Aggregation and minimax optimality in high-dimensional estimation

Alexandre B. Tsybakov*

Abstract. Aggregation is a popular technique in statistics and machine learning. Given a collection of estimators, the problem of linear, convex or model selection type aggregation consists in constructing a new estimator, called the aggregate, which is nearly as good as the best among them (or nearly as good as their best linear or convex combination), with respect to a given risk criterion. When the underlying model is sparse, which means that it is well approximated by a linear combination of a small number of functions in the dictionary, aggregation techniques turn out to be very useful in taking advantage of sparsity. On the other hand, aggregation is a general way of constructing adaptive nonparametric estimators, which is more powerful than the classical methods since it allows one to combine estimators of different nature. Aggregates are usually constructed by mixing the initial estimators or functions of the dictionary with data-dependent weights that can be defined is several possible ways. An important example is given by aggregates with exponential weights. They satisfy sharp oracle inequalities that allow one to treat in a unified way three different problems: Adaptive nonparametric estimation, aggregation and sparse estimation.

Mathematics Subject Classification (2010). Primary 62G05; Secondary 62J07.

Keywords. High-dimensional model, aggregation, sparsity, oracle inequality, minimax estimation, exponential weights.

1. Introduction

Aggregation of estimators in the regression model has been studied starting from [27, 7, 39, 33]. In this paper, we focus on the connection between aggregation and high-dimensional statistics. In particular, we show that some aggregation techniques, such as exponential weighting, achieve minimax rates in high-dimensional problems with sparsity constraints in an adaptive way. The results obtained for such methods are better than those available for the Lasso and related ℓ_1-penalized techniques. Furthermore, the procedure of exponential weighting achieves the task of universal aggregation.

*This work is supported by GENES, and by the French National Research Agency (ANR) under the grants Idex ANR -11- IDEX-0003-02, Labex ECODEC (ANR - 11-LABEX-0047), and IPANEMA (ANR-13-BSH1-0004-02). The author is grateful to Cowles Foundation and to the Department of Statistics of Yale University for their hospitality during the writing of this paper.

We consider the Gaussian regression model with fixed design. Suppose that we observe $\{(X_i, Y_i)\}_{i=1}^n$ such that

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

where $\mathcal{X}$ is an arbitrary set, $f : \mathcal{X} \to \mathbb{R}$ is an unknown function, $X_i \in \mathcal{X}$ are nonrandom, and ξ_i are independent random variables. Unless explicitly stated otherwise, we will assume that ξ_i are independent identically distributed (i.i.d.) Gaussian random variables with mean zero and variance σ^2, $\xi_i \sim \mathcal{N}(0, \sigma^2)$.

Let $\hat{f}$ be an estimator of f based on the observations $\{(X_i, Y_i)\}_{i=1}^n$. To measure the performance of $\hat{f}$, we use the squared error loss of the form

$$\|\hat{f} - f\|^2 = \frac{1}{n} \sum_{i=1}^n (\hat{f}(X_i) - f(X_i))^2$$

and we define the squared risk of estimator $\hat{f}$ as $E\|\hat{f} - f\|^2$ where E denotes the expectation. The pseudo-norm $\|f\|$ is referred to as the *empirical norm* of a function defined on $\mathcal{X}$. For vectors $b \in \mathbb{R}^n$, we will also consider the empirical ℓ_2-norm defined by $\|b\|^2 = \frac{1}{n} \sum_{i=1}^n b_i^2$, while $|b|_2^2 = \sum_{i=1}^n b_i^2$ defines the usual ℓ_2-norm $|b|_2$.

Assume that we are given a collection of functions $\{f_1, \dots, f_M\}$ called the *dictionary*, where $f_j : \mathcal{X} \to \mathbb{R}$. Assume also that we are given a subset Θ of $\mathbb{R}^M$. For $\theta = (\theta_1, \dots, \theta_M) \in \Theta$ we consider the linear combinations f_θ defined by

$$\mathsf{f}_\theta(x) \stackrel{\text{def}}{=} \sum_{j=1}^M \theta_j f_j(x), \quad x \in \mathcal{X}.$$

Functions f_θ are viewed as approximations of the unknown f. Choosing the dictionary $\{f_1, \dots, f_M\}$ to be rich enough and M sufficiently large, one can expect f_θ to be close to f under appropriate assumptions. This motivates the study of estimator of f having the form

$$\hat{f} = \mathsf{f}_{\hat{\theta}} = \sum_{j=1}^M \hat{\theta}_j f_j,$$

where $\hat{\theta}_j$ are suitable estimators of the coefficients θ_j. The overall aim is to mimimize the risk by choosing an optimal $\hat{\theta}_j$. However, depending on the assumptions that we make about the dictionary, the set Θ and f, we are led to different optimality properties. We introduce here three different settings and discuss the corresponding minimax optimality frameworks.

1.1. Setting 1: Linear Regression and Sparsity.

Assume that f is a linear combination of functions from the dictionary:

$$\exists \theta^* \in \mathbb{R}^M : \quad f(x) = \mathsf{f}_{\theta^*}(x) = \sum_{j=1}^M \theta_j^* f_j(x). \tag{2}$$

Then (1) takes the form of a linear regression model, i.e., it can be written as

$$y = X\theta^* + \xi, \tag{3}$$

where $y = (Y_1, \dots, Y_n)^T$, $\xi = (\xi_1, \dots, \xi_n)^T$, and $X \in \mathbb{R}^{n \times M}$ is a deterministic matrix with entries $f_j(X_i), i = 1, \dots, n, j = 1, \dots, M$. Estimation of f is now reduced to estimation of θ^* in (3). Classical theory of linear regression deals with the case $n \le M$, which is a necessary condition of identifiability of θ^* when we only know that $\theta^* \in \mathbb{R}^M$. However, motivated by several applications, recent years have witnessed an increasing interest in problems where M is greater than n and often $M \gg n$. In this case, f is not identifiable without additional assumptions on θ^*. A natural and most popular additional assumption is a sparsity constraint. It consists in restricting the parameter θ^* to the class $\Theta = B_0(s)$ where $B_0(s)$ is the ℓ_0-ball in $\mathbb{R}^M$:

$$B_0(s) = \{\theta \in \mathbb{R}^M : |\theta|_0 \le s\}, \quad s = 1, \dots, M. \tag{4}$$

Here,

$$|\theta|_0 \stackrel{\text{def}}{=} \sum_{j=1}^{M} I(\theta_j \neq 0)$$

is the "ℓ_0 norm". Vectors θ belonging to $B_0(s)$ are called s-sparse. It turns out that, under the s-sparsity restriction, estimation with reasonable accuracy is possible. A natural question arising in this context is: What is the optimal way to estimate θ^* if we know that $\theta^* \in B_0(s)$?

We will consider optimality in a minimax sense. Let $\hat{\theta}$ be an estimator of θ^*. The corresponding estimator of f is then $\hat{f} = \mathsf{f}_{\hat{\theta}}$ and, in view of (2) - (3), the squared risk takes the form

$$E\|\hat{f} - f\|^2 = E_{\theta^*}\left(\frac{1}{n}|X(\hat{\theta} - \theta^*)|_2^2\right).$$

Here and below, E_θ denotes the expectation with respect to the distribution of $y = X\theta + \xi$ where ξ is a Gaussian vector in $\mathbb{R}^n$ with i.i.d. components $\xi_i \sim \mathcal{N}(0, \sigma^2)$.

An estimator $\hat{\theta}$ is called minimax optimal on the class $B_0(s)$ if there exists a sequence of positive numbers $\psi_{n,M,s}$ such that, for all n and M, the following two conditions are satisfied:

$$\sup_{\theta^* \in B_0(s)} E_{\theta^*}\left(\frac{1}{n}|X(\hat{\theta} - \theta^*)|_2^2\right) \le C\psi_{n,M,s}, \tag{5}$$

$$\inf_T \sup_{\theta^* \in B_0(s)} E_{\theta^*}\left(\frac{1}{n}|X(T - \theta^*)|_2^2\right) \ge c\psi_{n,M,s} \tag{6}$$

where C and c are positive constants independent of n, M, s, and $\inf_T$ denotes the infimum over all estimators of θ^* based on the sample $\{(X_i, Y_i)\}_{i=1}^n$ satisfying the model (3). This property is commonly referred to as the *minimax optimality*. A sequence $\psi_{n,M,s}$ such that (5) and (6) hold is called *minimax rate of convergence*

(or *optimal rate of convergence*) on $B_0(s)$. Our main goal in this setting is to find a minimax optimal estimator $\hat{\theta}$ on the class $B_0(s)$. Along with $B_0(s)$, other classes can be considered, such as ℓ_q-balls $B_q(\delta) = \{\theta \in \mathbb{R}^M : |\theta|_q \le \delta\}$, where $|\theta|_q = (\sum_{j=1}^M |\theta_j|^q)^{\frac{1}{q}}$, $0 < q < \infty$, $\delta > 0$. The notion of optimality is defined for them analogously. This problem, in its particular case where $X^T X/n$ is the identity matrix, and $M = n$ (called the Gaussian sequence model) and with an asymptotic point of view ($n \to \infty$), has been in the focus of the statistical literature since the 1990ies [16, 17, 1]; for its non-asymptotic treatment, see [4]. We are interested here in a more general linear regression setting and we deal with the non-asymptotic minimax optimality. We also consider more general classes, such as intersection of ℓ_0-ball with ℓ_q-ball, $0 < q \le 2$, for which we propose an adaptive estimator. Here, adaptivity means that the estimator is independent of s, q, and of the radius δ of the ℓ_q-ball and it achieves the minimax rates simultateously for all $1 \le s \le M$, $\delta > 0$, $0 < q \le 2$.

