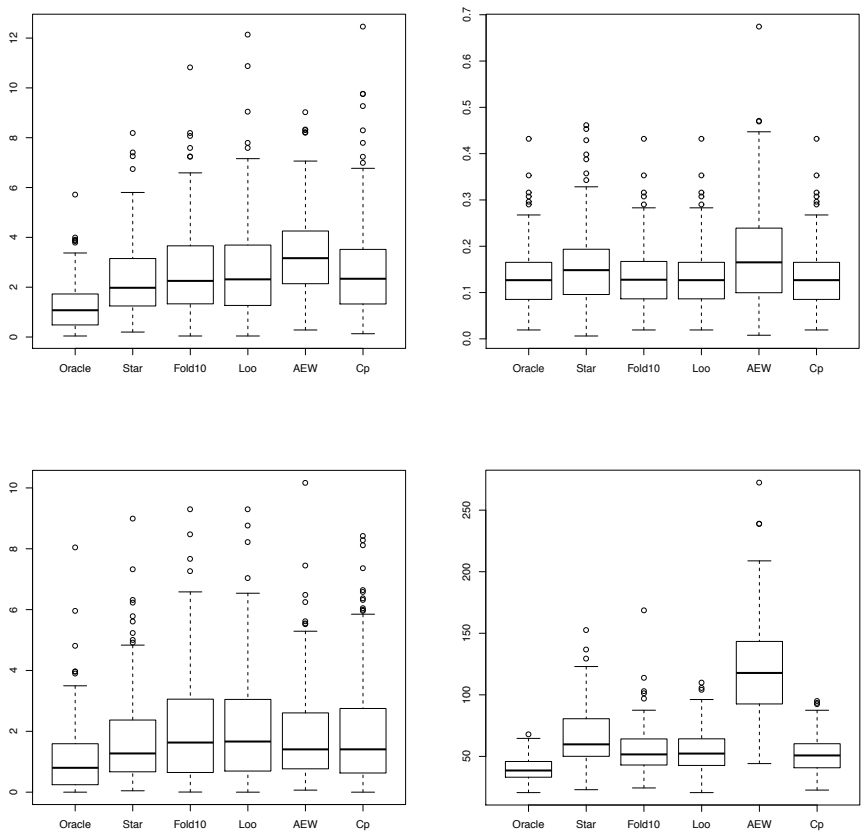


Figure 5: Prediction errors for Models 1 to 4 using the elastic-net dictionary (upper left: Model 1 with $\sigma = 3, n = 20$, upper right: Model 2 with $\sigma = 3, n = 20$, bottom left: Model 3 with $\sigma = 3, n = 20$ and bottom right: Model 4 with $n = 100$, $\sigma = 15$).

| | **Model 1** $n = 20$, $\sigma = 3$ | | **Model 2** $n = 20$, $\sigma = 3$ | | **Model 3** $n = 20$, $\sigma = 3$ | | **Model 4** $n = 100$, $\sigma = 15$ | |
|---|---|---|---|---|---|---|---|---|
| | Selected | Noise | Selected | Noise | Selected | Noise | Selected | Noise |
| Truth | 3 | 0 | 8 | 0 | 1 | 0 | 20 | 0 |
| 10-fold | 5.040 | 2.155 | 7.450 | 0 | 3.045 | 2.045 | 25.575 | 9.475 |
| Loo | 4.940 | 2.065 | 7.460 | 0 | 2.980 | 1.980 | 25.535 | 9.660 |
| Cp | 4.490 | 1.660 | 7.335 | 0 | 2.760 | 1.760 | 24.345 | 8.470 |
| Star | 4.355 | 1.475 | 7.485 | 0 | 2.080 | 1.080 | 24.090 | 8.755 |

Table 2: Accuracy of variable prediction in Models 1 to 4 (Elastic-Net dictionary)

