AGGREGATION AND HIGH-DIMENSIONAL STATISTICS
(preliminary notes of Saint-Flour lectures, July 8-20, 2013)

Alexandre B. Tsybakov (CREST-ENSAE)

March 23, 2014

1 Introduction

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, the aggregation techniques turn out to be very useful in taking advantage of sparsity. On the other hand, aggregation is a general technique of producing 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 computed is several possible ways. 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.

To be able to demonstrate the main ideas without excessive technicalities, throughout this course we will deal with a simple model, namely the Gaussian regression model with fixed design. Suppose that we observe $\{(Y_i, X_i)\}_{i=1}^n$ 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 the random errors ξ_i are i.i.d. Gaussian with mean zero and variance σ^2, $\xi_i \sim \mathcal{N}(0, \sigma^2)$.

The overall goal is to construct an estimator $\hat{f}$ for f based on the observations $\{(Y_i, X_i)\}_{i=1}^n$. To measure how good $\hat{f}$ is, 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 risk of estimator $\hat{f}$ as $E\|\hat{f} - f\|^2$. 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

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

Functions f_θ are thought to be approximations of the unknown f. Assuming the dictionary $\{f_1, \ldots, f_M\}$ to be rich enough and M sufficiently large, these approximations can be satisfactory. Therefore, the estimation of f may be reduced to estimating θ_j, leading an estimator

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

where $\hat{\theta}_j$ are suitable estimators of θ_j. Then the aim is to mimimize the risk by choosing an optimal $\hat{\theta}_j$. However, depending on the assumptions we make about the dictionary, the set Θ and f, we are lead to different optimality properties. We introduce below three scenarios and discuss how these assumptions influnce the construction of the estimators and the optimality framework.

1.1 Scenario 1: Linear Regression and Sparsity

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

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

Then we are in the usual linear regressoin framework, and the observations can be written in the following form

$$y = X\theta^* + \xi,$$

where

$$y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}, \qquad X = \begin{pmatrix} f_1(X_1) & \cdots & f_M(X_1) \\ \vdots & \vdots & \vdots \\ f_1(X_n) & \cdots & f_M(X_n) \end{pmatrix}. \tag{3}$$

Estimation of f is now reduced to estimation of θ^*. Classical theory of linear regression deals with cases where $n \le M$, which is a necessary condition of identifiability of θ^* when we only know that $\theta^* \in \mathbb{R}^M$. However, in recent years there is an increasing applied interest in the 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 on θ^*. It consists in restricting the parameter θ^* to the class $\Theta = B_0(s)$ where $B_0(s)$ is the ℓ_0-ball in $\mathbb{R}^M$:

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

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-sparisty restriction, estimation with reasonable accuracy is possible. We may ask ourselves the following question.

Question 1. *What is the optimal way to estimate* θ^* if we know that $\theta^* \in B_0(s)$*?*

Let $\hat{\theta}$ be an estimator of θ^*. The corresponding estimator of f is then

$$\hat{f} = f_{\hat{\theta}} = \sum_{j=1}^{M} \hat{\theta}_j f_j$$

and the squared risk defined above takes the form

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

This is known under the name of *prediction risk* for linear regression. The optimality is usually defined in a minimax sense. An estimator $\hat{\theta}$ is called optimal 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:

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

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

where C and c are positive constants independent of n, M, s, and $\inf_T$ denotes the minimum over all estimators of θ^* based on the sample $\{(Y_i, X_i)\}_{i=1}^n$. This 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 $B_0(s)$. To summarize, our main goal in this scenario 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 with $0 < q \le \infty$. This problem, in its simplest version where $X^T X/n$ is the identity (the Gaussian sequence model) and with asymptotic point of view, has been in the focus of statistical literature from the 1990ies, with the main developments due to Donoho and Johnstone. We are interested here in a more general linear regression setting and we deal with non-asymptotic minimax optimality.

1.2 Scenario 2: Nonparametric Regression

Let $f \in \mathcal{F}_{\beta,L}$ where $\mathcal{F}_{\beta,L}$, typically, is a class of smooth functions parametrized by $\beta > 0$ and $L > 0$. Roughly speaking, parameter β is the number of derivatives of f that are assumed bounded in some norm by constant L. In this scenario, it is usually assumed that the dictionary $\{f_1, \dots, f_M\}$ is composed of the first M functions of some orthonormal basis. For example, it can be the Fourier or wavelet basis. A key assumption in the nonparametric regression setting is that the true function f can be approximated by a linear combination of the basis functions. It can be stated, for example, in the following form.

Let $f \in \mathcal{F}_{\beta,L}$*. Then, for all* $M = 1, 2, \dots$ *there exists* $\theta^* = \theta^*(f) \in \mathbb{R}^M$ *such that*

$$\left\| f - \sum_{j=1}^{M} \theta_j^* f_j \right\| \le CM^{-\beta}, \tag{7}$$

where C *is a constant depending only on* β, L*.*

Here, in general, $f_{\theta^*} \stackrel{\text{def}}{=} \sum_{j=1}^{M} \theta_j^* f_j \ne f$, which is in contrast with the linear regression setting. Like in the linear regression case, we are interested in optimal estimation of f.

Question 2. *What is the minimax optimal estimator of f on the class $\mathcal{F}_{\beta,L}$?*

As before, a minimax optimal estimator $\hat{f}$ is the one that satisfies

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

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

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

Question 3. *How to construct an adaptive estimation procedure?*

An *adaptive estimator* is an estimator $\hat{f}$ which is independent of β and L and satisfies (8) with optimal rate of convergence $\psi_{n,\beta}$ for all pairs (β,L) in a wide range of values.

1.3 Scenario 3: Aggregation of estimators

The general mathematical framework of aggregation is introduced by Nemirovski in his Saint-Flour lectures in 1998 (published as Nemirovski (2000)). Nemirovskii (2000) outlined three problems: model selection type aggregation, convex aggregation, and linear aggregation.

More generally, the problem of aggregation is stated as follows. Suppose that we are given a collection of preliminary 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 the *aggregate*, which is approximately at least as good as the best linear combination $f_\theta=\sum_{j=1}^M\theta_j\hat{f}_j$ restricted to $\theta\in\Theta$. The best linear combination is defined as the one that solves the problem

$$\min_{\theta\in\Theta} E\|f-f_\theta\|^2$$

minimizing the squared risk. Unlike in the previous scenarios, here f_θ is a random function depending on the data. In contrast to those scenaria, we *do not assume* that $\|f-f_\theta\|$ is zero or small (see (2), (7)); it may happen that all f_θ for some Θ are very far from the true f. So, the choice of Θ is important for aggregation problems. Some examples of Θ are listed below.

1. *L-aggregation (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 $\hat{f}_1,\dots,\hat{f}_M$.

2. *C-aggregation (Convex aggregation):* Θ is the simplex

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

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

3. *MS-aggregation (Model Selection type aggregation):* $\Theta=\{e_1,\dots,e_M\}$ where e_i are the canonical basis vectors in $\mathbb{R}^M$. The aim of MS-aggregation is to construct an estimator $\tilde{f}$, which is approximately as good as the best among the initial estimators $\hat{f}_1,\dots,\hat{f}_M$.

4. *s-sparse aggregation:* $\Theta = B_0(s) \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^M : |\theta|_0 \le s\}$ where $s \in \{1, \dots, M\}$.

5. *L_q-aggregation:* $\Theta = B_q(\tau) \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^M : |\theta|_q \le \tau\}$ where $|\theta|_q = (\sum_{j=1}^M |\theta_j|^q)^{\frac{1}{q}}$ is the usual ℓ_q-norm.

Other types of aggregation will be discussed below as well. Note that for linear, convex and MS aggregation the sets Θ can be expressed as intersections of ℓ_0 and ℓ_1 balls. Indeed, for linear aggregation, $\Theta = \mathbb{R}^M = B_0(M)$, where $B_0(M)$ is the ℓ_0-ball of radius M. For convex aggregation, the simplex is included into $B_1^+(1)$ – an intersection of the ℓ_1-ball $B_1(1)$ with the cone of positive coordinates. For MS-aggregation, $\Theta = \{e_1, \dots, e_M\} = B_0(1) \cap B_1^+(1)$.

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

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

Based on the excess risk, we can introduce the concept of minimax optimality for aggregation. 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

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

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

Here, $\inf_{\hat{f}}$ is the minimum over all estimators, C and c are positive constants independent of n and M, and $\sup_{\hat{f}_1, \dots, \hat{f}_M}$, $\sup_f$ are the suprema over wide classes of preliminary estimators and functions f. In some cases, these will be all possible estimators and all possible f with no restriction; in other cases it will suffice to consider classes of $\hat{f}_1, \dots, \hat{f}_M$ and f satisfying a boundedness assumption in the empirical norm $\|\cdot\|$. If (11) and (12) hold for some sequence $\psi_{n,M}(\Theta)$, this sequence is called an *optimal rate of aggregation for the class* Θ. The questions arising in this context are as follows.

Question 4. *How to construct an optimal aggregate $\tilde{f}$ for a given class Θ?*

Question 5. *Is it possible to construct a* **universal aggregate**, *i.e., an aggregate which is optimal simultaneously for a large scale of classes Θ?*

The last question is of the same nature as Question 3 concerning adaptive nonparametric estimation.