1.2. Setting 2: Nonparametric Regression. Let $\mathcal{F}_{\beta,L}$ be a class of smooth functions on $\mathcal{X} \subseteq \mathbb{R}^d$ indexed by $\beta > 0$ and $L > 0$. Common examples of $\mathcal{F}_{\beta,L}$ are balls in a Sobolev or Besov space (see [34, 19] for more details). Assume that $f \in \mathcal{F}_{\beta,L}$. The parameter β typically stands for the number of derivatives of f that are assumed bounded in some norm by L, the radius of the ball. In this setting, the dictionary $\{f_1, \dots, f_M\}$ is usually composed of the first $M = n$ functions of some orthonormal basis. For example, it can be the Fourier or wavelet basis. A key property in the nonparametric regression setting (following from the definition of the class $\mathcal{F}_{\beta,L}$) is that f can be well approximated by a linear combination of basis functions. It can be stated, for example, as follows: For any $f \in \mathcal{F}_{\beta,L}$, and any $k = 1, \dots, n$, there exists $\theta^* = \theta^*(f) \in \mathbb{R}^k$ such that

$$\left\| f - \sum_{j=1}^{k} \theta_j^* f_j \right\| \le C(\beta, L) k^{-\beta}, \tag{7}$$

where $C(\beta, L)$ is a constant depending only on β, L. A minimax optimal estimator $\hat{f}$ is the estimator that satisfies, for all n,

$$\sup_{f \in \mathcal{F}_{\beta,L}} E\|\hat{f} - f\|^2 \le C\psi_{n,\beta}, \tag{8}$$

$$\inf_{\tilde{f}} \sup_{f \in \mathcal{F}_{\beta,L}} E\|\tilde{f} - f\|^2 \ge c\psi_{n,\beta}, \tag{9}$$

where C and c are positive constants independent of n, β, and $\inf_{\tilde{f}}$ denotes the infimum over all estimators of f based on the sample $\{(X_i, Y_i)\}_{i=1}^n$. A positive sequence $\psi_{n,\beta}$ such that (8) and (9) hold is called the *minimax rate of convergence* (or *optimal rate of convergence*) on $\mathcal{F}_{\beta,L}$.

Along with finding minimax optimal estimators in this setting, the second important issue is adaptivity: How to construct *adaptive estimators* that is estimators $\hat{f}$, which are independent of β and L and satisfy (8) with optimal rate of convergence $\psi_{n,\beta}$ for all pairs (β, L) in a wide range of values?

1.3. Setting 3: Aggregation of estimators. This setting will be the main object of study below. Suppose that we are given a collection of estimators $\hat{f}_1, \dots, \hat{f}_M$ of f and a subset Θ of $\mathbb{R}^M$. The goal is to find a new estimator $\tilde{f}$, called *aggregate*, which is approximately at least as good as the best linear combination $\mathsf{f}_\theta = \sum_{j=1}^M \theta_j \hat{f}_j$ with weights θ in the set Θ. The best linear combination is defined as the one that solves the minimization problem

$$\min_{\theta \in \Theta} E\|f - \mathsf{f}_\theta\|^2.$$

Unlike the previous two settings, here f_θ is a random function depending on the data, and we *do not assume* that $\|f - \mathsf{f}_\theta\|$ is zero or small (cf. (2), (7)); it may happen that all f_θ for some Θ are very far from f. Some common examples of Θ are the following.

(L) *Linear aggregation:* $\Theta = \mathbb{R}^M$. The aim of linear aggregation is to construct an estimator $\tilde{f}$, which is approximately as good as the best linear combination of the initial estimators $\hat{f}_1, \dots, \hat{f}_M$.

(C) *Convex aggregation:* Θ is the simplex

$$\Theta = \Lambda^M \stackrel{\text{def}}{=} \Big\{\theta \in \mathbb{R}^M : \theta_j \ge 0, \sum_{j=1}^M \theta_j = 1\Big\}.$$

Convex aggregation aims to find an estimator $\tilde{f}$, which is approximately as good as the best convex combination of the initial estimators $\hat{f}_1, \dots, \hat{f}_M$.

(MS) *Model Selection type aggregation:* $\Theta = \{e_1, \dots, e_M\}$ where e_i are the canonical basis vectors in $\mathbb{R}^M$. The MS-aggregation aims to mimic the best among the initial estimators $\hat{f}_1, \dots, \hat{f}_M$.

Other types of aggregation can be considered as well, for example, the *s-sparse aggregation* corresponding to $\Theta = B_0(s)$, or the ℓ_q-aggregation corresponding to $\Theta = B_q(\delta)$ with $0 < q < \infty$, $\delta > 0$.

The goal of aggregation is to mimic the best linear combination of initial estimators with weights restricted to a given set Θ of possible weights. The word "best" here is formalized as choosing $\tilde{f}$ with the smallest possible *excess risk* (also known under the name of *regret*) defined by

$$\mathcal{E}_\Theta(\tilde{f}, f) \stackrel{\text{def}}{=} E\|\tilde{f} - f\|^2 - \inf_{\theta \in \Theta} E\|\mathsf{f}_\theta - f\|^2. \tag{10}$$

Based on the excess risk, we can introduce the concept of minimax optimality for aggregation [33]. An estimator $\tilde{f}$ is called an *optimal aggregate for the class* Θ if there exists a sequence of positive numbers $\psi_{n,M}(\Theta)$ such that, for all n and M,

$$\sup_{\hat{f}_1, \dots, \hat{f}_M} \big\{ \sup_f \mathcal{E}_\Theta(\tilde{f}, f)\big\} \le C\psi_{n,M}(\Theta), \tag{11}$$

$$\sup_{\hat{f}_1, \dots, \hat{f}_M} \big\{ \inf_{\hat{f}} \sup_f \mathcal{E}_\Theta(\hat{f}, f)\big\} \ge c\psi_{n,M}(\Theta). \tag{12}$$

Here, $\inf_{\hat f}$ is the infimum over all estimators, C and c are positive constants independent of n and M, and $\sup_f$, $\sup_{\hat f_1,\dots,\hat f_M}$ are the suprema over all possible functions f and over wide classes of preliminary estimators. In some cases, these will be all possible estimators $\hat f_1,\dots,\hat f_M$ with no restriction; in other cases it will suffice to consider classes of $\hat f_1,\dots,\hat f_M$ such that $\hat f_j$'s are bounded in the empirical norm $\|\cdot\|$ uniformly over j. If (11) and (12) hold for some positive sequence $\psi_{n,M}(\Theta)$, this sequence is called an *optimal rate of aggregation for the class* Θ [33]. Two questions arise in this context. First, how to construct an optimal aggregate $\tilde f$ for a given class Θ? Second, is it possible to construct a *universal aggregate*, i.e., an aggregate which is optimal simultaneously for a large scale of classes Θ?

Inequalities (11) and (12) establish upper and lower bounds for the minimax regret, respectively. The upper bound (11) can be equivalently written in the form of *oracle inequality*[1]

$$E\|\tilde f - f\|^2 \le \inf_{\theta\in\Theta} E\|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(\Theta), \tag{13}$$

that should be valid for all $\hat f_1,\dots,\hat f_M$ in a wide class, and for all f. This garantees that the risk of the suggested aggregate $\tilde f$ is at least as good as the risk of the unknown *oracle* θ^* minimizing $E\|\mathsf{f}_\theta - f\|^2$, up to a remainder term of the order $\psi_{n,M}(\Theta)$, which characterizes the price to pay for aggregation. Lower bounds (12) ensure that this is the minimal price; the remainder term cannot be of smaller order whatever the aggregate is. For sparsity classes, for example, when $\Theta = B_0(s)$, the rate $\psi_{n,M}(\Theta)$ is a function of s; the corresponding oracle inequalities are called *sparsity oracle inequalities*.

2. Reduction to aggregation of functions

Aggregates are usually constructed in the form

$$\tilde f = \sum_{j=1}^M \hat\theta_j \hat f_j$$

where $\hat\theta_j$ are suitably chosen statistics measurable with respect to the data, and $\hat f_j$ are the preliminary estimators. In what follows, we will assume that $\hat\theta_j$ and estimators $\hat f_j$ are stochastically independent. This can be achieved by creating two independent samples from the initial sample $\{(X_i,Y_i)\}_{i=1}^n$ by randomization (*sample cloning*), cf. [27]. The estimators $\hat f_j$ are constructed from the first sample while the second one is used to perform aggregation, i.e., to compute the weights $\hat\theta_j$. To carry out the analysis of aggregation, it is enough to work conditionally on the first sample, so that $\hat f_j$ can be considered as deterministic functions. Thus, the problem reduces to aggregation of deterministic functions that we will denote as previously

[1] Here and in the sequel, we denote by C positive constants, possibly different on different appearances.

$f_j \stackrel{\text{def}}{=} \hat{f}_j$, $j = 1, \dots, M$. The procedure of sample cloning by randomization is based on the following elementary fact.

Lemma 2.1. *Let $Y_i = f(X_i) + \xi_i$. Let ω_i be a standard normal random variable independent of ξ_i. Set $Y_{i1} = Y_i + \sigma\omega_i$, and $Y_{i2} = Y_i - \sigma\omega_i$. Then we have $Y_{i1} = f(X_i) + \xi_{i1}$, and $Y_{i2} = f(X_i) + \xi_{i2}$, where $\xi_{i1} \sim \mathcal{N}(0, 2\sigma^2)$, $\xi_{i2} \sim \mathcal{N}(0, 2\sigma^2)$ and ξ_{i1} is independent of ξ_{i2}.*

Thus, by adding to and subtracting from the observations Y_i the variables $\sigma\omega_i$, we obtain two independent Gaussian n-samples $D_1 = \{(X_i, Y_{i1})\}_{i=1}^n$ and $D_2 = \{(X_i, Y_{i2})\}_{i=1}^n$, where $Y_{ik} = f(X_i) + \xi_{ik}$, $k = 1, 2$. The observations in both samples are of the same form as in the original sample $\{(X_i, Y_i)\}_{i=1}^n$, with the only difference that the variance of the noise is doubled.