Let us turn out to the situation where $\widehat{\mathcal{F}}$ is given by (7). Recall that $\widehat{f}_{n,1}$ is the ERM on $\widehat{F}_1$ using $D_{n,1}$. Consider $f_1$ such that $\|\widehat{f}_{n,1}-f_1\|_{L^2(\mu)}=\max_{f\in\widehat{F}_1}\|\widehat{f}_{n,1}-f\|_{L^2(\mu)}$, and note that

$$\|\widehat{f}_{n,1}-f_1\|_{L^2(\mu)}\le d\le 2\|\widehat{f}_{n,1}-f_1\|_{L^2(\mu)}.$$

The mid-point $f_2:=(\widehat{f}_{n,1}+f_1)/2$ belongs to $\mathrm{star}(\widehat{f}_{n,1},\widehat{F}_1)$. Using the parallelogram identity, we have for any $u,v\in L_2(\nu)$:

$$\mathbb{E}_\nu\Big(\frac{u+v}{2}\Big)^2\le\frac{\mathbb{E}_\nu(u^2)+\mathbb{E}_\nu(v^2)}{2}-\frac{\|u-v\|^2_{L_2(\nu)}}{4},$$

where for every $h\in L_2(\nu)$, $\mathbb{E}_\nu(h)=\mathbb{E}h(X,Y)$. In particular, for $u(X,Y)=\widehat{f}_{n,1}-Y$ and $v(X,Y)=f_1(X)-Y$, the mid-point is $(u(X,Y)+v(X,Y))/2=f_2(X)-Y$. Hence,

$$\begin{aligned}R(f_2)&=\mathbb{E}(f_2(X)-Y)^2=\mathbb{E}\Big(\frac{\widehat{f}_{n,1}(X)+f_1(X)}{2}-Y\Big)^2\\&\le\frac{1}{2}\mathbb{E}(\widehat{f}_{n,1}(X)-Y)^2+\frac{1}{2}\mathbb{E}(f_1(X)-Y)^2-\frac{1}{4}\|f_{n,1}-f_1\|^2_{L_2(\mu)}\\&\le\frac{1}{2}R(\widehat{f}_{n,1})+\frac{1}{2}R(f_1)-\frac{d^2}{16},\end{aligned}$$

where the expectations are taken conditioned on $D_{n,1}$. By Lemma 10 (see Appendix A.2 below), since $\widehat{f}_{n,1},f_1\in\widehat{F}_1$, we have

$$\frac{1}{2}R(\widehat{f}_{n,1})+\frac{1}{2}R(f_1)\le R(f^F)+c(\sigma_\varepsilon+b)\max(\phi d,b\phi^2),$$

and thus, since $f_2\in\widehat{\mathcal{F}}$

$$R(f^{\widehat{\mathcal{F}}})\le R(f_2)\le R(f^F)+c(\sigma_\varepsilon+b)\max(\phi d,b\phi^2)-cd^2.$$

Therefore,

$$\begin{aligned}\beta&=c(\sigma_\varepsilon+b)\max(d\phi,b\phi^2)-\big(R(f^F)-R(f^{\widehat{\mathcal{F}}})\big)\\&\le c(\sigma_\varepsilon+b)\max(\phi d,b\phi^2)-cd^2.\end{aligned}$$

Finally, if $d\ge c_{\sigma_\varepsilon,b}\phi$ then $\beta\le 0$, otherwise $\beta\le c_{\sigma_\varepsilon,b}\phi^2$. It concludes the proof of Theorem 2. □

### A.2 Tools from Empirical Process Theory and Technical Results

The following Theorem is a Talagrand's type concentration inequality (see Talagrand, 1996) for a class of unbounded functions.