Inequalities (11) and (12) establish upper and lower bounds for the minimax risk, respectively. The upper bounds (11) can be equivalently written in the form of *oracle inequalities*

$$E\|\tilde{f} - f\|^2 \le \inf_{\theta \in \Theta} E\|f_\theta - f\|^2 + C\psi_{n,M}(\Theta), \quad \forall\ \hat{f}_1, \dots, \hat{f}_M, f, \tag{13}$$

which say that the risk of the suggested aggregate $\tilde{f}$ is at least as good as the risk of the unknown *oracle* θ^* minimizing $E\|f_\theta - f\|^2$, up to a "small" remainder term of the order $\psi_{n,M}(\Theta)$ (a "price to pay for aggregation"). Lower bounds (12) say that this is the minimal price; the remainder term cannot be of a smaller order whatever is the aggregate. For the sparsity classes, for example, $\Theta = B_0(s)$, the rate $\psi_{n,M}(\Theta)$ is a function of s; the corresponding oracle inequalities are called *sparsity oracle inequalities*.

1.4 Outline

The main message of this course is that there are methods that solve problems described in Sections 1.1, 1.2, and 1.3 simultaneously. We will consider methods like the BIC, the Lasso, and the exponential weighting, provide oracle inequalities and discuss lower bounds for the three above scenarios in a unified framework. We will establish the optimal rates of aggregation. Anticipating, for the main types of aggregation they are given in the following table.

Problem	$\psi_{n,M}(\Theta)$
MS-aggregation	$\frac{\sigma^2 R}{n} \wedge \frac{\sigma^2 \log M}{n}$
C-aggregation	$\frac{\sigma^2 R}{n} \wedge \sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{M\sigma}{\sqrt{n}}\right)}$
L-aggregation	$\frac{\sigma^2 R}{n}$

Table 1.

We will also show that the technique of exponential weighting achieves universal aggregation.

2 From aggregation of estimators 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. The analysis is more involved if both $\hat{\theta}_j$ and the preliminary estimators $\hat{f}_j$ are constructed from the same sample $\{(Y_i, X_i)\}_{i=1}^n$. To avoid this, the idea put forward by Nemirovski (2000) is to obtain two independent samples from the initial one by randomization (*sample cloning*). Then 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 $f_j = \hat{f}_j$, $j = 1, \dots, M$. A limitation is that this type of randomization only applies to Gaussian model with known variance. Nevertheless, the idea of two-step procedures carries over to models with i.i.d. observations where one can do direct sample splitting (see, e.g., Rigollet and Tsybakov (2007); Lecué (2011)). Thus, in many cases aggregation of estimators can be achieved by reduction to aggregation of deterministic functions. Along with this approach, one can aggregate estimators using the same observations for estimation and aggregation. While for general estimators this would clearly result in overfitting, the idea proved to be successful for certain types of estimators, first for projection estimators (Leung and Barron (2006)) and more recently for a more general class of linear (affine) estimators (Dalalyan and Salmon (2011)).

The procedure of *sample cloning* by randomization is based on the following elementary lemma.

Lemma 1. *Let $Y_i = f(X_i) + \xi_i$. Let ω_i be a standard normal random variable independent of ξ_i. Set*

$$Y_{i1} = Y_i + \sigma\omega_i,$$
$$Y_{i2} = Y_i - \sigma\omega_i.$$

Then we have

$$Y_{i1} = f(X_i) + \xi_{i1},$$
$$Y_{i2} = f(X_i) + \xi_{i2},$$

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, we obtain two independent Gaussian samples $D_1 = \{(Y_{i1}, X_i)\}_{i=1}^n$ and $D_2 = \{(Y_{i2}, X_i)\}_{i=1}^n$, where $Y_{ik} = f(X_i) + \xi_{ik}$, $k = 1, 2$. Both samples are of the same form as the original one $\{(Y_i, X_i)\}_{i=1}^n$, 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

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

Clearly, to obtain (14) it suffices to show that, for any fixed functions $f_1, \dots, f_M, f$ (possibly satisfying some mild assumptions), we have

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

where f_θ is a linear combination of $f_1, \dots, f_M$, 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. Then, the minimax framework modifies only in that the excess risk takes the form

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

(no expectation in the term $\inf_{\theta\in\Theta} \|f_\theta - f\|^2$). In this setting, an estimator $\tilde{f}$ is 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 $\hat{f}_j$'s are replaced by f_j's. The upper bound on the maximum excess risk is equivalent to the oracle inequality

$$E\|\tilde{f} - f\|^2 \le \inf_{\theta\in\Theta} \|f_\theta - f\|^2 + C\psi_{n,M}(\Theta), \quad \forall\ f_1, \dots, f_M, f. \tag{17}$$

Once such an oracle inequality is established, we can obtain upper bounds for the minimax risk in Scenarios 1 and 2 as simple corollaries. Indeed, those scenarios introduce additional strong restrictions on f, in particular, that the oracle risk $\inf_{\theta\in\Theta} \|f_\theta - f\|^2$ is either 0 (for Scenario 1) or admits a given bound, cf. (7) (for Scenario 2).

3 Least squares aggregation

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

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

and we define the LS aggregate as

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

We are going to show that this idea works for linear and convex aggregation but fails for MS-aggregation. Recall that we denote by X the matrix

$$X = \begin{pmatrix} f_1(X_1) & \cdots & f_M(X_1) \\ \vdots & \vdots & \vdots \\ f_1(X_n) & \cdots & f_M(X_n) \end{pmatrix}.$$

Proposition 2 (Linear aggregation)**.** *Let $\hat{\theta}^{LS} \stackrel{\text{def}}{=} \hat{\theta}^{LS}(\mathbb{R}^M)$ be a least squares estimator for the class $\Theta = \mathbb{R}^M$. Then for all $f, f_1, \dots, f_M$ we have*

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

where $R = \operatorname{Rank}(X)$.

Proof. In what follows, with a slight abuse of notation, 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, with the notation from (3), the model of observations (1) can be written as $y = f + \xi$. Also, $f_\theta = X\theta$ for all θ and, in particular, $f_{\hat{\theta}^{LS}} = X\hat{\theta}^{LS} = Ay$ where A is the orthogonal projector on $\operatorname{Im}(X)$. Since $y = f + \xi$ we have

$$\|f_{\hat{\theta}^{LS}} - f\|^2 = \|Ay - f\|^2 = \|A(f + \xi) - f\|^2,$$

which yileds

$$E\|f_{\hat{\theta}^{LS}} - f\|^2 = \|Af - f\|^2 + E\|A\xi\|^2.$$

Since A is the projector on $\operatorname{Im}(X)$,

$$\|Af - f\|^2 = \min_{v \in \operatorname{Im}(X)} \|v - f\|^2 = \min_{\theta \in \mathbb{R}^M} \|X\theta - f\|^2 = \min_{\theta \in \mathbb{R}^M} \|f_\theta - f\|^2.$$

On the other hand,

$$E\|A\xi\|^2 = \frac{\sigma^2}{n} \operatorname{Tr}(A) = \frac{\sigma^2 R}{n}$$

and the proposition follows.

We now turn to convex aggregation and consider the corresponding LS aggregate

$$\hat{\theta}^{LS}_{\text{conv}} = \operatorname*{argmin}_{\theta \in \Lambda^M} \|y - f_\theta\|^2$$

where Λ^M is the simplex.

Proposition 3 (Convex aggregation)**.** *For all f and all dictionaries $f_1, \dots, f_M$ such that $\|f_j\| \le L$, $j = 1, \dots, M$, we have*

$$E\|f_{\hat{\theta}^{LS}_{\text{conv}}} - f\|^2 \le \min_{\theta \in \Lambda^M} \|f_\theta - f\|^2 + 4\sigma L\sqrt{\frac{2\log M}{n}}\,.$$

Proof. Set for brevity $\tilde{f} = f_{\hat{\theta}^{LS}_{\text{conv}}}$. First, by a simple algebra, for any $g = f_\theta$ with $\theta \in \Lambda^M$, using that $\|y - \tilde{f}\|^2 \le \|y - g\|^2$ and $y = f + \xi$, we deduce that

$$\|\tilde{f} - f\|^2 \le \|f - g\|^2 + 2 < \tilde{f} - g, \xi >$$

where

$$< f, g > \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f(X_i)g(X_i)\,.$$

Thus,

$$E\|\tilde{f} - f\|^2 \le \|f - f_\theta\|^2 + 2E < \tilde{f} - f_\theta, \xi > .$$

Now,

$$E < \tilde{f} - f_\theta, \xi > \le E \max_{\theta' \in \Lambda^M} < f_{\theta'} - f_\theta, \xi > = E \max_{1 \le j \le M} < f_j - f_\theta, \xi > \,.$$

Note that

$$\begin{aligned}
\|f_\theta\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{M}\theta_j f_j(X_i)\right)^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M}\theta_j f_j^2(X_i) \quad (\text{since } \theta \in \Lambda^M) \\
&= \sum_{j=1}^{M}\theta_j\|f_j\|^2 \le L^2.
\end{aligned}$$

Therefore, $\|f_j - f_\theta\| \le 2L$. On the other hand,

$$\eta_j \stackrel{\text{def}}{=} < f_j - f_\theta, \xi > \sim \mathcal{N}(0, \bar{\sigma}^2)$$

where $\bar{\sigma}^2 = \sigma^2\|f_j - f_\theta\|^2/n$. Hence, using Lemma 28 we obtain that

$$\begin{aligned}
E \max_{1 \le j \le M} < f_j - f_\theta, \xi > &= E \max_{1 \le j \le M} \eta_j \le \bar{\sigma}\sqrt{2\log M} \\
&= \sigma\|f_j - f_\theta\|\sqrt{\frac{2\log M}{n}} \le 2L\sigma\sqrt{\frac{2\log M}{n}}\,,
\end{aligned}$$

and the proposition follows.