Now, we use D_1 to construct preliminary estimators $\hat{f}_1, \dots, \hat{f}_M$ and we use D_2 to determine the weights $\hat{\theta}_1, \dots, \hat{\theta}_M$. Denoting by $E_{(k)}$ the expectations with respect to the distribution of D_k for $k = 1, 2$, we may write the oracle inequality (13) that we need to prove in the form

$$E_{(1)}E_{(2)}\|\tilde{f} - f\|^2 \le \inf_{\theta\in\Theta} E_{(1)}\|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(\Theta). \tag{14}$$

To obtain (14), it suffices to show that, for fixed functions $f_1, \dots, f_M, f$, we have

$$E_{(2)}\|\tilde{f} - f\|^2 \le \inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(\Theta), \tag{15}$$

where $\mathsf{f}_\theta = \sum_{j=1}^M \theta_j f_j$, and $\tilde{f} = \sum_{j=1}^M \hat{\theta}_j f_j$ with $\hat{\theta}_j$ measurable with respect to D_2.

Thus, using the sample cloning device, we can reduce aggregation of estimators to its special case, which is *aggregation of fixed functions*. This will be the setting considered in the rest of the paper. In this case, the minimax framework of aggregation (cf. Setting 3 in the Introduction) changes only in that the excess risk takes the form

$$\mathcal{E}_\Theta(\tilde{f}, f) \stackrel{\text{def}}{=} E\|\tilde{f} - f\|^2 - \inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 \tag{16}$$

(no expectation in the term $\inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2$). Accordingly, an estimator $\tilde{f}$ is called an *optimal aggregate for the class* Θ if there exists a sequence of positive numbers $\psi_{n,M}(\Theta)$ such that (11) and (12) are satisfied where instead of $\hat{f}_j$ we have fixed functions f_j. Upper bounds on the maximum excess risk are then equivalent to oracle inequalities

$$E\|\tilde{f} - f\|^2 \le \inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(\Theta). \tag{17}$$

Such an oracle inequality being established, we can obtain upper bounds for the minimax risks in Settings 1 and 2 as simple corollaries. Indeed, those settings impose additional strong restrictions on f; the oracle risk $\inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2$ is 0 in Setting 1, and it admits a bound such as (7) in Setting 2. In Setting 3, the oracle risk can be arbitrary, therefore we use only the excess risk to measure the performance of aggregates.

3. Least squares aggregation

A first simple idea is to construct aggregates by minimizing the least squares (LS) criterion. Given a set Θ and a collection of deterministic functions $f_1, \dots, f_M$, we take

$$\hat{\theta}^{LS}(\Theta) \in \operatorname*{argmin}_{\theta \in \Theta} \|y - \mathsf{f}_\theta\|^2$$

and we define the LS aggregate as

$$\tilde{f} = \mathsf{f}_{\hat{\theta}^{LS}(\Theta)} = \sum_{j=1}^{M} \hat{\theta}_j^{LS}(\Theta) f_j .$$

Proposition 3.1. *Let $\hat{\theta}^{LS} \stackrel{\text{def}}{=} \hat{\theta}^{LS}(\mathbb{R}^M)$ be a global least squares estimator. Assume that $E(\xi_i) = 0$, $E(\xi_i \xi_j) = 0$, if $i \neq j$ for $i, j = 1, \dots, n$. If $E(\xi_i^2) = \sigma^2$, $i = 1, \dots, n$, then for all $f, f_1, \dots, f_M$, and integers $n \geq 1$, $M \geq 1$, we have*

$$E\|\mathsf{f}_{\hat{\theta}^{LS}} - f\|^2 = \min_{\theta \in \mathbb{R}^M} \|\mathsf{f}_\theta - f\|^2 + \frac{\sigma^2 R}{n} . \tag{18}$$

where $R = \mathrm{Rank}(X)$ denotes the rank of matrix X. Furthermore, if $E(\xi_i^2) \leq \sigma^2$, $i = 1, \dots, n$, then for any convex set $\Theta \subset \mathbb{R}^M$,

$$E\|\mathsf{f}_{\hat{\theta}^{LS}(\Theta)} - f\|^2 \leq \min_{\theta \in \Theta} \|\mathsf{f}_\theta - f\|^2 + \frac{4\sigma^2 R}{n} . \tag{19}$$

Proof. We prove only (19) since (18) follows from a simple orthogonal decomposition, cf., e.g., [31]. Below, we will denote by f and f_θ not only the functions from $\mathcal{X}$ to $\mathbb{R}$ but also the n-vectors of values of these functions at points $X_1, \dots, X_n$. Then, the model of observations (1) can be written as $y = f + \xi$, and $\mathsf{f}_\theta = X\theta$ for all θ. Set for brevity $\tilde{f} = \mathsf{f}_{\hat{\theta}^{LS}(\Theta)}$. From the definition of this estimator we get by a simple algebra that, for any $\theta \in \Theta$,

$$\|\tilde{f} - f\|^2 \leq \|\mathsf{f}_\theta - f\|^2 + 2\langle \tilde{f} - \mathsf{f}_\theta, \xi \rangle \tag{20}$$

where $\langle f, g \rangle \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n f(X_i) g(X_i)$. On the other hand, $\tilde{f} - \mathsf{f}_\theta \in \mathrm{Im}(X)$, and thus $\langle \tilde{f} - \mathsf{f}_\theta, \xi \rangle = \langle \tilde{f} - \mathsf{f}_\theta, A\xi \rangle$ where A is the orthogonal projector on $\mathrm{Im}(X)$. This and (20) imply

$$\|\tilde{f} - f\|^2 \leq \|\mathsf{f}_\theta - f\|^2 + \frac{1}{2}\|\tilde{f} - \mathsf{f}_\theta\|^2 + 2\|A\xi\|^2 . \tag{21}$$

Since Θ is convex, for all $\theta' \in \Theta$ we have $\|f - \mathsf{f}_{\theta^*}\|^2 + \|\mathsf{f}_{\theta'} - \mathsf{f}_{\theta^*}\|^2 \leq \|\mathsf{f}_{\theta'} - f\|^2$ where $\theta^* = \operatorname*{argmin}_{\theta \in \Theta} \|\mathsf{f}_\theta - f\|^2$. This inequality with $\theta' = \hat{\theta}^{LS}(\Theta)$, and (21) with $\theta = \theta^*$ imply that $\|\tilde{f} - f\|^2 \leq \|\mathsf{f}_{\theta^*} - f\|^2 + 4\|A\xi\|^2$. Now, (19) follows by taking here the expectations and noticing that $E\|A\xi\|^2 \leq \frac{\sigma^2 R}{n}$. □

Proposition 3.2. *Let Θ be a subset of the simplex Λ^M, and let $\xi_1, \dots, \xi_n$ be independent zero mean σ-subgaussian random variables, i.e., $E\exp(s\xi_i) \le \exp(s^2\sigma^2/2)$ for all $s > 0$, $i = 1, \dots, n$. Then, for all f, all integers $n \ge 1$, $M \ge 2$, and all dictionaries $\{f_1, \dots, f_M\}$ such that $\max_{j=1,\dots,M} \|f_j\| \le L$, we have*

$$E\|\mathsf{f}_{\hat\theta^{LS}(\Theta)} - f\|^2 \le \inf_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 + 2\sigma L\sqrt{\frac{2\log M}{n}}\,.$$

Proof. In view of (20), it suffices to prove that $E\langle \tilde f, \xi\rangle \le \sigma L\sqrt{\frac{2\log M}{n}}$ where $\tilde f = \mathsf{f}_{\hat\theta^{LS}(\Theta)}$. But $E\langle \tilde f, \xi\rangle \le E\max_{\theta'\in\Lambda^M}\langle f_{\theta'}, \xi\rangle = E\max_{1\le j\le M}\langle f_j, \xi\rangle$ and the random variable $\langle f_j, \xi\rangle$ is $\bar\sigma$-subgaussian with $\bar\sigma = \sigma\|f_j\|/\sqrt{n} \le \sigma L/\sqrt{n}$. By the standard properties of subgaussian variables, $E\max_{1\le j\le M}\langle f_j, \xi\rangle \le \bar\sigma\sqrt{2\log M}$. □

Consider now convex aggregation and MS-aggregation by the LS method. The corresponding weights are $\hat\theta^{LS}_{\mathrm{conv}} \stackrel{\mathrm{def}}{=} \hat\theta^{LS}(\Lambda^M)$, and $\hat\theta^{LS}_{\mathrm{MS}} \stackrel{\mathrm{def}}{=} \hat\theta^{LS}(\{e_1, \dots, e_M\})$. The MS-aggregate selects one function in the dictionary:

$$\mathsf{f}_{\hat\theta^{LS}_{\mathrm{MS}}} = f_{\hat\jmath} \quad \text{where} \quad \hat\jmath \in \mathrm{argmin}_{1\le j\le M}\|y - f_j\|^2.$$

The following corollaries are straightforward in view of Propositions 3.1 and 3.2.