**Theorem 6 (Theorem 4, Adamczak, 2008)** *Assume that $X,X_1,\dots,X_n$ are independent random variables and $F$ is a countable set of functions such that $\mathbb{E}f(X)=0,\forall f\in F$ and $\|\sup_{f\in F}f(X)\|_{\psi_1}<+\infty$. Define*

$$Z:=\sup_{f\in F}\Big|\frac{1}{n}\sum_{i=1}^n f(X_i)\Big|$$

*and*

$$\sigma^2 = \sup_{f\in F}\mathbb{E}f(X)^2 \text{ and } b := \|\max_{i=1,\dots,n}\sup_{f\in F}|f(X_i)|\|_{\psi_1}.$$

*Then, for any* $\eta\in(0,1)$ *and* $\delta>0$*, there is* $c=c_{\eta,\delta}$ *such that for any* $x>0$*:*

$$\mathbb{P}\Big[Z\geq(1+\eta)\mathbb{E}Z+\sigma\sqrt{2(1+\delta)\frac{x}{n}}+cb\Big(\frac{x}{n}\Big)\Big]\leq 4e^{-x}$$
$$\mathbb{P}\Big[Z\leq(1-\eta)\mathbb{E}Z-\sigma\sqrt{2(1+\delta)\frac{x}{n}}-cb\Big(\frac{x}{n}\Big)\Big]\leq 4e^{-x}.$$

Now we state some technical Lemmas, used in the proof of Theorem 2. Given a sample $(Z_i)_{i=1}^n$, we set the random empirical measure $P_n := n^{-1}\sum_{i=1}^n\delta_{Z_i}$. For any function $f$ define $(P-P_n)(f) := n^{-1}\sum_{i=1}^n f(Z_i)-\mathbb{E}f(Z)$ and for a class of functions $F$, define $\|P-P_n\|_F := \sup_{f\in F}|(P-P_n)(f)|$. In all what follows, we denote by $c$ an absolute positive constant, that can vary from place to place. Its dependence on the parameters of the setting is specified in place.

**Lemma 7** *Define*

$$d(F) := \mathrm{diam}(F,L^2(\mu)),\quad \sigma^2(F)=\sup_{f\in F}\mathbb{E}[f(X)^2],\quad \mathcal{C}=\mathrm{conv}(F),$$

*and* $\mathcal{L}_{\mathcal{C}}(\mathcal{C})=\{(Y-f(X))^2-(Y-f^{\mathcal{C}}(X))^2 : f\in\mathcal{C}\}$*, where* $f^{\mathcal{C}}\in\mathrm{argmin}_{g\in\mathcal{C}}R(g)$*. If* (2) *holds, we have*

$$\mathbb{E}\Big[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^n f^2(X_i)\Big]\leq c\max\Big(\sigma^2(F),\frac{b^2\log M}{n}\Big),\ \textit{and}$$
$$\mathbb{E}\|P_n-P\|_{\mathcal{L}_{\mathcal{C}}(\mathcal{C})}\leq cb\sqrt{\frac{\log M}{n}}\max\Big(b\sqrt{\frac{\log M}{n}},d(F)\Big).$$

*If* (3) *holds, we have*

$$\mathbb{E}\Big[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^n f^2(X_i)\Big]\leq c\max\Big(\sigma^2(F),\frac{b^2\log M}{n}\Big),\ \textit{and}$$
$$\mathbb{E}\|P_n-P\|_{\mathcal{L}_{\mathcal{C}}(\mathcal{C})}\leq cb\sqrt{\frac{\log M\log n}{n}}\max\Big(b\sqrt{\frac{\log M\log n}{n}},d(F)\Big).$$

**Proof** First, consider the case (3). Define

$$r^2=\sup_{f\in F}\frac{1}{n}\sum_{i=1}^n f(X_i)^2,$$

and note that $\mathbb{E}_X(r^2)\leq\mathbb{E}_X\|P-P_n\|_{F^2}+\sigma(F)^2$, where $F^2 := \{f^2 : f\in F\}$. Using the Giné-Zinn symmetrization argument, see Giné and Zinn (1984), we have

$$\mathbb{E}_X\|P-P_n\|_{F^2}\leq\frac{c}{n}\mathbb{E}_X\mathbb{E}_g\Big[\sup_{f\in F}\Big|\sum_{i=1}^n g_i f^2(X_i)\Big|\Big],$$