Now introduce a new aggregate that switches between linear and convex LS aggregates:

$$\tilde{f}^* \stackrel{\text{def}}{=} \begin{cases} f_{\hat{\theta}^{LS}} & \text{if } \frac{\sigma^2 R}{n} < 4L\sigma\sqrt{\frac{2\log M}{n}}\,, \\ f_{\hat{\theta}^{LS}_{\text{conv}}} & \text{if } \frac{\sigma^2 R}{n} \ge 4L\sigma\sqrt{\frac{2\log M}{n}}\,. \end{cases}$$

The following corollary is straightforward in view of Propositions 2 and 3. It allows to obtain for $\tilde{f}^*$ the fastest of the two rates.

Corollary 4 (Convex aggregation). *For all f and all dictionaries $f_1, \dots, f_M$ such that $\|f_j\| \le L$, $j = 1, \dots, M$, we have*

$$E\|\tilde{f}^* - f\|^2 \le \min_{\theta \in \Lambda^M} \|f_\theta - f\|^2 + \frac{\sigma^2 R}{n} \wedge 4\sigma L \sqrt{\frac{2 \log M}{n}}\,.$$

Note that, up to a minor logarithmic discrepancy, the aggregate $\tilde{f}^*$ achieves the target optimal rate of convex aggregation given in Table 1. However, for MS-aggregation the situation is different. In this case, Θ is a finite set and the least squares estimator of f is defined by

$$\hat{f}^{MS} = f_{\hat{j}} \quad \text{where} \quad \hat{j} = \operatorname{argmin}_{1 \le j \le M} \|y - f_j\|^2.$$

Repeating the argument of Proposition 3 we obtain the following oracle inequality.

Proposition 5 (MS-aggregation). *For all f and all $f_1, \dots, f_M$ such that $\|f_j\| \le L$, $j = 1, \dots, M$, we have*

$$E\|\hat{f}^{MS} - f\|^2 \le \min_{1 \le j \le M} \|f_j - f\|^2 + 4\sigma L \sqrt{\frac{2 \log M}{n}}\,.$$

We see that the optimal rate for MS-aggregation, which is of the order $(\log M)/n$ (cf. Table 1) is not achieved and the LS-aggregate $\hat{f}^{MS}$ exhibits much poorer behavior. This is not due to the techniques of the proof. In fact, the rate $\sqrt{(\log M)/n}$ given in Proposition 5 is the best that one can obtain for $\hat{f}^{MS}$. The following result shows that this defect is intrinsic not only for the least squares estimator but also for any method that selects only one function in the dictionary. 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 6 (Suboptimality of selectors). *Assume that*

$$(\sigma \vee 1)\sqrt{(\log M)/n} \le C_0 \tag{18}$$

for $0 < C_0 < 1$ small enough. Then, there exists a dictionary $\{f_1, \dots, f_M\}$ with $\|f_j\| \le 1$, $j = 1, \dots, M$, such that the following holds. For any selector $\hat{S}_n$, and in particular, for any selector based on penalized empirical risk minimization, there exist a regression function f such that $\|f\| \le 1$ and

$$E\|\hat{S}_n - f\|^2 \ge \min_{1 \le j \le M} \|f_j - f\|^2 + C_* \sigma \sqrt{\frac{\log M}{n}} \tag{19}$$

for some positive constant C_*.

It follows from the lower bound (19) that *selecting* one of the functions in a finite dictionary to solve the problem of model selection is suboptimal in the sense that it exhibits a too large remainder term, of the order $\sqrt{(\log M)/n}$. It turns out that we can do better if we take a *mixture*, that is a convex combination of the functions in the dictionary. We will see below that under a particular choice of weights in this convex combination, namely the *exponential weights*, one can achieve oracle inequalities with the optimal rate $(\log M)/n$.

Proof of Theorem 6. Consider a random matrix $\mathbb{X}$ of size $n \times M$ such that its elements $\mathbb{X}_{i,j}, i = 1, \ldots, n,\ j = 1, \ldots, M$ are i.i.d. Rademacher random variables, i.e., random variables taking values 1 and -1 with probability $1/2$. Moreover, assume that

$$\frac{2}{n} \log\left(1 + \frac{eM}{2}\right) < C_1 \,. \tag{20}$$

for some positive constant $C_1 < 1/2$. Note that (20) follows from (18) if C_0 is chosen small enough. Theorem 5.2 in Baraniuk et al (2008) [see also Subsection 5.2.1 in Rigollet and Tsybakov (2011)] implies that if (20) holds for C_1 small enough, then there exists a nonempty set $\mathcal{M}$ of matrices obtained as realizations of the matrix $\mathbb{X}$ that enjoy the following *weak restricted isometry* property. For any $X \in \mathcal{M}$, there exist constants $\bar{\kappa} \geq \underline{\kappa} > 0$, such that for any $\lambda \in \mathbb{R}^M$ with at most 2 nonzero coordinates,

$$\underline{\kappa}^2 |\lambda|_2^2 \leq \frac{|X\lambda|_2^2}{n} \leq \bar{\kappa}^2 |\lambda|_2^2 \,, \tag{21}$$

when (20) is satisfied. For $X \in \mathcal{M}$, let $\phi_1, \ldots, \phi_M$ be any functions on $\mathcal{X}$ satisfying

$$\phi_j(X_i) = x_{i,j} \,, \quad i = 1, \ldots, n,\ j = 1, \ldots, M \,,$$

where $x_{i,j}$ are the entries of X. Note that $\|\phi_j\| = 1$ since $x_{i,j} \in \{-1, 1\}$.

Fix $\tau > 0$ to be chosen later and set

$$f_j = \tau\,(1 + \alpha)\,\phi_j \,, \quad j = 1, \ldots, M,$$

where we set for brevity $\alpha = (\sigma/3)\sqrt{\frac{\log M}{\bar{\kappa}^2 n}}$. Moreover, consider the functions

$$\eta_j = \tau\alpha\phi_j \,, \quad j = 1, \ldots, M.$$

Using (18) we choose τ small enough to ensure that $\|\eta_j\| \leq 1$ and $\|f_j\| \leq 1$ for any $j = 1, \ldots, M$.

For any function g, we write for brevity $R_j(g) = \|g - \eta_j\|^2$. Set also $\mathcal{H} = \{f_1, \ldots, f_M\}$. It is easy to check that

$$\min_{f \in \mathcal{H}} R_j(f) = R_j(f_j) = \|f_j - \eta_j\|^2 \,. \tag{22}$$

We now reduce our estimation problem to a testing problem as follows. Let $\psi \in \{1, \ldots, M\}$ be the random variable, or *test*, defined by $\psi = j$ if and only if $\hat{S}_n = f_j$. Then, $\psi \neq j$ implies that there exists $k \neq j$ such that $\hat{S}_n = f_k$, so that

$$\begin{aligned} \|\hat{S}_n - \eta_j\|^2 - \|f_j - \eta_j\|^2 &= \|f_k - f_j\|^2 + 2\langle f_k - f_j, f_j - \eta_j\rangle \\ &= \tau^2(1+\alpha)^2\|\phi_j - \phi_k\|^2 + 2\tau^2(1+\alpha)(\langle\phi_j, \phi_k\rangle - 1) \\ &\geq \tau^2\alpha\|\phi_j - \phi_k\|^2. \end{aligned}$$

From (21), we find that $\|\phi_j - \phi_k\|^2 \geq 2\underline{\kappa}^2$ so that

$$\|\hat{S}_n - \eta_j\|^2 - \|f_j - \eta_j\|^2 \geq \frac{2\tau^2\underline{\kappa}^2\sigma}{3\bar{\kappa}}\sqrt{\frac{\log M}{n}} \stackrel{\text{def}}{=} \nu_{n,M}.$$

Therefore, we conclude that $\psi \neq j$ implies that

$$R_j(\hat{S}_n) - \min_{f \in \mathcal{H}} R_j(f) \geq \nu_{n,M}.$$

Hence,

$$\max_{1\le j\le M} P_j\left\{R_j(\hat{S}_n) - \min_{f\in\mathcal{H}} R_j(f) \ge \nu_{n,M}\right\} \ge \inf_{\psi} \max_{1\le j\le M} P_j(\psi \ne j)\,, \tag{23}$$

where the infimum is taken over all tests taking values in $\{1,\dots,M\}$ and P_j denotes the joint distribution of $Y_1,\dots,Y_n$ that are independent Gaussian random variables with variance σ^2 and means $\eta_j(X_1),\dots,\eta_j(X_n)$ respectively. It follows from Proposition 2.3 and Theorem 2.5 in Tsybakov (2009) that if for any $1 \le j,k \le M$, the Kullback-Leibler divergence between P_j and P_k satisfies

$$\mathcal{K}(P_j,P_k) < \frac{\log M}{8}\,, \tag{24}$$

then there exists a constant $C>0$ such that

$$\inf_{\psi} \max_{1\le j\le M} P_j(\psi \ne j) \ge C\,. \tag{25}$$

To check (24), observe that, choosing $\tau \le 1$ and applying (21), we get

$$\mathcal{K}(P_j,P_k) = \frac{n}{2\sigma^2}\|\eta_j-\eta_k\|^2 = \frac{\tau^2\log M}{18\bar{\kappa}^2}\|\phi_j-\phi_k\|^2 < \frac{\log M}{8}\,.$$

Therefore, in view of (23) and (25), we find using the Markov inequality that for any selector $\hat{S}_n$,

$$\max_{1\le j\le M} E_j\left[R_j(\hat{S}_n) - \min_{f\in\mathcal{H}} R_j(f)\right] \ge C\nu_{n,M} = C_*\sigma\sqrt{\frac{\log M}{n}}\,,$$

where E_j denotes the expectation with respect to P_j. This proves the theorem.