Corollary 3.3 (Convex aggregation). *For all f, all integers $n \ge 1$, $M \ge 2$, and all dictionaries $\{f_1, \dots, f_M\}$ such that $\max_{j=1,\dots,M} \|f_j\| \le L$, we have*

$$E\|\mathsf{f}_{\hat\theta^{LS}_{\mathrm{conv}}} - f\|^2 \le \min_{\theta\in\Lambda^M} \|\mathsf{f}_\theta - f\|^2 + \left(\frac{4\sigma^2 R}{n} \wedge 2\sigma L\sqrt{\frac{2\log M}{n}}\right).$$

Corollary 3.4 (MS-aggregation). *For all f, all integers $n \ge 1$, $M \ge 2$, and all dictionaries $\{f_1, \dots, f_M\}$ such that $\max_{j=1,\dots,M} \|f_j\| \le L$, we have*

$$E\|\mathsf{f}_{\hat\theta^{LS}_{\mathrm{MS}}} - f\|^2 \le \min_{1\le j\le M} \|f_j - f\|^2 + 2\sigma L\sqrt{\frac{2\log M}{n}}\,.$$

The rate of aggregation $\frac{\sigma^2 R}{n}$ of the global least squares estimator given in (18) is the optimal rate of linear aggregation, see Section 8 below and [33, 6, 31]. Also, the rate of the convex aggregate $\mathsf{f}_{\hat\theta^{LS}_{\mathrm{conv}}}$ given in Corollary (3.3) is the optimal rate of convex aggregation up to a minor discrepancy in the expression under the logarithm [33, 31]. However, for MS-aggregation the situation is different. The optimal rate for MS-aggregation is of the order $(\log M)/n$ [33, 31], while the LS-aggregate $\mathsf{f}_{\hat\theta^{LS}_{\mathrm{MS}}}$ achieves only the rate $\sqrt{(\log M)/n}$ according to Corollary 3.4. Moreover, it turns out that $\mathsf{f}_{\hat\theta^{LS}_{\mathrm{MS}}}$ cannot do better; the upper bound of Corollary 3.4 is tight for it. Indeed, the next theorem shows that not only the least squares MS-aggregate but also any method that selects a single function in the dictionary cannot have faster rate. This includes methods of model selection by penalized empirical risk minimization. We call estimators $\hat S_n$ taking values in $\{f_1, \dots, f_M\}$ the *selectors*.

Theorem 3.5. [32] *Assume that $n \geq 1$, $M \geq 2$ are such that*

$$(\sigma \vee 1)\sqrt{(\log M)/n} \leq C_0$$

for $0 < C_0 < 1$ small enough. Then, there exists a dictionary $\{f_1, \dots, f_M\}$ with $\|f_j\| \leq 1$, $j = 1, \dots, M$, such that the following holds. For any selector $\hat{S}_n$, there exists a regression function f such that $\|f\| \leq 1$ and

$$E\|\hat{S}_n - f\|^2 \geq \min_{1 \leq j \leq M} \|f_j - f\|^2 + C_* \sigma \sqrt{\frac{\log M}{n}}$$

for some positive constant C_*.

A related result about suboptimality of selectors when $\hat{f}_j$ are preliminary estimators rather than fixed functions is proved in [18].

Thus, we see that choosing one of the functions in a finite dictionary to solve the problem of model selection is suboptimal in the sense that the rate $\sqrt{(\log M)/n}$ is too slow. A natural idea is to extend the class of estimators by taking a convex combination of the functions in the dictionary rather than selecting one function. It turns out that this is sufficient; under a particular choice of weights in this convex combination, namely the exponential weights, one can achieve oracle inequalities with the optimal rate $(\log M)/n$.

4. Exponentially weighted aggregates

Let $f_1, \dots, f_M$ be a given dictionary of functions. Consider the exponentially weighted aggregate

$$\hat{f}^{EW} \stackrel{\text{def}}{=} \mathsf{f}_{\hat{\theta}^{EW}} = \sum_{j=1}^{M} \hat{\theta}_j^{EW} f_j$$

where the weights $\hat{\theta}^{EW} = (\hat{\theta}_1^{EW}, \dots, \hat{\theta}_M^{EW})$ are defined as

$$\hat{\theta}_j^{EW} = \frac{\exp(-n\hat{r}_j/\beta)\pi_j}{\sum_{k=1}^{M} \exp(-n\hat{r}_k/\beta)\pi_k}\,.$$

Here, $\hat{r}_j = \|y - f_j\|^2$ is the empirical risk corresponding to function f_j, $\beta > 0$ is a tuning parameter, and $\pi_1, \dots, \pi_M$ is a set of prior probabilities, $\pi_k > 0$, $\sum_{k=1}^{M} \pi_k = 1$. This definition dates back at least to [37] where the method was introduced in the context of the theory of prediction of deterministic individual sequences. It is now a popular tool in that theory, cf. [8, 18] where one can find further references.

Note that

$$\hat{\theta}^{EW} = \operatorname*{argmin}_{\theta \in \Lambda^M} \left(\sum_{j=1}^{M} \theta_j \hat{r}_j + \frac{\beta}{n} \mathcal{K}(\theta, \pi) \right) \tag{22}$$

where $\mathcal{K}(\theta,\pi)=\sum_{j=1}^M \theta_j \log\frac{\theta_j}{\pi_j}\ge 0$ (with the convention that $0\cdot\log 0=0$) is the Kullback-Leibler divergence between the discrete probability measures defined by the probability vectors $\theta\in\Lambda^M$ and $\pi\in\Lambda^M$. Since, by Jensen's inequality, $\sum_{j=1}^M \theta_j\hat r_j \ge \|y-\mathsf{f}_\theta\|^2$, we see that $\hat\theta^{EW}$ minimizes, over the simplex Λ^M, an upper bound on the empirical risk penalized by the Kullback-Leibler divergence from π:

$$\|y-\mathsf{f}_\theta\|^2+\frac{\beta}{n}\mathcal{K}(\theta,\pi).$$

So, intuitively, the method penalizes the solution for being far from the prior π.

Theorem 4.1. *For $\beta\ge 4\sigma^2$, and for all $f, f_1,\dots,f_M$, and integers $n\ge 1$, $M\ge 1$, we have*

$$E\|\hat f^{EW}-f\|^2 \le \min_{\theta\in\Lambda^M}\left(\sum_{j=1}^M \theta_j\|f-f_j\|^2+\frac{\beta}{n}\mathcal{K}(\theta,\pi)\right). \tag{23}$$

If the ξ_i are not Gaussian but rather i.i.d. symmetric random variables such that $P(|\xi_i|\le B)=1$ for some finite $B>0$, then (23) holds for any $\beta\ge 4B^2$.

The proof of this theorem can be found in [12, 13, 15] as a special case of more general results relaxing the assumptions on the distribution of ξ_i and allowing for continuous priors (see also [14]). More recent work [29, 9, 23] proposes estimators other than $\hat f^{EW}$ satisfying analogous oracle inequalities both in expectation and in probability.

Note that the right-hand side of (23) is similar to (22). The only difference is that in (22) we have the empirical risks $\hat r_j=\|y-f_j\|^2$ rather than the deterministic discrepancies $\|f-f_j\|^2$. Thus, the minimization problem in (22) is an empirical analog of the right-hand side of (23). An immediate corollary of Theorem 4.1 is the following.

Theorem 4.2. *For $\beta\ge 4\sigma^2$, and for all $f, f_1,\dots,f_M$, and integers $n\ge 1$, $M\ge 1$, we have*

$$E\|\hat f^{EW}-f\|^2 \le \min_{1\le j\le M}\left(\|f-f_j\|^2+\frac{\beta}{n}\log\frac{1}{\pi_j}\right).$$

In particular, if $\pi_j=\frac{1}{M}$, $j=1,\dots,M$, and $M\ge 2$,

$$E\|\hat f^{EW}-f\|^2 \le \min_{1\le j\le M}\|f-f_j\|^2+\frac{\beta}{n}\log M.$$

Thus, the exponentially weighted aggregate achieves the optimal rate of the order $(\log M)/n$, which cannot be attained by the selectors. A result similar to Theorem 4.2 was proved in [24] for the case where f_j are not arbitrary fixed functions but rather the least squares estimators on linear subspaces of $\mathbb{R}^M$. In [24], these estimators are constructed from the same sample y that is used to compute the weights, and the weights are are different:

$$w_j=\frac{\exp\left(-\frac{n\hat r_j}{\beta}-\frac{\dim(j)}{2}\right)\pi_j}{\sum_{k=1}^M \exp\left(-\frac{n\hat r_k}{\beta}-\frac{\dim(k)}{2}\right)\pi_k} \tag{24}$$

where $\dim(j)$ is the dimension of the space on which the jth least squares estimator projects. Extension of the results of [24] to affine estimators are given in [11, 10].

5. Sparsity pattern aggregation

We call a *sparsity pattern* any binary vector $p \in \mathcal{P} \stackrel{\text{def}}{=} \{0,1\}^M$. Denote by $|p| = |p|_0$ the number of ones in p. To each sparsity pattern $p = (p_1, \dots, p_M) \in \mathcal{P}$, we associate a linear subspace S^p of $\mathbb{R}^M$:

$$S^p \stackrel{\text{def}}{=} \operatorname{span}\{e_j : p_j = 1\}, \qquad \dim(S^p) = |p|.$$

From the initial sample y, we clone two randomized independent samples $y^{(1)} \in \mathbb{R}^n$ and $y^{(2)} \in \mathbb{R}^n$ with $\mathcal{N}(0, 2\sigma^2)$ errors as described in Section 2. For each $p \in \mathcal{P}$, we construct a least squares estimator $\hat{\theta}_p$ on S^p based on the first sample $y^{(1)}$:

$$\hat{\theta}_p = \operatorname*{argmin}_{\theta \in S^p} \|y^{(1)} - \mathsf{f}_\theta\|^2. \tag{25}$$

Set $\hat{r}_p = \|y^{(2)} - \mathsf{f}_{\hat{\theta}_p}\|^2$ and define a vector $\hat{\theta}^{SPA} = (\hat{\theta}^{SPA}_p,\ p \in \mathcal{P})$ with components

$$\hat{\theta}^{SPA}_p = \frac{\exp(-\eta \hat{r}_p/\beta)\pi_p}{\sum_{p' \in \mathcal{P}} \exp(-\eta \hat{r}_{p'}/\beta)\pi_{p'}}, \quad \forall p \in \mathcal{P}.$$