where $(g_i)$ are i.i.d. standard normal. The process $f \mapsto Z_{2,f} = \sum_{i=1}^n g_i f^2(X_i)$ is Gaussian, with intrinsic distance

$$\mathbb{E}_g|Z_{2,f} - Z_{2,f'}|^2 = \sum_{i=1}^n (f(X_i)^2 - f'(X_i)^2)^2 \leq d_{n,\infty}(f,f')^2 \times 4nr^2,$$

where $d_{n,\infty}(f,f') = \max_{i=1,\dots,n} |f(X_i) - f'(X_i)|$. Using (3) we have $d_{n,\infty}(f,f') \leq 2b$ for any $f, f' \in F$, so using Dudley's entropy integral, we have

$$\mathbb{E}_g\|P - P_n\|_{F^2} \leq \frac{c}{\sqrt{n}} \int_0^{2b} \sqrt{\log N(F, d_{n,\infty}, t)}dt \leq cr\sqrt{\frac{\log M}{n}}.$$

So, we get

$$\mathbb{E}_X\|P - P_n\|_{F^2} \leq cb\sqrt{\frac{\log M}{n}}\mathbb{E}_X[r] \leq cb\sqrt{\frac{\log M}{n}}\sqrt{\mathbb{E}_X[r^2]},$$

which entails that

$$\mathbb{E}_X(r^2) \leq c\max\Big(\frac{b^2\log M}{n} + \sigma(F)^2\Big).$$

Let us turn to the part of the Lemma concerning $\mathbb{E}\|P - P_n\|_{\mathcal{L}_\mathcal{C}(\mathcal{C})}$. Recall that $\mathcal{C} = \mathrm{conv}(F)$ and write for short $\mathcal{L}_f(X,Y) = \mathcal{L}_\mathcal{C}(f)(X,Y) = (Y - f(X))^2 - (Y - f^\mathcal{C}(X))^2$ for each $f \in \mathcal{C}$, where we recall that $f^\mathcal{C} \in \mathrm{argmin}_{g\in\mathcal{C}} R(g)$. Using the same argument as before we have

$$\mathbb{E}\|P - P_n\|_{\mathcal{L}_\mathcal{C}(\mathcal{C})} \leq \frac{c}{n}\mathbb{E}_{(X,Y)}\mathbb{E}_g\Big[\sup_{f\in\mathcal{C}}\Big|\sum_{i=1}^n g_i\mathcal{L}_f(X_i,Y_i)\Big|\Big].$$

Consider the Gaussian process $f \in \mathcal{C} \to Z_f := \sum_{i=1}^n g_i\mathcal{L}_f(X_i,Y_i)$ indexed by $\mathcal{C}$. For every $f, f' \in \mathcal{C}$, the intrinsic distance of $(Z_f)_{f\in\mathcal{C}}$ satisfies

$$\begin{aligned}
\mathbb{E}_g|Z_f - Z_{f'}|^2 &= \sum_{i=1}^n (\mathcal{L}_f(X_i,Y_i) - \mathcal{L}_{f'}(X_i,Y_i))^2 \\
&\leq \max_{i=1,\dots,n} |2Y_i - f(X_i) - f'(X_i)|^2 \times \sum_{i=1}^n (f(X_i) - f'(X_i))^2 \\
&= \max_{i=1,\dots,n} |2Y_i - f(X_i) - f'(X_i)|^2 \times \mathbb{E}_g|Z'_f - Z'_{f'}|^2,
\end{aligned}$$

where $Z'_f := \sum_{i=1}^n g_i(f(X_i) - f^\mathcal{C}(X_i))$. Therefore, by Slepian's Lemma, we have for every $(X_i,Y_i)_{i=1}^n$:

$$\mathbb{E}_g\Big[\sup_{f\in\mathcal{C}} Z_f\Big] \leq \max_{i=1,\dots,n} \sup_{f,f'\in\mathcal{C}} |2Y_i - f(X_i) - f'(X_i)| \times \mathbb{E}_g\Big[\sup_{f\in\mathcal{C}} Z'_f\Big],$$