4 Sparsity and high dimensional regression

Let us go back to Scenario 1 (sparse linear regression). We assume that $f = f_{\theta^*}$ for some $\theta^* \in \mathbb{R}^M$, and θ^* is s-sparse. Using Proposition 2 we obtain that the least squares estimator satisfies

$$\begin{aligned} E\|f_{\hat{\theta}^{LS}} - f\|^2 = E\|f_{\hat{\theta}^{LS}} - f_{\theta^*}\|^2 &= E\left(\frac{1}{n}|X(\hat{\theta}^{LS}-\theta^*)|_2^2\right) \\ &= \min_{\theta\in\mathbb{R}^M} \frac{1}{n}|X(\theta-\theta^*)|_2^2 + \frac{\sigma^2(M\wedge n)}{n} \\ &= \frac{\sigma^2(M\wedge n)}{n} \end{aligned}$$

whenever the matrix X is of full rank $M\wedge n$. This result is useless in high-dimensional problems when $M > n$ since the remainder term is not small. The sparsity s is not involved in the expression for the risk. So, the global least squares cannot take advantage of sparsity, even if the target vector is very sparse, i.e., $s \ll M\wedge n$. On the other hand, imagine that some "oracle" discloses to us the set of non-zero components of the target vector $J(\theta^*) = \{j : \theta_j^* \ne 0\}$. Then we can use the least squares estimator restricted to the linear subspace of vectors with non-zero components in $J(\theta^*)$. Denoting this estimator by $\hat{\theta}^{LS,J(\theta^*)}$ and applying again Proposition 2 we find

$$E\left(\frac{1}{n}|X(\hat{\theta}^{LS,J(\theta^*)}-\theta^*)|_2^2\right) \le \frac{\sigma^2|\theta^*|_0}{n} \le \frac{\sigma^2 s}{n}$$

where we have used that $\mathrm{Card}(J(\theta^*)) = |\theta^*|_0$ and that θ^* is s-sparse. This bound is much better, it takes advantage of sparsity and can be very small when $s \ll n$. Unfortunately, $\hat{\theta}^{LS,J(\theta^*)}$ is not an estimator. It is an *oracle*; it depends on the unknown θ^* and cannot be computed from the data.

A natural question in this context is whether one can construct a true estimator $\tilde{\theta}$ such that

$$E\left(\frac{1}{n}|X(\tilde{\theta}-\theta^*)|_2^2\right) \le \frac{\sigma^2|\theta^*|_0}{n} \ ?$$

We will see that this is "almost" possible. In particular, we will exhibit an estimator $\tilde{\theta}$ such that

$$E\left(\frac{1}{n}|X(\tilde{\theta}-\theta^*)|_2^2\right) \le C\frac{\sigma^2|\theta^*|_0}{n}\log\left(\frac{M}{|\theta^*|_0}\right) \tag{26}$$

for some constant C and all $0 < |\theta^*|_0 < M$. The additional logarithmic factor in (26) characterizes the (modest) "price" to pay for the lack of knowledge of the set $J(\theta^*)$. We will see that this factor cannot be avoided in a minimax sense on the class of all s-sparse vectors. Inequality (26) is an example of *sparsity oracle inequality*.

4.1 Sparsity in Gaussian sequence model

To give an idea how to construct estimators $\tilde{\theta}$ satisfying (26), we consider a simple but instructive case when the columns of matrix X are orthonormal.

Assumption (ORT). *Matrix X is such that $\frac{1}{n}X^TX = I_M$ where I_M is the $M \times M$ identity matrix, $M \ge 2$.*

This assumption implies that $M \le n$ since otherwise X^TX is degenerate.

Using the model $y = X\theta^* + \xi$ we may write

$$\begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix} \stackrel{\mathrm{def}}{=} \frac{1}{n}X^Ty = \frac{1}{n}X^TX\theta^* + \frac{1}{n}X^T\xi = \theta^* + \zeta,$$

where $\zeta = \frac{1}{n}X^T\xi$ is a Gaussian random vector in $\mathbb{R}^M$ with mean zero and covariance matrix

$$\mathrm{V}(\zeta) = \frac{1}{n^2}E(X^T\xi\xi^TX) = \frac{\sigma^2}{n}I_M.$$

Thus, the components ζ_j of ζ are i.i.d. Gaussian random variables that can be written in the form $\zeta_j = \varepsilon\eta_j$ where $\varepsilon = \frac{\sigma}{\sqrt{n}}$ and $\eta_1, \dots, \eta_M$ are i.i.d. standard normal.

We see that, under Assumption (ORT), we have a sequence of "new" observations $y_1, \dots, y_M$ of the form

$$y_j = \theta_j^* + \varepsilon\eta_j, \quad j = 1, \dots, M, \quad \varepsilon = \frac{\sigma}{\sqrt{n}}, \tag{27}$$

where θ_j^* is the jth component of θ^* and η_j are i.i.d. $\mathcal{N}(0,1)$ random variables. The model (27) is called the *Gaussian sequence model* and has a simple "signal + noise" interpretation.

In the rest of this subsection, we will "forget" the initial model $y = X\theta^* + \xi$ and work with a sequence of observations $y_1, \dots, y_M$ satisfying (27). Note first that, for any θ, in view of Assumption (ORT),

$$\frac{1}{n}|X\theta|_2^2 = \frac{1}{n}\theta^TX^TX\theta = |\theta|_2^2,$$

so that the squared risk of an arbitrary estimator $\hat{\theta}$ simplifies to

$$E\left(\frac{1}{n}|X(\hat{\theta}-\theta^*)|_2^2\right) = E|\hat{\theta}-\theta^*|_2^2. \tag{28}$$

As discussed above, under the sparsity assumption on θ^*, it is crucial to detect the set of non-zero components $J(\theta^*)$. For the Gaussian sequence model (27), such a detection is based on a very simple idea to keep only the indices j such that the absolute values $|y_j|$ are large enough. To quantify the notion of "large enough value", we will refer to the following property (cf. Lemma 27 below): *If η_j are standard Gaussian random variables then $\max_{1\le j\le M}|\eta_j| \le \sqrt{2\log M}$ with probability close to 1 for large M.*

Intuitively, the value $\sqrt{2\log M}$ characterizes the "noise level". The observation y_j is "under the noise level", or is difficult to distinguish from the noise if $|y_j| \le \varepsilon\sqrt{2\log M}$. On the contrary, if $|y_j| > c\varepsilon\sqrt{\log M}$ for some constant $c > \sqrt{2}$, then it is almost impossible to have $\theta_j^* = 0$. Thus, all indices j such that $|y_j| > \varepsilon\sqrt{2\log M} = \sigma\sqrt{\frac{2\log M}{n}}$ belong to the set $J(\theta^*)$ with probability close to 1 for large M.

These remarks lead us to estimation of coefficients θ_j^* by *thresholding*. It means that we use a suitable estimator of θ_j^* (for example, the least squares and maximal likelihood estimator equal to y_j) for indices j such that $|y_j| > c\sigma\sqrt{\frac{\log M}{n}}$ and we estimate by 0 all the coefficients θ_j^* such that $|y_j|$ is under the "noise level" $c\sigma\sqrt{\frac{\log M}{n}}$. A basic realization of this idea is given by the *hard thresholding* estimator

$$\hat{\theta}_j^H = y_j I(|y_j| > \tau),$$

wher $\tau > 0$ is the threshold, typically chosen of the order $\sqrt{\frac{\log M}{n}}$. The following theorem summarizes the main properties of the hard thresholding estimator $\hat{\theta}^H = (\hat{\theta}_1^H, \dots, \hat{\theta}_M^H)$.

Theorem 7. *Consider the linear regression model under Assumption (ORT). Then the following holds.*

(i) **(Oracle inequality in expectation)** *If $\tau = \sigma\sqrt{\frac{2\log M}{n}}$ and $\theta^* \neq 0$, then*

$$E|\hat{\theta}^H - \theta^*|_2^2 \le 2\sigma^2\frac{|\theta^*|_0}{n}\log M\left(1+\frac{4}{\sqrt{\log M}}\right).$$

(ii) **(Oracle inequality in probablility)** *If $\tau = A\sigma\sqrt{\frac{\log M}{n}}, A > 2\sqrt{2}$, then with probability at least $1 - M^{1-A^2/8}$ we have:*

$$|\hat{\theta}^H - \theta^*|_2^2 \le \frac{9}{4}A^2\sigma^2\Big(\frac{|\theta^*|_0}{n}\log M\Big).$$

(iii) **(Selection of variables)** *If $\tau = B\sigma\sqrt{\frac{\log M}{n}}$ with $B > \sqrt{2}$ and*

$$\min_{j:\theta_j^*\neq 0}|\theta_j^*| > 2\tau,$$

then, with probability at least $1 - M^{1-B^2/2}$ we have:

$$\hat{J} = J(\theta^*),$$

where $J(\theta^*) = \{j : \theta_j^* \neq 0\}$ and $\hat{J} = \{j : \hat{\theta}_j^H \neq 0\}$.