Here, $\{\pi_p, p \in \mathcal{P}\}$ is a prior probability measure on $\mathcal{P}$ with $\pi_p \geq 0$ (not necessarily $\pi_p > 0$; values $\pi_p = 0$ are possible, as opposed to the priors in Section 4). The *sparsity pattern aggregate* is defined by

$$\hat{f}^{SPA} \stackrel{\text{def}}{=} \sum_{p \in \mathcal{P}} \hat{\theta}^{SPA}_p \mathsf{f}_{\hat{\theta}_p}.$$

From Theorem 4.2 (where we replace σ^2 by $2\sigma^2$ to account for the sample cloning) we get that if $\beta = 8\sigma^2$, then

$$\forall f: \quad E\|\hat{f}^{SPA} - f\|^2 \leq \min_{p \in \mathcal{P}: \pi_p \neq 0} \left[E\|f_{\hat{\theta}_p} - f\|^2 + \frac{8\sigma^2}{n} \log \frac{1}{\pi_p} \right] \tag{26}$$

while from Proposition 3.1 (again, replacing σ^2 by $2\sigma^2$),

$$\forall f: \quad E\|\mathsf{f}_{\hat{\theta}_p} - f\|^2 \leq \min_{\theta \in S^p} \|\mathsf{f}_\theta - f\|^2 + \frac{2\sigma^2 |p|}{n}. \tag{27}$$

Consider the prior distribution

$$\pi_p = \begin{cases} \left(\binom{M}{|p|} e^{|p|} H \right)^{-1} & \text{if } |p| \leq R, \\ 1/2 & \text{if } |p| = M, \\ 0 & \text{otherwise,} \end{cases} \tag{28}$$

where $H > 0$ is the normalizing constant such that $\sum_{p\in\mathcal{P}} \pi_p = 1$.

Denote by $\hat{f}^{ES}$ the sparsity pattern aggregate $\hat{f}^{SPA}$ with prior $\{\pi_p, p \in \mathcal{P}\}$ given in (28). Following [31], we will call $\hat{f}^{ES}$ the *Exponential Screening* (ES) estimator. The corresponding vector of weights is denoted by $\hat{\theta}^{ES}$. Algorithms of computation of this estimator via Markov Chain Monte-Carlo schemes are discussed and analyzed in [31, 32].

Combining (26) – (28) leads to the following sparsity oracle inequality, cf. [31].

Theorem 5.1. *Let $\hat{f}^{ES}$ be the Exponential Screening estimator with $\beta = 8\sigma^2$. Then for all f, and all integers $n \geq 1$, $M \geq 1$, we have*

$$E\|\hat{f}^{ES} - f\|^2 \leq \min_{\theta\in\mathbb{R}^M} \left[\|\mathsf{f}_\theta - f\|^2 + \frac{C\sigma^2}{n} \left(R \wedge |\theta|_0 \log\left(\frac{eM}{|\theta|_0}\right)\right)\right] + \frac{C'\sigma^2}{n} \qquad (29)$$

where C, C' are absolute positive constants, and $R = \mathrm{Rank}(X)$. Here and in the sequel, $0 \cdot \log\infty = 0$ by convention.

In [31], the oracle inequality (29) is proved for a slightly different estimator, with modified weights (24) and without sample cloning. A weaker result of this form, not taking into account the rank of X, is given in [3]. Inequalities close to (29) that hold with high probability but without accounting for the rank of X are obtained for some estimators different from $\hat{f}^{ES}$ in [10].

Theorem 5.1 is the main result that will allow us to show that one and the same estimator $\hat{f}^{ES}$ achieves the minimax rates of convergence in the three different settings described in the Introduction. Moreover, it achieves these rates adaptively to the parameters of the classes in the first two settings and to the choice of Θ (universal aggregation) in Setting 3. The rest of the paper is devoted to deriving these properties as corollaries of Theorem 5.1.

Remark 5.2. All the results stated below for the estimators $\hat{f}^{ES} = \mathsf{f}_{\hat{\theta}^{ES}}$ and $\hat{\theta}^{ES}$ are also valid for any other estimators $\mathsf{f}_{\hat{\theta}}$ and $\hat{\theta}$ such that (29) holds with $\mathsf{f}_{\hat{\theta}}$ in place of $\hat{f}^{ES}$. Indeed, only (29) will be used in the subsequent argument.

6. Sparsity oracle inequalities on ℓ_0-balls

Theorem 5.1 and monotonicity of the function $x \mapsto x\log(eM/x)$ imply the following upper bounds.

Theorem 6.1. [31] *Let $\hat{f}^{ES}$ be the exponential screening estimator with $\beta = 8\sigma^2$ and let $\hat{\theta}^{ES}$ denote the corresponding vector of weights. Then for all f, and all integers $n \geq 1$, $M \geq 1$, $1 \leq s \leq M$,*

$$E\|\hat{f}^{ES} - f\|^2 \leq \min_{\theta\in B_0(s)} \|\mathsf{f}_\theta - f\|^2 + C\sigma^2 \left(\frac{s}{n}\log\left(\frac{eM}{s}\right) \wedge \frac{R}{n}\right) \qquad (30)$$

where $C > 0$ is an absolute constant. If the model is linear: $y = X\theta + \xi$, then

$$\sup_{\theta \in B_0(s)} E_\theta |X(\hat{\theta}^{ES} - \theta)|_2^2 / n \le C\sigma^2 \left(\frac{s}{n} \log \left(\frac{eM}{s} \right) \wedge \frac{R}{n} \right) . \tag{31}$$

The bounds of Theorem 6.1 are sparsity oracle inequalities. They cannot be improved in a minimax sense, see Section 8 below. Analogous oracle inequalities with leading constant 1 can be also established for the Lasso and related techniques [22] but they need strong assumptions on the dictionary $\{f_1, \dots, f_M\}$ such as the restricted isometry or restricted eigenvalue conditions. In contrast to this, the ES estimator satisfies the sparsity oracle inequalities *under no assumption on the dictionary*. Moreover, Theorem 6.1 shows that the ES estimator simultaneously takes advantage of two types of sparsity: small number of non-zero entries of θ (ℓ_0 norm) and small rank of matrix X. This is not available for the least squares estimators on $B_0(s)$ studied in [28, 36] among others. Note also that the least squares estimators on $B_0(s)$ cannot achieve the excess risk bound (30) with leading constant 1 required for the aggregation setting.

7. Estimation on ℓ_q-balls and on intersection of ℓ_0- and ℓ_q-balls

From Theorem 5.1, we can also deduce oracle inequalities and upper bounds on the risk of the estimator $\hat{f}^{ES}$ on ℓ_q-balls with $0 < q \le 2$. They follow from (29) using the "Maurey argument" as first noticed in [5, 6] for $q = 1$. The proof for $q = 1$ is based on the next lemma (cf. [5, 6, 31]).

Lemma 7.1. *Let $\|f_j\| \le L$, $j = 1, \dots, M$, and $1 \le m \le M$. Then, for any f and any $\theta \in \mathbb{R}^M$ there exists $\theta' \in \mathbb{R}^M$ such that $|\theta'|_0 \le m$ and*

$$\|\mathsf{f}_{\theta'} - f\|^2 \le \|\mathsf{f}_\theta - f\|^2 + \frac{L^2 |\theta|_1^2}{m} .$$

The case $0 < q < 1$ was considered in [38, 35, 10], and the case $1 < q \le 2$ in [35]. Deriving bounds on the risk over ℓ_q-balls with $0 < q < 1$ from (29) can be done based on the following extension of Lemma 7.1.

Lemma 7.2. *Let $\|f_j\| \le L$, $j = 1, \dots, M$, and $1 \le m \le M$. Then, for any f, any $0 < q \le 1$ and any $\theta \in \mathbb{R}^M$ there exists $\bar{\theta} \in \mathbb{R}^M$ such that $|\bar{\theta}|_0 \le 2m$ and*

$$\|\mathsf{f}_{\bar{\theta}} - f\|^2 \le \|\mathsf{f}_\theta - f\|^2 + L^2 |\theta|_q^2 m^{1-2/q} .$$

Proof. By Lemma 7.1, for any $h : \mathcal{X} \to \mathbb{R}$ and any $\theta'' \in \mathbb{R}^M$ there exists $\theta' \in \mathbb{R}^M$ such that $|\theta'|_0 \le m$ and

$$\|\mathsf{f}_{\theta'} - h\|^2 \le \|\mathsf{f}_{\theta''} - h\|^2 + \frac{|\theta''|_1^2 L^2}{m} . \tag{32}$$

Take any $\theta \in \mathbb{R}^M$ and let $J \subseteq \{1, \dots, M\}$ be the set of indices corresponding to the m largest in absolute value components of θ. Take any $f : \mathcal{X} \to \mathbb{R}$ and use (32) with $\theta'' = \theta_{J^c}$, $h = f - \mathsf{f}_{\theta_J}$ where $\theta_J = (\theta_j I(j \in J),\ j = 1, \dots, M)$. Then (32) takes the form

$$\|\mathsf{f}_{\theta'+\theta_J} - f\|^2 \le \|\mathsf{f}_\theta - f\|^2 + \frac{|\theta_{J^c}|_1^2 L^2}{m}\,. \tag{33}$$

Set $\bar\theta = \theta' + \theta_J$. By construction, $|\bar\theta|_0 \le 2m$. Finally, note that

$$|\theta|_{(j)} \le \frac{|\theta|_q}{j^{1/q}} \tag{34}$$

where $|\theta|_{(j)}$ is the jth largest absolute value of the components of θ. Using (34) we get the following bound, which together with (33) yields the lemma:

$$|\theta_{J^c}|_1 = \sum_{j \ge m+1} |\theta|_{(j)} \le |\theta|_{(m)}^{1-q} \sum_{j \ge m+1} |\theta|_{(j)}^q \le \left(\frac{|\theta|_q}{m^{1/q}}\right)^{1-q} |\theta|_q^q = |\theta|_q m^{1-1/q}.$$

□

Lemmas 7.1 and 7.2 combined with Theorem 5.1 imply the following result.