and since for every $f = \sum_{j=1}^M \alpha_j f_j \in \mathcal{C}$, where $\alpha_j \geq 0, \forall j = 1,\dots,M$ and $\sum\alpha_j = 1$, $Z'_f = \sum_{j=1}^M \alpha_j Z_{f_j}$, we have

$$\mathbb{E}_g\Big[\sup_{f\in\mathcal{C}} Z'_f\Big] \leq \mathbb{E}_g\Big[\sup_{f\in F} Z'_f\Big].$$

Moreover, we have, using Dudley's entropy integral argument,

$$\frac{1}{n}\mathbb{E}_g\Big[\sup_{f\in F} Z'_f\Big] \leq \frac{c}{\sqrt{n}}\int_0^{\Delta_n(F')} \sqrt{N(F,\|\cdot\|_n,t)}dt \leq c\sqrt{\frac{\log M}{n}}r',$$

where $F' := \{f - f^{\mathcal{C}} : f \in F\}$ and $\Delta_n(F') := \mathrm{diam}(F', \|\cdot\|_n)$ and

$$r'^2 := \sup_{f \in F'} \frac{1}{n} \sum_{i=1}^n f(X_i)^2.$$

Hence, we proved that

$$\mathbb{E}\|P - P_n\|_{\mathcal{L}_{\mathcal{C}}(C)} \leq c\sqrt{\frac{\log M}{n}} \sqrt{\mathbb{E}\Big[\max_{i=1,\dots,n} |2Y_i - f(X_i) - f'(X_i)|^2\Big]} \sqrt{\mathbb{E}(r'^2)}.$$

Using Pisier's inequality for $\psi_1$ random variables and the fact that $\mathbb{E}(U^2) \leq 4\|U\|_{\psi_1}$ for any $\psi_1$-random variable $U$, together with (3), we obtain that

$$\mathbb{E}\Big[\max_{i=1,\dots,n} \sup_{f,f' \in \mathcal{C}} |2Y_i - f(X_i) - f'(X_i)|^2\Big] \leq cb^2 \log(n). \tag{8}$$

So, we finally obtain

$$\mathbb{E}\|P - P_n\|_{\mathcal{L}_{\mathcal{C}}(C)} \leq c\sqrt{\frac{\log n \log M}{n}} \sqrt{\mathbb{E}(r'^2)},$$

and the conclusion follows from the first part of the Lemma, since $\sigma(F') \leq d(F)$. The case (2) is easier and follows from the fact that the left hand side of (8) is smaller than $4b$. ■

Lemma 7 combined with Theorem 6 leads to the following corollary.

**Corollary 8** *Let $d(F) = \mathrm{diam}(F, L^2(\mu))$, $\mathcal{C} := \mathrm{conv}(F)$ and $\mathcal{L}_f(X,Y) = (Y - f(X))^2 - (Y - f^{\mathcal{C}}(X))^2$ for any $f \in \mathcal{C}$.*

*If* (3) *holds, we have, with probability larger than $1 - 4e^{-x}$, that for every $f \in \mathcal{C}$:*

$$\Big|\frac{1}{n}\sum_{i=1}^n \mathcal{L}_f(X_i,Y_i) - \mathbb{E}\mathcal{L}_f(X,Y)\Big| \leq c(\sigma_\varepsilon + b)\sqrt{\frac{(\log M + x)\log n}{n}} \max\Big(b\sqrt{\frac{(\log M + x)\log n}{n}}, d(F)\Big).$$

*If* (2) *holds, we have, with probability larger than $1 - 2e^{-x}$, that for every $f \in \mathcal{C}$:*

$$\Big|\frac{1}{n}\sum_{i=1}^n \mathcal{L}_f(X_i,Y_i) - \mathbb{E}\mathcal{L}_f(X,Y)\Big| \leq cb\sqrt{\frac{\log M + x}{n}} \max\Big(b\sqrt{\frac{\log M + x}{n}}, d(F)\Big).$$