Proof.
(i). If $\theta_j^* = 0$, then

$$|\hat{\theta}_j^H - \theta_j^*| = |y_j| I(|y_j| > \tau) = \varepsilon |\eta_j| I(|\eta_j| > \sqrt{2 \log M}),$$

while for $\theta_j^* \neq 0$ we have the bound

$$|\hat{\theta}_j^H - \theta_j^*| = |y_j I(|y_j| > \tau) - \theta_j^*| \leq |y_j - \theta_j^*| + |y_j| I(|y_j| \leq \tau) \leq \varepsilon |\eta_j| + \tau.$$

Therefore,

$$\begin{aligned} E|\hat{\theta}^H - \theta|_2^2 &= \sum_{j=1}^{M} E|\hat{\theta}_j^H - \theta_j^*|^2 \\ &\leq M\varepsilon^2 E\left[\eta_1^2 I(|\eta_1| > \sqrt{2 \log M})\right] + |\theta^*|_0 E\left[(\varepsilon|\eta_j| + \tau)^2\right]. \end{aligned} \tag{29}$$

Since $E(\eta_1^2) = 1$ and $E|\eta_1| = \sqrt{2/\pi}$,

$$\begin{aligned} E\left[(\varepsilon|\eta_j| + \tau)^2\right] &= \varepsilon^2 + \frac{4\varepsilon\tau}{\sqrt{2\pi}} + \tau^2 \\ &= \varepsilon^2 \left(1 + 4\sqrt{\frac{\log M}{\pi}} + 2 \log M\right). \end{aligned} \tag{30}$$

By Lemma 26

$$E\left[\eta_1^2 I(|\eta_1| > \sqrt{2 \log M})\right] \leq \sqrt{\frac{2}{\pi}} \left(\frac{1}{\sqrt{2 \log M}} + 2\sqrt{2 \log M}\right) M^{-1}. \tag{31}$$

Plugging (30) and (31) in (29) and using that $1 + \frac{1}{\sqrt{\pi \log M}} \leq 4\sqrt{\frac{\log M}{\pi}}$ for all $M \geq 2$ and the inequality $6/\sqrt{\pi} \leq 4$, we obtain the result.

(ii). Set $r = \frac{A\sigma}{2}\sqrt{\frac{\log M}{n}} = \frac{A}{2}\varepsilon\sqrt{\log M}$. Consider the random event

$$\mathcal{A} = \{|y_j - \theta_j^*| \leq r, j = 1, \ldots, M\}.$$

By Lemma 27, the probability of the complementary event $\mathcal{A}^c$ satisfies

$$\begin{aligned} P(\mathcal{A}^c) &= P\left\{\max_{1 \leq j \leq M} |\zeta_j| > r\right\} \\ &= P\left\{\varepsilon \max_{1 \leq j \leq M} |\eta_j| > \frac{A}{2}\varepsilon\sqrt{\log M}\right\} \\ &\leq M^{1 - A^2/8}. \end{aligned}$$

On the event $\mathcal{A}$ we have, in view of Lemma 29,

$$|y_j I(|y_j| > 2r) - \theta_j^*| \leq 3 \min(|\theta_j^*|, r).$$

Using that $r = \tau/2$ this implies

$$|\hat{\theta}^H - \theta|_2^2 = \sum_{j=1}^{M} |\hat{\theta}_j^H - \theta_j^*|^2 \leq 9 \sum_{j=1}^{M} \min\left(|\theta_j^*|^2, \frac{\tau^2}{4}\right)$$
$$= 9 \sum_{j:\theta_j^* \neq 0} \frac{\tau^2}{4} = 9|\theta^*|_0 \frac{\tau^2}{4}.$$

(iii). Set $B = A/2$. Then r defined in the proof of part (ii) has the form $r = \tau$. Consider the event $\mathcal{A}$ defined in the proof of part (ii).
Let us show that $\hat{J} \subseteq J(\theta^*)$ on the event $\mathcal{A}$. Let $\hat{\theta}_j^H \neq 0$. In this case,

$$\hat{\theta}_j^H = y_j \Leftrightarrow |y_j| > \tau \Leftrightarrow |\theta_j^* + \varepsilon\eta_j| > \tau,$$

which implies $|\theta_j^*| > \tau - |\varepsilon\eta_j| \geq \tau - r = 0$ on the event $\mathcal{A}$. Therefore, $\theta_j^* \neq 0$.
Let us show that $J(\theta^*) \subseteq \hat{J}$ on the event $\mathcal{A}$. Let $\theta_j^* \neq 0$. Then $|\theta_j^*| > 2\tau$, which yields

$$|y_j| = |\theta_j^* + \varepsilon\eta_j| > 2\tau - |\varepsilon\eta_j|$$
$$\geq 2\tau - r = \tau$$

on the event $\mathcal{A}$. On the other hand, by definition of $\hat{\theta}^H$,

$$|y_j| > \tau \Rightarrow \hat{\theta}_j^H = y_j$$

Thus, $\hat{\theta}_j^H \neq 0$ with probability 1.

There exist other thresholding estimators behaving similarly as described in Theorem 7. For example, if τ is the same threshold, the *soft thresholding estimator* defined as

$$\hat{\theta}_j^S = \max\left(1 - \frac{\tau}{|y_j|}, 0\right) y_j, \quad j = 1, \ldots, M, \tag{32}$$

and the *non-negative garrotte estimator*[1], defined as

$$\hat{\theta}_j^G = \max\left(1 - \frac{\tau^2}{y_j^2}, 0\right) y_j \quad j = 1, \ldots, M, \tag{33}$$

have similar risk and selection of variables behavior.

We can equivalently define the soft and hard thresholding estimators in terms of optimization programs as described below.

Proposition 8. *The soft and hard thresholding estimators are solutions to the following optimization problems*

$$\hat{\theta}^H = \underset{\theta \in \mathbb{R}^M}{\operatorname{argmin}} \sum_{j=1}^{M} (y_j - \theta_j)^2 + \tau^2 |\theta|_0, \tag{34}$$

$$\hat{\theta}^S = \underset{\theta \in \mathbb{R}^M}{\operatorname{argmin}} \sum_{j=1}^{M} (y_j - \theta_j)^2 + 2\tau |\theta|_1. \tag{35}$$

[1]This estimator is closely related to the James-Stein estimator.

Furthermore, under Assumption (ORT), we can express these two estimators as follows:

$$\hat{\theta}^H = \underset{\theta\in\mathbb{R}^M}{\operatorname{argmin}}\left(\frac{1}{n}|y - X\theta|_2^2 + \tau^2|\theta|_0\right), \tag{36}$$

$$\hat{\theta}^S = \underset{\theta\in\mathbb{R}^M}{\operatorname{argmin}}\left(\frac{1}{n}|y - X\theta|_2^2 + 2\tau|\theta|_1\right). \tag{37}$$

Indeed, since we assume that $\frac{1}{n}X^TX = I_M$ (Assumption (ORT)) and we use the notation $\begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix} \overset{\text{def}}{=} \frac{1}{n}X^Ty$, we may write

$$\begin{aligned}
\sum_{j=1}^{M}(y_j - \theta_j)^2 &= \left|\frac{1}{n}X^Ty - \theta\right|_2^2 \\
&= |\theta|_2^2 - \frac{2}{n}\theta^TX^Ty + \frac{1}{n^2}y^TXX^Ty \\
&= \frac{1}{n}|X\theta|_2^2 - \frac{2}{n}\theta^TX^Ty + \frac{1}{n}y^TXX^Ty \\
&= \frac{1}{n}|X\theta - y|_2^2 + \frac{1}{n}y^TXX^Ty - \frac{1}{n}|y|_2^2 \\
&= \frac{1}{n}|y - X\theta|_2^2 + c
\end{aligned}$$

where c is a constant independent of θ.

An important observation is that the estimators (36) and (37) can be used with general matrices X and therefore can be applied in full generality in Scenarios 1 - 3 and not only in the Gaussian sequence model. For general X, the estimator defined by (36) is called the BIC estimator and that defined by (37) is called the Lasso estimator. So, the BIC and Lasso are natural extensions of the hard and soft thresholding estimators respectively.