Theorem 7.3. *Assume that $\|f_j\| \le 1$, $j = 1, \dots, M$, and $0 < q \le 1$. Let $\hat f^{ES}$ be the exponential screening estimator with $\beta = 8\sigma^2$ and let $\hat\theta^{ES}$ denote the corresponding vector of weights. Then for any f, and any $\delta > 0$, and integers $n \ge 1$, $M \ge 2$,*

$$E\|\hat f^{ES} - f\|^2 \le \min_{\theta \in B_q(\delta)} \|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(B_q(\delta)) \tag{35}$$

where $C > 0$ is an absolute constant, and

$$\psi_{n,M}(B_q(\delta)) = \sigma^{2-q}\delta^q \left[\frac{1}{n}\log\left(1 + \left(\frac{\sigma}{\delta}\right)^q \frac{M}{n^{q/2}}\right)\right]^{1-q/2} \wedge \frac{\sigma^2 R}{n}\,. \tag{36}$$

Furthermore, if the model is linear, $y = X\theta + \xi$, then for any $n \ge 1$, $M \ge 2$, $1 \le s \le M$, and $\delta > 0$ we have

$$\sup_{\theta \in B_0(s) \cap B_q(\delta)} \frac{1}{n} E_\theta |X(\hat\theta^{ES} - \theta)|_2^2 \le C\bar\psi_{n,M}(\delta, s, q) \tag{37}$$

where $C > 0$ is an absolute constant, and

$$\bar\psi_{n,M}(\delta, s, q) = \psi_{n,M}(B_q(\delta)) \wedge \frac{\sigma^2 s}{n}\log\left(\frac{eM}{s}\right) \wedge \left(\delta^2 + \frac{\sigma^2}{n}\right).$$

Proof. By Theorem 5.1, for an absolute constant $C > 0$ and any $1 \le m \le M/2$,

$$\begin{aligned} E\|\hat f^{ES} - f\|^2 &\le \min_{\theta \in \mathbb{R}^M} \left[\|\mathsf{f}_\theta - f\|^2 + \frac{C\sigma^2}{n}|\theta|_0 \log\left(\frac{eM}{|\theta|_0}\right)\right] + \frac{C\sigma^2}{n} \\ &\le \min_{\theta : |\theta|_0 \le 2m} \|\mathsf{f}_\theta - f\|^2 + \frac{C\sigma^2 m}{n}\log\left(\frac{eM}{2m}\right) \end{aligned}$$

where we have used the monotonicity of the mapping $x \mapsto x \log(eM/x)$. This and Lemma 7.2 imply

$$E\|\hat{f}^{ES} - f\|^2 \le \min_{\theta \in B_q(\delta)} \|\mathsf{f}_\theta - f\|^2 + C\left(\delta^2 m^{1-2/q} + \frac{\sigma^2 m}{n} \log\left(\frac{eM}{m}\right)\right). \qquad (38)$$

Minimizing the right hand side of (38) in m we obtain the first term on the right hand side of (36). The minimum with $\sigma^2 R/n$ comes from Theorem 5.1. Thus (35) follows. To show (37), we note that replacing the minimum on the right hand side of (29) by the value at $\theta = 0$ and using that $f = \mathsf{f}_\theta$ for $\theta \in B_q(\delta)$ yields

$$\sup_{\theta \in B_q(\delta)} \frac{1}{n} E_\theta |X(\hat{\theta}^{ES} - \theta)|_2^2 \le \sup_{\theta \in B_q(\delta)} \|\mathsf{f}_\theta\|^2 + \frac{C\sigma^2}{n} \le \delta^2 + \frac{C\sigma^2}{n} \qquad (39)$$

where we have used that $\|\mathsf{f}_\theta\| \le |\theta|_1 \max_j \|f_j\| \le \delta$ for $\theta \in B_q(\delta)$ if $0 \le q \le 1$. Finally, (37) is straightforward in view of the last display, (35), and (31). □

The rates on ℓ_q-balls with $0 < q \le 1$ follow from (37) by setting there $s = M$ (then $B_0(s) = \mathbb{R}^M$). Note also that the upper bounds of Theorem 7.3 are optimal in a minimax sense, cf. Section 8 below. Theorem 7.3 shows that the estimator $\hat{\theta}^{ES}$ attains minimax rates over all $B_0(s) \cap B_q(\delta)$ with $0 < q \le 1$ (cf. (37)), adaptively to s, δ, q, and, in addition, $\hat{\theta}^{ES}$ achieves optimal rates of aggregation on these sets (cf. (35)). For $1 < q \le 2$ we only show that $\hat{\theta}^{ES}$ accomplishes the first task – minimax rates over ℓ_q-balls, under the boundedness assumption on the maximal eigenvalue $\lambda_{\max}(X^T X/n)$ of matrix $X^T X/n$.

Theorem 7.4. *Assume that $\lambda_{\max}(X^T X/n) \le L^2$, and $1 < q \le 2$. Let $\hat{\theta}^{ES}$ denote the vector of weights of the exponential screening estimator with $\beta = 8\sigma^2$. If the model is linear, $y = X\theta + \xi$, then for any $n \ge 1$, $M \ge 2$, $1 \le s \le M$, and $\delta > 0$ we have*

$$\sup_{\theta \in B_0(s) \cap B_q(\delta)} \frac{1}{n} E_\theta |X(\hat{\theta}^{ES} - \theta)|_2^2 \le C\bar{\psi}_{n,M}(L\delta, s, q) \qquad (40)$$

$C > 0$ is an absolute constant.

Proof. In this case, we get the analog (39) with $L\delta$ instead of δ since $\|\mathsf{f}_\theta\|^2 \le L^2|\theta|_2^2 \le (L\delta)^2$ for $\theta \in B_q(\delta)$, $1 < q \le 2$. To complete the proof, it suffices to show that, for any $\theta \in \mathbb{R}^M$ there exists $\bar{\theta} \in \mathbb{R}^M$ such that $|\bar{\theta}|_0 \le m$ and

$$|X(\bar{\theta} - \theta)|_2^2/n = \|\mathsf{f}_{\bar{\theta}} - \mathsf{f}_\theta\|^2 \le L^2 |\theta|_q^2 m^{1-2/q}. \qquad (41)$$

This replaces Lemma 7.2 when the model is linear, i.e., $f = \mathsf{f}_\theta$. Given (41), the argument follows the same lines as in the proof of (37). To prove (41), take $\bar{\theta} = \theta_J$ where $J \subseteq \{1, \dots, M\}$ is the set of indices corresponding to the m largest in absolute value components of θ. Then $\|\mathsf{f}_{\bar{\theta}} - \mathsf{f}_\theta\|^2 \le \lambda_{\max}(X^T X/n)|\bar{\theta} - \theta|_2^2 \le L^2 |\theta_{J^c}|_2^2$. Using (34) we deduce (41) from the chain of inequalities

$$|\theta_{J^c}|_2^2 = \sum_{j \ge m+1} |\theta|_{(j)}^2 \le |\theta|_{(m)}^{2-q} \sum_{j \ge m+1} |\theta|_{(j)}^q \le \left(\frac{|\theta|_q}{m^{1/q}}\right)^{2-q} |\theta|_q^q = |\theta|_q^2 m^{1-2/q}.$$

□

8. Minimax lower bounds

The rates for the minimax risk on the intersection of ℓ_0- and ℓ_q-balls obtained in the previous section are optimal as shows the next theorem.

Theorem 8.1. *Let $M \geq 1, n \geq 1, 1 \leq s \leq M$, $M \leq n$, and $\delta \geq c_* \sigma/\sqrt{n}$ for some constant $c_* > 0$. Let either $\Theta_{\delta,s,q} = B_0(s) \cap B_q(\delta)$ and $0 < q \leq 1$ or $\Theta_{\delta,s,q} = \delta\Lambda^M \cap B_0(s)$ and $q = 1$. Then there exists a dictionary $f_1, \dots, f_M$ with $\max_{1\leq j\leq M} \|f_j\| \leq 1$ such that*

$$\inf_{\hat{f}} \sup_{\theta\in\Theta_{\delta,s,q}} E_\theta \|\hat{f} - \mathsf{f}_\theta\|^2 \geq C\bar{\psi}_{n,M}(\delta, s, q)$$

where $C > 0$ is a constant independent of n, M, δ, s.

The proof of Theorem 8.1 is given in [31]($q = 1$), and in [38]($0 < q < 1$). These papers also describe the additional conditions on the matrix X, for which the lower bound holds when not necessarily $M \leq n$. Some bounds on the sparse eigenvalues of $X^T X/n$ are then required. Minimax lower bounds for the case $\delta = \infty$, corresponding to the class $B_0(s)$ are studied in [6, 2, 28, 36]. The other extreme case $s = M$, corresponding to the class $B_q(\delta)$, $0 < q \leq 1$, is studied in [28] under specific asymptotics on n, M, δ that do not provide the general form of $\bar{\psi}_{n,M}(\delta, M, q)$; related results are given in [40] for different risk. For the diagonal case when $X^T X/n$ is the identity matrix and $M = n$, upper and lower bounds under some specific asymptotics separately on $B_0(s)$ and on $B_q(\delta)$ are proved in [16, 17, 1]; they are extended to non-asymptotic bounds in [4, 21].

Remark 8.2. If we assume that $\max_{1\leq j\leq M} \|f_j\| \leq L$ instead of $\max_{1\leq j\leq M} \|f_j\| \leq 1$, the lower bound of Theorem 8.1 remains valid with δ replaced by $L\delta$. This remark concerns also the upper bounds of Theorem 7.3.