**Proof** Applying Theorem 6 to

$$Z := \sup_{f \in \mathcal{C}} \Big|\frac{1}{n}\sum_{i=1}^n \mathcal{L}_f(X_i,Y_i) - \mathbb{E}\mathcal{L}_f(X,Y)\Big|,$$

we obtain that, with a probability larger than $1 - 4e^{-x}$:

$$Z \leq c\Big(\mathbb{E}Z + \sigma(\mathcal{C})\sqrt{\frac{x}{n}} + b_n(\mathcal{C})\frac{x}{n}\Big),$$

where

$$\sigma(\mathcal{C})^2 = \sup_{f\in\mathcal{C}} \mathbb{E}[\mathcal{L}_f(X,Y)^2], \quad \text{and}$$
$$b_n(\mathcal{C}) = \Big\| \max_{i=1,\dots,n} \sup_{f\in\mathcal{C}} |\mathcal{L}_f(X_i,Y_i) - \mathbb{E}[\mathcal{L}_f(X,Y)]| \Big\|_{\psi_1}.$$

Since

$$\mathcal{L}_f(X,Y) = 2\varepsilon(f^{\mathcal{C}}(X) - f(X)) + (f^{\mathcal{C}}(X) - f(X))(2f_0(X) - f(X) - f^{\mathcal{C}}(X)), \tag{9}$$

we have using Assumptions 1 and (3):

$$\mathbb{E}[\mathcal{L}_f(X,Y)^2] \le (4\sigma_\varepsilon^2 + 2b^2)\|f - f^{\mathcal{C}}\|^2_{L^2(\mu)},$$

meaning that

$$\sigma(\mathcal{C})^2 \le (4\sigma_\varepsilon^2 + 2b^2)d(F).$$

Since $\mathbb{E}(|Z|) \le \|Z\|_{\psi_1}$, we have $b_n(\mathcal{C}) \le 2\log(n+1)\|\sup_{f\in\mathcal{C}} |\mathcal{L}_f(X,Y)|\|_{\psi_1}$. Moreover, using again (9), we obtain that

$$b_n(\mathcal{C}) \le 16\log(n+1)b^2.$$

Putting all this together, and using Lemma 7, we arrive at

$$Z \le c(\sigma_\varepsilon + b)\sqrt{\frac{(\log M + x)\log n}{n}} \max\Big(b\sqrt{\frac{(\log M + x)\log n}{n}}, d(F)\Big),$$

with probability larger than $1 - 4e^{-x}$ for any $x > 0$. In the bounded case (2) the proof is easier, and one can use the original Talagrand's concentration inequality. ■

**Lemma 9** *Let $\mathcal{L}_f(X,Y) = (Y - f(X))^2 - (Y - f^F(X))^2$ for any $f \in F$.*

*If* (3) *holds, we have with probability larger than $1 - 4e^{-x}$, that for every $f \in F$:*

$$\Big|\frac{1}{n}\sum_{i=1}^n \mathcal{L}_f(X_i,Y_i) - \mathbb{E}\mathcal{L}_f(X,Y)\Big| \le c(\sigma_\varepsilon + b)\sqrt{\frac{(\log M + x)\log n}{n}} \max\Big(b\sqrt{\frac{(\log M + x)\log n}{n}}, \|f - f^F\|\Big).$$

*Also, with probability at least $1 - 4e^{-x}$, we have for every $f, g \in F$:*

$$\big|\|f-g\|_n^2 - \|f-g\|^2\big| \le cb\sqrt{\frac{(\log M + x)\log n}{n}} \max\Big(b\sqrt{\frac{(\log M + x)\log n}{n}}, \|f-g\|\Big).$$