4.2 Sparsity oracle inequality for the BIC

We now return to the general regression model $y = f + \xi$. Let $\tau > 0$ be a given threshold. The original *BIC estimator* is defined as follows

$$\begin{aligned}
\hat{\theta}^{BIC} &= \underset{\theta\in\mathbb{R}^M}{\operatorname{argmin}}\left(\frac{1}{n}|y - X\theta|_2^2 + \tau^2|\theta|_0\right) \\
&= \underset{\theta\in\mathbb{R}^M}{\operatorname{argmin}}\left(\|y - f_\theta\|^2 + \tau^2|\theta|_0\right).
\end{aligned}$$

Note that it can be considered not only as an estimator for Scenario 1 but it also generates a nonparametric estimator $f_{\hat{\theta}^{BIC}}$ for Scenario 2 and an aggregate $f_{\hat{\theta}^{BIC}}$ for Scenario 3. To get sharper bounds on the risk, it is convenient to slightly modify the BIC by replacing the term $\tau^2|\theta|_0$ by a penalty function $\text{pen}(\theta)$ defined by

$$\text{pen}(|\theta|_0) = \frac{2\sigma^2}{n}\left(1 + C_1\sqrt{L(\theta)} + \frac{C_2}{\epsilon}L(\theta)\right)|\theta|_0 \tag{38}$$

where C_1, C_2 are suitable positive constants, $\epsilon > 0$ is an arbitrary positive number, and

$$L(\theta) = \log\left(\frac{eM}{|\theta|_0 \vee 1}\right).$$

We will consider this penalty instead of $\tau^2|\theta|_0$ and use a modified definition of BIC:

$$\tilde{\theta}^{BIC} = \underset{\theta\in\mathbb{R}^M}{\operatorname{argmin}} \left(\frac{1}{n}|y - X\theta|_2^2 + \operatorname{pen}(|\theta|_0) \right). \tag{39}$$

Both versions of the BIC are penalized least squares estimators where the penality is imposed on the size of the support of θ. However, the BIC optimization problem is NP-hard. To see this, we can reformulate the BIC program as follows

$$\begin{aligned} \min_{\theta\in\mathbb{R}^M} \left(\frac{1}{n}|y - X\theta|_2^2 + \operatorname{pen}(|\theta|_0) \right) &= \min_{0\le m\le M} \min_{\theta:|\theta|_0=m} \left(\frac{1}{n}|y - X\theta|_2^2 + \operatorname{pen}(|\theta|_0) \right) \\ &= \min_{0\le m\le M} \left(\min_{\theta:|\theta|_0=m} \frac{1}{n}|y - X\theta|_2^2 + \operatorname{pen}(m) \right). \end{aligned}$$

Thus, we have to solve $\sum_{m=0}^{M} \binom{M}{m} = 2^M$ possible least squares problems. Despite the computational unfeasibility, the theoretical properties of the BIC estimator can be analyzed in detail. In particular, it satisfies the oracle inequalities given in the next theorem.

Theorem 9 (Oracle Inequality for BIC)**.** *Fix $\epsilon > 0$. Let $\tilde{\theta}^{BIC}$ be defined in (38)–(39) with sufficiently large C_1 and C_2 and let $\tilde{f}^{BIC} = f_{\tilde{\theta}^{BIC}}$. Then there exists a constant $C > 0$ such that, for all f,*

$$E\|\tilde{f}^{BIC} - f\|^2 \le (1+\epsilon) \min_{\theta\in\mathbb{R}^M} \left(\|f_\theta - f\|^2 + \frac{C}{\epsilon} \frac{\sigma^2|\theta|_0}{n} \log\left(\frac{eM}{|\theta|_0 \vee 1} \right) \right) + \frac{C\sigma^2}{n}. \tag{40}$$

In addition, there exists a constant $C > 0$ such that, for any $0 < \delta < 1$ with probability at least $1 - \delta$,

$$\forall\ f: \quad \|\tilde{f}^{BIC} - f\|^2 \le (1+\epsilon) \min_{\theta\in\mathbb{R}^M} \left[\|f_\theta - f\|^2 + \frac{C}{\epsilon} \frac{\sigma^2|\theta|_0}{n} \log\left(\frac{eM}{|\theta|_0 \vee 1} \right) \right] + \frac{C\sigma^2}{n} \log\left(\frac{1}{\delta} \right). \tag{41}$$

In particular, if $f(x) = f_{\theta^*}(x)$ with $\theta^* \neq 0$,

$$E\left(\frac{1}{n} \left| X(\tilde{\theta}^{BIC} - \theta^*) \right|_2^2 \right) \le C \frac{\sigma^2}{n} |\theta^*|_0 \log\left(\frac{eM}{|\theta^*|_0 \vee 1} \right). \tag{42}$$

The oracle inequality in expectation (40) is proved in Birgé and Massart (2007) (see also Johnstone (2013)). For the proof of the inequality in probability (41), see Bunea, Tsybakov and Wegkamp (2004).

Remarks.

1. Inequalities of Theorem 9 are *sparsity oracle inequalities* since the remainder term depends only on $|\theta|_0$. For instance, the "in expectation" version (40) is of the form

$$E\|\hat{f} - f\|^2 \le K \min_{\theta\in\mathbb{R}^M} \left(\|f_\theta - f\|_2^2 + \Delta_{n,M}(\theta) \right) \tag{43}$$

where $\hat{f}$ is an estimator of f, c is a constant, and $\Delta_{n,M} > 0$ only depends on $|\theta|_0$[2].

[2]If $\Delta_{n,M}$ depends on $|\theta|_0$ and other features of θ, then the corresponding oracle inequality is sometimes referred to as a *balanced oracle inequality*.

2. The sparsity oracle inequalities of Theorem 9 are *not sharp*, i.e., the leading constant K is greater than 1. In particular, we cannot obtain a meaningful bound on the excess risk using inequality (40). Indeed, since it is of the form (43) with $K > 1$ the excess risk can be only bounded as

$$\mathcal{E}_\Theta(\hat{f}, f) = E\|\hat{f} - f\|^2 - \min_{\theta\in\Theta} \|f_\theta - f\|^2 \le (K-1) \min_{\theta\in\Theta} \|f_\theta - f\|^2 + K \sup_{\theta\in\Theta} \Delta_{n,M}(\theta).$$

But this bound is useless in the aggregation context because we have no control of the minimum $\min_{\theta\in\Theta} \|f_\theta - f\|^2$ (it can be arbitrarily large).

3. The oracle inequalities of Theorem 9 hold under no assumption on the dictionary $f_1, \dots, f_M$, and (except for inequality (42)) under no assumption of f.

4. Inequality (42) gives a solution to the question announced above, cf. (26). It contains an oracle term $C\frac{\sigma^2}{n}|\theta^*|_0$ multiplied by $\log\left(\frac{eM}{|\theta^*|_0 \vee 1}\right)$. This factor represents the price to pay for not knowing the set of non-zero components of θ^*.

5. Instead of the penalty (39) implemented above, we can also use the penalty

$$\mathrm{pen}(\theta) = C\sigma^2|\theta|_0 \log M.$$

This leads to oracle inequalities similar to those of Theorem 9 except for the logarithmic factors that become slightly suboptimal. More precisely, $\log\left(\frac{eM}{|\theta|_0 \vee 1}\right) \to \log M$ for this type of penalty.

4.3 Sparsity oracle inequality for the Lasso

As in the previous subsection, here we consider the general regression model $y = f + \xi$. Let $\hat{\theta}^L$ be the Lasso estimator

$$\begin{aligned}\hat{\theta}^L &= \operatorname*{argmin}_{\theta\in\mathbb{R}^M} \left(\frac{1}{n}|y - X\theta|_2^2 + 2\tau|\theta|_1\right) \\ &= \operatorname*{argmin}_{\theta\in\mathbb{R}^M} \left(\|y - f_\theta\|^2 + 2\tau|\theta|_1\right)\end{aligned}$$

where $\tau > 0$ is a tuning parameter. Similarly to the BIC estimator, it can be considered not only as an estimator for parametric Scenario 1 but also it generates a nonparametric estimator $f_{\hat{\theta}^L}$ for Scenario 2 and an aggregate $f_{\hat{\theta}^L}$ for Scenario 3. The following theorem is a modification of a result in Koltchinskii, Lounici and Tsybakov (2011). It provides a sparsity oracle inequality in probability with leading constant 1 for the Lasso estimator.

Theorem 10. *Let ξ be i.i.d. random variables, $\xi_i \sim \mathcal{N}(0, \sigma^2)$ and let $\|f_i\| \le 1$, $j = 1, \dots, M$. Let $\hat{\theta}^L$ be the Lasso estimator with the tuning parameter $\tau = A\sigma\sqrt{\frac{\log M}{n}}$, $A = \frac{t}{\delta}$, $t > \sqrt{2}$, $0 < \delta < 1$. Then, with probability at least $1 - M^{1-t^2/2}$ we have*

$$\forall f: \qquad \|f_{\hat{\theta}^L} - f\|^2 \le \min_{\theta\in\mathbb{R}^M} \left[\|f_\theta - f\|^2 + C\min\left(\sigma^2\mu^2(\theta)|\theta|_0 \frac{\log M}{n}, \sigma|\theta|_1\sqrt{\frac{\log M}{n}}\right)\right] \tag{44}$$

where $C > 0$ depends only on t and δ,

$$\mu(\theta) = \inf\left\{\mu > 0 : |\Delta_{J(\theta)}|_1 \leq \mu\sqrt{\frac{|\theta|_0}{n}}|X\Delta|_2, \forall \Delta \in C_\theta\right\}$$

with

$$C_\theta = \left\{\Delta \in \mathbb{R}^M : |\Delta_{J^c(\theta)}|_1 \leq \frac{1+\delta}{1-\delta}|\Delta_{J(\theta)}|_1\right\}.$$

Proof. Set for brevity $\hat{\theta}^L = \hat{\theta}$ and

$$G(\theta) = \|y - f_\theta\|^2 + 2\tau|\theta|_1.$$

Then

$$\hat{\theta} = \underset{\theta\in\mathbb{R}^M}{\operatorname{argmin}}\ G(\theta).$$

Denote by $(\cdot,\cdot)$ the inner product in $\mathbb{R}^M$, and set

$$< f, g > \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f(X_i)g(X_i).$$

We now recall the following general fact from convex analysis.