9. Nonparametric estimation and group sparsity

Consider now Setting 2 of the Introduction (nonparametric regression). Assume that $f \in \mathcal{F}_{\beta,L}$ and the class $\mathcal{F}_{\beta,L}$ is such that assumption (7) holds. Theorem 5.1 with $M = n$, and this assumption imply that

$$E\|\hat{f}^{ES} - f\|^2 \leq C(\beta, L)^2 k^{-2\beta} + \frac{C\sigma^2}{n} k \log n \tag{42}$$

for any $k \leq n$. Minimizing this bound in k and taking the suprema we obtain

$$\sup_{f\in\mathcal{F}_{\beta,L}} E\|\hat{f}^{ES} - f\|^2 \leq C'(\beta, L) \left(\frac{\log n}{n}\right)^{-2\beta/(2\beta+1)} \tag{43}$$

where $C'(\beta, L)$ is a constant depending only on β and L. Thus, the estimator $\hat{f}^{ES}$ attains (adaptively in β, L) the rate $n^{-2\beta/(2\beta+1)}$ up to a logarithmic factor. Note

that $n^{-2\beta/(2\beta+1)}$ is the optimal rate of convergence of the squared risk for major classes $\mathcal{F}_{\beta,L}$ satisfying assumption (7) [34]. The extra logarithmic factor in (43) can be avoided by using, instead of $\hat{f}^{ES}$, a group exponentially weighted aggregate, which is of independent interest and is defined as follows.

Let $B_1, \dots, B_K$ be given subsets of $\{1, \dots, M\}$ called the groups. Consider $\theta \in \mathbb{R}^M$ such that $\mathrm{supp}(\theta) \subseteq B = \bigcup_{k=1}^K B_k$ where $\mathrm{supp}(\theta)$ is the set of indices of non-zero components of θ. For any such θ, we denote by $J(\theta)$ a subset of $\{1, \dots, K\}$ of smallest cardinality among all J such that $\mathrm{supp}(\theta) \subseteq B_J = \bigcup_{k \in J} B_k$. Define

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

where $|\cdot|$ denotes the cardinality. For any subset J of $\{1, \dots, K\}$, denote by p^J the sparsity pattern in $\mathcal{P}$ with coordinates

$$p_j^J = \begin{cases} 1 & \text{if } j \in \bigcup_{k \in J} B_k, \\ 0 & \text{otherwise,} \end{cases}$$

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

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

To each sparsity pattern $p \in \mathcal{P}_{\mathrm{G}}$, we assign a least squares estimator $\hat{\theta}_p$ constructed from the first subsample $y^{(1)}$, cf. (25). Define the following prior probability distribution on $\mathcal{P}_{\mathrm{G}}$:

$$\pi_p^{\mathrm{G}} = \left[\binom{K}{|J|} e^{|J|} H' \right]^{-1}, \quad \forall\, p = p^J, \quad J \subseteq \{1, \dots, K\},$$

where $H' = \sum_{k=0}^K e^{-k}$, and consider the exponentially weighted aggregate

$$\hat{f}^{\mathrm{G}} = \sum_{p \in \mathcal{P}_{\mathrm{G}}} \hat{\theta}_p^{\mathrm{G}} \mathsf{f}_{\hat{\theta}_p} , \tag{44}$$

where $\hat{\theta}^{\mathrm{G}} = (\hat{\theta}_p^{\mathrm{G}},\ p \in \mathcal{P}_{\mathrm{G}})$ is a vector with components

$$\hat{\theta}_p^{\mathrm{G}} = \frac{\exp(-\eta \hat{r}_p/\beta)\pi_p^{\mathrm{G}}}{\sum_{p' \in \mathcal{P}} \exp(-\eta \hat{r}_{p'}/\beta)\pi_{p'}^{\mathrm{G}}}, \quad \forall p \in \mathcal{P}_{\mathrm{G}}.$$

Theorem 9.1. [32] *Let $\beta = 8\sigma^2$. Then for any f, and any integers $n \geq 1$, $M \geq 1$, we have*

$$E\|\hat{f}^{\mathrm{G}} - f\|^2 \leq \inf_{\substack{\theta \in \mathbb{R}^M: \\ \mathrm{supp}(\theta) \subseteq B}} \left\{ \|\mathsf{f}_\theta - f\|^2 + \frac{C\sigma^2}{n} \left(B(\theta) + g(\theta) \log\left(\frac{eK}{g(\theta)} \right) + 1 \right) \right\} \tag{45}$$

where $C > 0$ is an absolute constant.

Remark that Theorem 9.1 is stated for arbitrary groups B_j. They can overlap and not necessarily cover the whole set $\{1, \dots, M\}$.

Using (45), one can prove that the aggregate $\tilde{f}^{\mathrm{G}}$ achieves the optimal rate of convergence under the group sparsity setting [32]. Note that upper bounds for the risk of the Group Lasso estimators in [20, 26] as well as in the earlier papers cited therein depart from this optimal rate at least by a logarithmic factor. Moreover, they are obtained under strong assumptions on the dictionary $\{f_1, \dots, f_M\}$ such as restricted isometry or restricted eigenvalue type conditions, while (45) is valid under no assumption on the dictionary.

We now apply Theorem 9.1 for Setting 2 of the Introduction. Let $M = n$, and let all groups B_j be of the same size $T = \lceil (\log n)^2 \rceil$ and form a partition of $\{1, \dots, n\}$, so that, w.l.o.g., $n = KT$. Let the class $\mathcal{F}_{\beta,L}$ be such that (7) holds. Denote by $\tilde{f}^{\mathrm{G}}$ the estimator (44) with this choice of parameters. In particular, functions f_j are those from (7). Fix any $f \in \mathcal{F}_{\beta,L}$. Set $k_* = \lceil n^{1/(2\beta+1)} \rceil$, and let θ^* be the vector in $\mathbb{R}^n$ whose first k_* components are the values $\theta^*_1, \dots, \theta^*_{k_*}$ from (7), and other components are 0. Then $g(\theta^*) \le k_*/T + 1$, and $B(\theta^*) \le k_* + T$. Plugging these values into the right hand side of (45) and using (7) with $k = k_*$, we find

$$E\|\tilde{f}^{\mathrm{G}} - f\|^2 \le C(\beta, L)^2 k_*^{-2\beta} + \frac{C\sigma^2(k_* + T)}{n}\left(1 + \frac{1}{T}\log\left(\frac{en}{k_* + T}\right)\right), \qquad (46)$$

which immediately implies the next corollary.

Corollary 9.2. *For any class of functions $\mathcal{F}_{\beta,L}$ such that (7) holds, we have*

$$\sup_{f \in \mathcal{F}_{\beta,L}} E\|\tilde{f}^{\mathrm{G}} - f\|^2 \le c(\beta, L, \sigma^2) n^{-2\beta/(2\beta+1)} \qquad (47)$$

where $c(\beta, L, \sigma^2)$ is a constant depending only on β, L, and σ^2.

Remark that (43) and (47) are adaptive results. Indeed, the estimators $\hat{f}^{ES}$ and $\tilde{f}^{\mathrm{G}}$ do not depend on the parameters β and L, and satisfy these upper bounds simultaneously for all classes $\mathcal{F}_{\beta,L}$ such that (7) holds.

10. Universal aggregation

Along with the three main types of aggregation (MS, C, L) described in the Introduction, two other natural examples are of interest: the s-sparse aggregation (L_s) [6], and the convex s-sparse aggregation (C_s) [25]. As summarized in Table 1, the sets Θ for the five types of aggregation are either ℓ_0-balls or intersections of ℓ_0-balls with the simplex Λ^M.

Problem	Θ	Description of the oracle
(MS)	$\Theta_{(\mathrm{MS})} = B_0(1) \cap \Lambda^M$	Best in dictionary
(C)	$\Theta_{(\mathrm{C})} = \Lambda^M$	Best convex combination
(L)	$\Theta_{(\mathrm{L})} = \mathbb{R}^M = B_0(M)$	Best linear combination
(L_s)	$\Theta_{(\mathrm{L}_s)} = B_0(s)$	Best s-sparse linear combination
(C_s)	$\Theta_{(\mathrm{C}_s)} = B_0(s) \cap \Lambda^M$	Best s-sparse convex combination

Table 1.

The next theorem follows from (30), (35) with $q = \delta = 1$ and the inclusion $\Lambda^M \subset B_1(1)$.

Theorem 10.1. [31] *Assume that* $\max_{1\le j\le M} \|f_j\| \le 1$. *Then for any* $M \ge 2, n \ge 1, 1 \le s \le M$, *and* $\Theta \in \{\Theta_{(\mathrm{MS})}, \Theta_{(\mathrm{C})}, \Theta_{(\mathrm{L})}, \Theta_{(\mathrm{L}_s)}, \Theta_{(\mathrm{C}_s)}\}$ *the exponential screening estimator with* $\beta = 8\sigma^2$ *satisfies the following oracle inequality*

$$E\|\hat{f}^{ES} - f\|^2 \le \min_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 + C\psi_{n,M}(\Theta)\,.$$

Here, $C > 0$ *is an absolute constant,*

$$\psi_{n,M}(\Theta) = \psi^*_{n,M}(\Theta) \wedge \frac{\sigma^2 R}{n}$$

and $\psi^*_{n,M}(\Theta)$ *is given in Table 2.*

Theorem 8.1 implies that the rates $\psi_{n,M}(\Theta)$ given in Theorem 10.1 are optimal rates of aggregation for the corresponding classes Θ. Indeed, to show (12) with the same rates, it suffices to use the lower bounds for the minimax risk, since obviously

$$\inf_{\hat f} \sup_{f} \left(E\|\hat f - f\|^2 - \min_{\theta\in\Theta} \|\mathsf{f}_\theta - f\|^2 \right) \ge \inf_{\hat f} \sup_{\theta\in\Theta} E_\theta \|\hat f - \mathsf{f}_\theta\|^2. \tag{48}$$

On the other hand, the sets Θ in Theorem 10.1 are either ℓ_0-balls or intersections of ℓ_0-balls with the ℓ_1-simplex Λ^M, and the lower bounds for the minimax risk on these sets are available from Theorem 8.1.