*If* (2) *holds, we have, with probability larger than $1 - 2e^{-x}$, that for every $f \in F$:*

$$\Big|\frac{1}{n}\sum_{i=1}^n \mathcal{L}_f(X_i,Y_i) - \mathbb{E}\mathcal{L}_f(X,Y)\Big| \le cb\sqrt{\frac{\log M + x}{n}} \max\Big(b\sqrt{\frac{\log M + x}{n}}, \|f - f^F\|\Big),$$

*and with probability at least $1 - 2e^{-x}$, that for every $f, g \in F$:*

$$\big|\|f-g\|_n^2 - \|f-g\|^2\big| \le cb\sqrt{\frac{\log M + x}{n}} \max\Big(b\sqrt{\frac{\log M + x}{n}}, \|f-g\|\Big).$$

**Proof** [Proof of Lemma 9] The proof uses exactly the same arguments as that of Lemma 7 and Corollary 8, and thus is omitted. ■

**Lemma 10** *Let* $\widehat{F}_1$ *be given by* (4) *and recall that* $f^F \in \operatorname{argmin}_{f\in F} R(f)$ *and let* $d(\widehat{F}_1) = \operatorname{diam}(\widehat{F}_1, L_2(\mu))$.

*If* (3) *holds, we have with probability at least* $1-4\exp(-x)$ *that* $f^F \in \widehat{F}_1$*, and any function* $f \in \widehat{F}_1$ *satisfies*

$$R(f) \leq R(f^F) + c(\sigma_\varepsilon + b)\sqrt{\frac{(\log M + x)\log n}{n}} \max\left(b\sqrt{\frac{(\log M + x)\log n}{n}}, d(\widehat{F}_1)\right).$$

*If* (2) *holds, we have with probability at least* $1-2\exp(-x)$ *that* $f^F \in \widehat{F}_1$*, and any function* $f \in \widehat{F}_1$ *satisfies*

$$R(f) \leq R(f^F) + cb\sqrt{\frac{\log M + x}{n}} \max\left(b\sqrt{\frac{\log M + x}{n}}, d(\widehat{F}_1)\right).$$

**Proof** The proof follows the lines of the proof of Lemma 4.4 in Lecué and Mendelson (2009), together with Lemma 9, so we don't reproduce it here. ■

## References

Radosław Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:no. 34, 1000–1034, 2008. ISSN 1083-6489.

Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37:1591, 2009. URL doi:10.1214/08-AOS623.

Olivier Catoni. *Statistical Learning Theory and Stochastic Optimization*. Ecole d'été de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics. Springer, N.Y., 2001.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *COLT*, pages 97–111, 2007.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. ISSN 0090-5364. With discussion, and a rejoinder by the authors.

Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *Ann. Probab.*, 12(4): 929–998, 1984. ISSN 0091-1798.

Anatoli Juditsky, Alexander V. Nazin, Alexandre B. Tsybakov, and Nicolas Vayatis. Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii*, 41(4):78–96, 2005. ISSN 0555-2923.

Anatoli Juditsky, Philippe Rigollet, and Alexandre B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 2008. ISSN 0090-5364. doi: 10.1214/07-AOS546. URL http://dx.doi.org/10.1214/07-AOS546.

Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, 145(3-4):591–613, 2009. ISSN 0178-8051. doi: 10.1007/s00440-008-0180-8. URL https://dx.doi.org/10.1007/s00440-008-0180-8.

Guillaume Lecué and Shahar Mendelson. On the optimality of the aggregate with exponential weights for small temperatures. *Submitted*, 2010.

Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006. ISSN 0018-9448.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Michel Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996. ISSN 0020-9910.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58 (1):267–288, 1996. ISSN 0035-9246.

Alexandre. B. Tsybakov. Optimal rates of aggregation. *Computational Learning Theory and Kernel Machines. B.Schölkopf and M.Warmuth, eds. Lecture Notes in Artificial Intelligence*, 2777:303–313, 2003. Springer, Heidelberg.

Yuhong Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 2000. ISSN 0090-5364. doi: 10.1214/aos/1016120365. URL http://dx.doi.org/10.1214/aos/1016120365.

Yuhong Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004. ISSN 1350-7265.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x. URL http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.