Lemma 11. *For any convex function $G : \mathbb{R} \to \mathbb{R}^M$ we have: $\hat{\theta} \in \underset{\theta\in\mathbb{R}^M}{\operatorname{argmin}}\ G(\theta)$ if and only if $0 \in \partial G(\hat{\theta})$, where $\partial G(\hat{\theta})$ is the subdiffrential of G at point $\hat{\theta}$.*

The condition $0 \in \partial G(\hat{\theta})$ of this lemma obviously implies: *there exists $B \in \partial G(\hat{\theta})$ such that $(B, \hat{\theta} - \theta) = 0$, for all $\theta \in \mathbb{R}^M$*. In the sequel, we will use this property.

In our case,

$$\nabla(\|y - f_\theta\|^2) = \nabla(\frac{1}{n}|y - X\theta|_2^2) = -\frac{2}{n}X^T(y - X\theta),$$

and thus

$$\begin{aligned}(\nabla(\|y - f_{\hat{\theta}}\|^2), \hat{\theta} - \theta) &= -\frac{2}{n}(X^T(y - X\hat{\theta}), \hat{\theta} - \theta) \\ &= -\frac{2}{n}(X(\hat{\theta} - \theta), y - X\hat{\theta}) = -2 < f_{\hat{\theta}-\theta}, y - f_{\hat{\theta}} > .\end{aligned}$$

Applying Lemma 11, we get that there exists $\hat{V} \in \partial(|\hat{\theta}|_1)$ such that

$$-2\langle f_{\hat{\theta}-\theta}, y - f_{\hat{\theta}}\rangle + 2\tau(\hat{V}, \hat{\theta} - \theta) = 0. \tag{45}$$

Let V be any element of $\partial(|\theta|_1)$. It follows from (45) that

$$-2\langle f_{\hat{\theta}-\theta}, y - f_{\hat{\theta}}\rangle + 2\tau(\hat{V} - V, \hat{\theta} - \theta) = -2\tau(V, \hat{\theta} - \theta). \tag{46}$$

We now use the following fact from convex analysis applied to the function $g(\theta) = |\theta|_1$.

Lemma 12. *For any convex function $g : \mathbb{R} \to \mathbb{R}^M$, we have*

$$(V - V', \theta - \theta') \geq 0, \quad \forall \theta, \theta' \in \mathbb{R}^M,$$

for all $V \in \partial g(\theta), V' \in \partial g(\theta')$.

From Lemma 12 and (46) we find

$$-2\langle f_{\hat{\theta}-\theta}, y - f_{\hat{\theta}} \rangle \leq -2\tau(V, \hat{\theta} - \theta).$$

Since $y = f + \xi$, we can rewrite this in the form:

$$2\langle f_{\hat{\theta}-\theta}, f_{\hat{\theta}} - f \rangle \leq -2\tau(V, \hat{\theta} - \theta) + 2\langle \xi, f_{\hat{\theta}-\theta} \rangle \tag{47}$$

for any $V \in \partial(|\theta|_1)$ and any $\theta \in \mathbb{R}^M$. Next, elementary argument yields

$$2\langle f_{\hat{\theta}-\theta}, f_{\hat{\theta}} - f \rangle = 2\langle f_{\hat{\theta}} - f_\theta, f_{\hat{\theta}} - f \rangle = \|f_{\hat{\theta}} - f_\theta\|^2 + \|f_{\hat{\theta}} - f\|^2 - \|f_\theta - f\|^2. \tag{48}$$

Fix some $\theta \in \mathbb{R}^M$ and let $J = J(\theta)$ be the set of non-zero components of θ. Write

$$V = V_J + V_{J^c}$$

where $V_J \in \mathbb{R}^M$ is the vector with components $V_j I(j \in J)$, $j = 1, \ldots, M$, where V_j are the components of V, and $J^c = \{1, \ldots, M\} \backslash J$ is the complement of J. Then

$$(V, \hat{\theta} - \theta) = (V_J, \hat{\theta} - \theta) + (V_{J^c}, \hat{\theta} - \theta) = (V_J, \hat{\theta} - \theta) + (V_{J^c}, \hat{\theta})$$

since the components of V_{J^c} vanish on the support of θ. On the other hand, V is any element of $\partial(|\theta|_1)$, and thus the components of V satisfy

$$\begin{cases} |V_j| \leq 1, \ j \in J^c, \\ V_j = \text{sign}(\theta_j), \ j \in J. \end{cases}$$

Choose V such that $V_j = \text{sign}(\hat{\theta}_j)$ for $j \in J^c$. This is possible, since V_j can be any values satisfying $|V_j| \leq 1$ for $j \in J^c$. Then

$$(V, \hat{\theta} - \theta) = (V_J, \hat{\theta} - \theta) + |\hat{\theta}_{J^c}|_1 = (V_J, \Delta) + |\Delta_{J^c}|_1 = (V_J, \Delta_J) + |\Delta_{J^c}|_1$$

where $\Delta = \hat{\theta} - \theta$ and we used the fact that $|\hat{\theta}_{J^c}|_1 = |\Delta_{J^c}|_1$. This and (47) imply

$$2\langle f_{\hat{\theta}-\theta}, f_{\hat{\theta}} - f \rangle \leq 2\tau|\Delta_J|_1 - 2\tau|\Delta_{J^c}|_1 + 2(H, \Delta), \tag{49}$$

where $H = \frac{1}{n} X^T \xi = \begin{pmatrix} H_1 \\ \vdots \\ H_M \end{pmatrix}$ with $H_j = \frac{1}{n} \sum_{i=1}^n f_j(X_i)\xi_i$. We have used here the identity $\langle \xi, f_{\hat{\theta}-\theta} \rangle = (H, \hat{\theta} - \theta)$. Note now that if

$$\langle f_{\hat{\theta}-\theta}, f_{\hat{\theta}} - f \rangle \leq 0$$

then, in view of (48), we get

$$\|f_{\hat{\theta}} - f\|^2 \leq \|f_\theta - f\|^2$$

and the result of the theorem follows in a trivial way. So, it is enough to consider the case $\langle f_{\hat{\theta}-\theta}, f_{\hat{\theta}} - f\rangle \geq 0$. But in this case, in view of (49),

$$\tau|\Delta_{J^c}|_1 \leq \tau|\Delta_J|_1 + |H|_\infty|\Delta|_1.$$

Assume for the moment that

$$|H|_\infty \leq \delta\tau$$

for some $0 < \delta < 1$. Then, since $|\Delta|_1 = |\Delta_J|_1 + |\Delta_{J^c}|_1$, we have

$$|\Delta_{J^c}|_1 \leq \frac{1+\delta}{1-\delta}|\Delta_J|_1.$$

In other words, it suffices to consider $\Delta \in C_\theta$, where

$$C_\theta = \left\{\Delta \in \mathbb{R}^M : |\Delta_{J^c}|_1 \leq \frac{1+\delta}{1-\delta}|\Delta_J|_1\right\}$$

and $J = J(\theta)$ is the set of non-zero components of θ.

We now return to (49), and bound the terms on the right-hand side of (49). Using that $|H|_\infty \leq \delta\tau$ and $\Delta \in C_\theta$ we get

$$\begin{aligned} 2\tau|\Delta_J|_1 - 2\tau|\Delta_{J^c}|_1 + 2(H,\Delta) &\leq 2\tau|\Delta_J|_1 - 2\tau|\Delta_{J^c}|_1 + 2|H|_\infty|\Delta|_1 \\ &= 2\tau|\Delta_J|_1 - 2\tau|\Delta_{J^c}|_1 + 2\delta\tau(|\Delta_J|_1 + |\Delta_{J^c}|_1) \\ &\leq 2\tau(1+\delta)|\Delta_J|_1. \end{aligned} \tag{50}$$

This and (49) imply

$$2\langle f_{\hat{\theta}-\theta}, f_{\hat{\theta}} - f\rangle \leq 2\tau(1+\delta)|\Delta_J|_1.$$

Combining this with (48) we get

$$\|f_{\hat{\theta}} - f\|^2 \leq \|f_\theta - f\|^2 - \|f_{\hat{\theta}} - f_\theta\|^2 + 2\tau(1+\delta)|\Delta_J|_1. \tag{51}$$

Since $\Delta \in C_\theta$, we get

$$|\Delta_J|_1 \leq \mu(\theta)\sqrt{\frac{|\theta|_0}{n}}\ |X\Delta|_2 = \mu(\theta)\sqrt{|\theta|_0}\ \|f_\Delta\| = \mu(\theta)\sqrt{|\theta|_0}\ \|f_{\hat{\theta}} - f_\theta\|.$$

This and the elementary inequality $2ab \leq a^2 + b^2$ yield

$$2\tau(1+\delta)|\Delta_J|_1 \leq 2\tau(1+\delta)\mu(\theta)\sqrt{|\theta|_0}\ \|f_{\hat{\theta}} - f_\theta\| \leq \tau^2(1+\delta)^2\mu^2(\theta)|\theta|_0 + \|f_{\hat{\theta}} - f_\theta\|^2. \tag{52}$$

Combining (51) and (52) we obtain

$$\forall\theta, \forall f: \qquad \|f_{\hat{\theta}} - f\|^2 \leq \|f_\theta - f\|^2 + \tau^2(1+\delta)^2\mu^2(\theta)|\theta|_0. \tag{53}$$