If assumption $\max_{1\le j\le M} \|f_j\| \le 1$ in Theorem 10.1 is replaced by $\max_{1\le j\le M} \|f_j\| \le L$, the rates in Table 2 remain valid with the only difference that $\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{M\sigma}{\sqrt{n}}\right)}$ should be replaced by $L\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{M\sigma}{L\sqrt{n}}\right)}$.

We see that the problem of aggregation is closely related to that of minimax estimation on the intersection of ℓ_0- and ℓ_1-balls. Indeed, for both problems upper bounds for the risk and for the excess risk are attained by one and the same estimator, which is the exponential screening estimator. Furthermore, the optimal

Problem	$\psi^*_{n,M}(\Theta)$
(MS)	$\frac{\sigma^2 \log M}{n}$
(C)	$\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{M\sigma}{\sqrt{n}}\right)}$
(L)	$\frac{\sigma^2 R}{n}$
(L_s)	$\frac{\sigma^2 s}{n} \log\left(\frac{eM}{s}\right)$
(C_s)	$\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{M\sigma}{\sqrt{n}}\right)} \wedge \frac{\sigma^2 s}{n} \log\left(\frac{eM}{s}\right)$

Table 2.

rates of aggregation in Theorem 10.1 are similar to the minimax rates on the intersection of the corresponding ℓ_0- and ℓ_1-balls (cf. Section 7).

Using (48), the upper bounds (30), (35), and Theorem 8.1 we also find that the estimator $\hat{f}^{ES}$ attains the optimal rates of ℓ_q-aggregation:

Theorem 10.2. *Let the assumptions of Theorems 8.1 and 10.1 be satisfied, and let $\delta \geq c_*$ where $c_* > 0$ is a constant independent of n, M. Then the estimator $\hat{f}^{ES}$ with $\beta = 8\sigma^2$ is an optimal aggregate for the classes $\Theta = B_q(\delta)$, and $\Theta = B_q(\delta) \cap B_0(s)$. The optimal rates of aggregation for these classes are, respectively, $\psi_{n,M}(B_q(\delta))$, and $\bar{\psi}_{n,M}(\delta, s, q)$.*

In summary, the exponential screening estimator enjoys the property of *universal aggregation*, i.e., it attains optimal rates of aggregation simultaneously on all the classes Θ considered in this section.

References

[1] Abramovich, F., Benjamini, Y., Donoho, D. L., Johnstone, I. M. . Adapting to Unknown Sparsity by Controlling the False Discovery Rate. *Ann. Statist.* **34** (2006), 584–653.

[2] Abramovich, F., Grinshtein, V. MAP Model Selection in Gaussian Regression. *Electronic J. of Statistics* **4** (2010), 932–949.

[3] Alquier, P., Lounici, K. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic J. of Statistics* **5** (2011), 127–145.

[4] Birgé, L., Massart, P. Gaussian model selection. *J. Eur. Math. Soc.* **3** (2001), 203–268.

[5] Bunea, F., Tsybakov, A. B., Wegkamp, M. Aggregation for Regression Learning, 2004. `arXiv:math.ST/0410214`

[6] Bunea, F., Tsybakov, A. B., Wegkamp, M. Aggregation for Gaussian Regression. *Annals of Statistics* **35** (2007), 1674–1697.

[7] Catoni, O. *Statistical Learning Theory and Stochastic Optimization.* Saint-Flour Summer School in Probability XXXI, 2001. Lecture Notes in Mathematics 1851. Springer, Berlin, 2004.

[8] Cesa-Bianchi, N., Lugosi, G. *Prediction, Learning, and Games.* Cambridge Univ. Press, 2006.

[9] Dai, D, Rigollet, P., Zhang, T. Deviation Optimal Learning using Greedy Q-aggregation. *Annals of Statistics* **40** (2012), 1878-1905.

[10] Dai, D, Rigollet, P., Xia, L., Zhang, T. Aggregation of Affine Estimators. *Electronic J. of Statistics* **8** 2014, 302-327.

[11] Dalalyan, A. S., Salmon, J. Sharp Oracle Inequalities for Aggregation of Affine Estimators. *Annals of Statistics* **40** (2012), 2327-2355.

[12] Dalalyan, A., Tsybakov, A. B. Aggregation by Exponential Weighting and Sharp Oracle Inequalities. In *Proc. COLT 2007*, Lecture Notes in Artificial Intelligence 4539, 97–111. Springer, Berlin-Heidelberg, 2007.

[13] Dalalyan, A., Tsybakov, A. B. Aggregation by Exponential Weighting, Sharp PAC-Bayesian Bounds and Sparsity. *Machine Learning* **72** (2008), 39–61.

[14] Dalalyan, A., Tsybakov, A. B. Mirror Averaging with Sparsity Priors. *Bernoulli* **18** (2012), 914–944.

[15] Dalalyan, A., Tsybakov, A.B. Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences* **78** (2012), 1423-1443.

[16] Donoho, D. L., Johnstone, I. M., Hoch, J. C., Stern, A. S. Maximum Entropy and the Nearly Black Object. *J. Roy. Statist. Soc. Ser. B* **54** (1992), 41–81.

[17] Donoho, D. L., Johnstone, I. M. Minimax Risk Over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** (1994), 277–303.

[18] Gerchinovitz, S. Prediction of Individual Sequences and Prediction in the Statistical Framework : some Links around Sparse Regression and Aggregation Techniques. PhD thesis, Université Paris Sud, 2011.

[19] Härdle, W., Kerkyacharian, G., Picard, D.,Tsybakov, A. *Wavelets, Approximation, and Statistical Applications.* Lecture Notes in Statistics 129. Springer, NY, 1998.

[20] Huang, J., Zhang, T. The benefit of group sparsity. *Annals of Statistics* **38** (2010), 1978–2004.

[21] Johnstone, I. *Gaussian Estimation: Sequence and Wavelet Models.* Draft of a book. `http://www-stat.stanford.edu/~imj/GE06-11-13.pdf`, 2011.

[22] Koltchinskii, V., Lounici, K., Tsybakov, A. B. Nuclear Norm Penalization and Optimal Rates for Noisy Low Rank Matrix Completion. *Annals of Statistics* **39** (2011), 2302–2329.

[23] Lecué, G., Rigollet, P. Optimal Learning with Q-aggregation. *Annals of Statistics* **42** (2014), 211-224.

[24] Leung, G., Barron, A. R. Information Theory and Mixing Least-squares Regressions. *IEEE Trans. Inform. Theory* **52** (2006), 3396–3410.

[25] Lounici, K. Generalized Mirror Averaging and D-convex Aggregation. *Mathematical Methods of Statistics* **16** (2007), 246–259.

[26] Lounici, K., Pontil, M., Tsybakov, A. B., van de Geer, S. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** (2011), 2164–2204.

[27] Nemirovski, A. *Topics in Non-parametric Statistics.* Saint-Flour Summer School in Probability XXVIII, 1998. Lecture Notes in Mathematics 1738. Springer, NY, 2000.

[28] Raskutti, G., Wainwright, M. J.,Yu, B. Minimax Rates of Estimation for High-dimensional Linear Regression over l_q-balls. *IEEE Trans. Inform. Th.* **57** (2011), 6976–6994.

[29] Rigollet, P. Kullback-Leibler Aggregation and Misspecified Generalized Linear Models. *Annals of Statistics* **40** (2012), 639-665.

[30] Rigollet, P., Tsybakov, A. B. Linear and Convex Aggregation of Density Estimators. *Mathematical Methods of Statistics* **16** (2007), 260–280.

[31] Rigollet, P. , Tsybakov, A. B. Exponential Screening and Optimal Rates of Sparse Estimation. *Annals of Statistics* **39** (2011), 731-771.

[32] Rigollet, P. , Tsybakov, A. B. Sparse Estimation by Exponential Weighting. *Statistical Science* **27** (2012), 558-575.

[33] Tsybakov, A. B. Optimal Rates of Aggregation. In *Proc. COLT 2003*, Lecture Notes in Computer Science 2777, 303–313. Springer, NY, 2003.

[34] Tsybakov, A. B. *Introduction to Nonparametric Estimation.* Springer, NY, 2009.

[35] Tsybakov, A. B. *Aggregation and High-dimensional Statistics.* Lecture notes of Saint-Flour Summer School in Probability, July 2013.

[36] Verzelen, N. Minimax Risks for Sparse Regressions: Ultra-high Dimensional Phenomenons. *Electron. J. Stat.* **6** (2012), 38–90.

[37] Vovk, V. Aggregating Strategies. In *Proc. 3rd Annual Workshop on Computational Learning Theory*, 372–383. Morgan Kaufmann, San Mateo, CA, 1990.

[38] Wang, Z., Paterlini, S., Gao, F., Yang, Y. Adaptive Minimax Estimation over Sparse l_q-hulls, 2011. `arXiv:1108.1961`

[39] Yang, Y. Aggregating Regression Procedures to Improve Performance. *Bernoulli* **10** (2004), 25– 47.

[40] Ye, F., Zhang, C.-H. Rate Minimaxity of the Lasso and Dantzig Selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research* **11** (2010), 3519–3540.

Laboratoire de Statistique, CREST-ENSAE, 3, av. Pierre Larousse 92245 Malakoff, France

E-mail: alexandre.tsybakov@ensae.fr