Note that this inequality is proved for all $\theta \in \mathbb{R}^M$ and all f, *under the assumption that*

$$|H|_\infty \leq \delta\tau.$$

Let us now show that $|H|_\infty \le \delta\tau$ holds with probability at least $1 - M^{1-t^2/2}$. Consider the random event

$$\mathcal{A} = \{|H|_\infty \le \delta\tau\} .$$

The probability of its complement $P(\mathcal{A}^c)$ is estimated as follows

$$P(\mathcal{A}^c) = P\left(\max_{1\le j\le M}\left|\frac{1}{n}\sum_{i=1}^n f_j(X_i)\xi_i\right| > \delta\tau\right) \le \sum_{j=1}^M P\left(\left|\frac{1}{n}\sum_{i=1}^n f_j(X_i)\xi_i\right| > \delta\tau\right) .$$

Here, for each j,

$$\frac{1}{n}\sum_{i=1}^n f_j(X_i)\xi_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\|f_j\|^2\right) .$$

It follows similarly to Lemma 27 that, since $\|f_j\|^2 \le 1$ for all j, we get

$$P(|H|_\infty \ge \delta\tau) = P\left(|H|_\infty > \sigma t\sqrt{\frac{\log M}{n}}\right) \le M^{1-t^2/2} .$$

To finish the proof of the theorem, we show that, on the same event $\mathcal{A}$,

$$\forall\theta, \forall f: \qquad \|f_{\hat\theta} - f\|^2 \le \|f_\theta - f\|^2 + C'\sigma|\theta|_1\sqrt{\frac{\log M}{n}} \tag{54}$$

for a constant $C' > 0$ depending only on t and δ. Indeed, since $G(\hat\theta) \le G(\theta)$ for all $\theta \in \mathbb{R}^M$, we get, by a simple algebra,

$$\|f_{\hat\theta} - f\|^2 \le \|f_\theta - f\|^2 + 2\langle\xi, f_{\hat\theta} - f_\theta\rangle + 2\tau|\theta|_1 - 2\tau|\hat\theta|_1. \tag{55}$$

Since $\langle\xi, f_{\hat\theta} - f_\theta\rangle = (H, \hat\theta - \theta)$ and $|H|_\infty \le \delta\tau$ on $\mathcal{A}$, we find

$$2\langle\xi, f_{\hat\theta} - f_\theta\rangle \le 2\delta\tau|\hat\theta - \theta|_1 + 2\tau(|\theta|_1 - |\hat\theta|_1) \le 2\tau(1+\delta)|\theta|_1. \tag{56}$$

Combining (55) and (56) we get (54) with $C' = 2t(1 + 1/\delta)$. Finally, the theorem follows from (53) and (54). The constant C in (44) can be taken equal to $\max(t^2(1+1/\delta)^2, 2t(1+1/\delta))$.

5 Mixing with exponential weights

Let $f_1, \dots, f_M$ be given functions forming a dictionary. Set

$$\hat r_j = \|y - f_j\|^2.$$

This is the empirical risk of f_j. The exponentially weighted aggregate is defined by

$$\hat f^{EW} = \sum_{j=1}^M \hat\theta_j^{EW} f_j = f_{\hat\theta^{EW}}$$

where $\hat\theta^{EW} = (\hat\theta_1^{EW}, \dots, \hat\theta_M^{EW})$ with

$$\hat\theta_j^{EW} = \frac{\exp(-n\hat r_j/\beta)\pi_j}{\sum_{k=1}^M \exp(-n\hat r_k/\beta)\pi_k}$$

for some $\beta > 0$ and some set of prior probabilities $\pi_k > 0$, $\sum_{k=1}^M \pi_k = 1$. This definition has been brought to Machine Learning by Vovk (1990), Littlestone and Warmuth (1994).

There exist two heuristic interpretations of exponential weighting.

1. *Quasi-bayesian interpretation.* The weights $\hat{\theta}^{EW}$ define a posterior distribution (which is the Gibbs distribution if π_k are uniform) in the "phantom model"

$$Y_i = f_{\theta^*}(X_i) + \xi_i', \quad i = 1, \dots, n,$$

where ξ_i' are i.i.d. $\mathcal{N}(0, \frac{\beta}{2})$ random variables, $\theta^* \in \{e_1, \dots, e_M\}$, and π_j are prior probabilities of e_j.

2. *Variational interpretation.* It is not hard to check that $\hat{\theta}^{EW}$ is a solution of the following minimization problem:

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

where $\mathcal{K}(\theta, \pi) = \sum_{j=1}^M \theta_j \log \frac{\theta_j}{\pi_j}$ is the Kullback-Lieibler divergence between θ and π, and

$$\Lambda^M = \left\{ \theta : \theta_j \geq 0, \ \sum_{j=1}^M \theta_j = 1 \right\}$$

is a simplex. Note that

$$\sum_{j=1}^M \theta_j \hat{r}_j = \sum_{j=1}^M \theta_j \|y - f_j\|^2 \underset{Jensen}{\geq} \|y - f_\theta\|^2.$$

Thus, $\hat{\theta}^{EW}$ minimizes an upper approximation of the empirical risk penalized by Kullback-Leibler divergence from π:

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

Note that $\mathcal{K}(\theta, \pi) \geq 0$ and $\mathcal{K}(\theta, \pi) = 0 \Leftrightarrow \theta = \pi$. So, we penalize the solution for being too far from the prior π.

In what follows we set for brevity

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

The following proposition goes back to Vovk (1990) who considered a deterministic model. Indeed, no assumption on the distribution of y is needed.

Proposition 13. *The value $\hat{r} = \sum_{j=1}^M w_j \hat{r}_j$ satisfies*

$$\hat{r} \leq \min_{1 \leq j \leq M} \left(\hat{r}_j + \frac{\beta}{n} \log \frac{1}{\pi_j} \right).$$

As a consequence, for all y,

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

Note that if $\pi_j = \frac{1}{M}$, $j = 1, \ldots, M$ (the uniform prior), then $\frac{\beta}{n} \log \frac{1}{\pi_j} = \frac{\beta \log M}{n}$, which is the optimal rate of MS-aggregation. But the bound is for the *empirical risk* $\|y - \hat{f}^{EW}\|^2$ and not for the risk $E\|f - \hat{f}^{EW}\|^2$. Also, on the RHS we have the empirical risk $\|y - f_j\|^2$ and not the discrepancy $\|f - f_j\|^2$ as expected in our oracle inequalities.

Proof. Take logarithms of the both sides of the equation

$$w_j = \frac{\exp(-n\hat{r}_j/\beta)\pi_j}{Z}.$$

Then, for any k and j, we have

$$-\log Z = \frac{n\hat{r}_k}{\beta} + \log \frac{1}{\pi_k} + \log w_k, \qquad -\log Z = \frac{n\hat{r}_j}{\beta} + \log \frac{1}{\pi_j} + \log w_j,$$

so that

$$\frac{n\hat{r}_k}{\beta} = \frac{n\hat{r}_j}{\beta} + \log \frac{1}{\pi_j} - \log \frac{1}{\pi_k} + \log w_j - \log w_k.$$

Thus, using that $\log w_j \le 0$, we get

$$\hat{r} = \sum_{k=1}^{M} w_k \hat{r}_k \le \hat{r}_j + \frac{\beta}{n} \log \frac{1}{\pi_j} - \frac{\beta}{n} \underbrace{\sum_{k=1}^{M} w_k \log \frac{w_k}{\pi_k}}_{K(w,\pi)}.$$

Since $K(w, \pi) \ge 0$ the first result of the proposition follows. The second result is obtained from the first one using the inequalities:

$$\|y - \underbrace{\hat{f}^{EW}}_{=\sum_{j=1}^{M} w_j f_j}\|^2 = \|\sum_{j=1}^{M} w_j (y - f_j)\|^2 \underset{Jensen}{\le} \sum_{j=1}^{M} w_j \|y - f_j\|^2 = \sum_{j=1}^{M} w_j \hat{r}_j = \hat{r}.$$

The next proposition is inspired by the argument in Leung and Barron (2006).

Proposition 14. *(i) If $\beta = 4\sigma^2$, then $\hat{r} - \sigma^2$ is an unbiased estimator of the risk :*

$$E\|\hat{f}^{EW} - f\|^2 = E(\hat{r}) - \sigma^2.$$

(ii) If $\beta > 4\sigma^2$, then

$$E\|\hat{f}^{EW} - f\|^2 \le E(\hat{r}) - \sigma^2.$$

Proof. First, recall that

$$\hat{f}^{EW}(\cdot) = \sum_{j=1}^{M} w_j f_j(\cdot)$$

with

$$w_j = w_j(y) = \frac{\exp(-\frac{n}{\beta}\hat{r}_j)\pi_j}{Z}$$

where $Z = \sum_{k=1}^{M} \exp(-\frac{n}{\beta}\hat{r}_k)\pi_k$, $\hat{r}_j = \|y - f_j\|^2$. By Stein unbiased risk estimation formula (see e.g. Tsybakov (2009), p. 157), the statistic

$$\hat{R} \stackrel{\text{def}}{=} \|y - \hat{f}^{EW}\|^2 + \frac{2\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}^{EW}(X_i)}{\partial Y_i} - \sigma^2$$

is an unbiased estimator of the risk $E\|f - \hat{f}^{EW}\|^2$, i.e.,

$$E(\hat{R}) = E\|f - \hat{f}^{EW}\|^2. \tag{57}$$

Let us compute $\hat{R}$. Note that in the definition of $\hat{f}^{EW}$ only the weights w_j depend on $Y_1, \ldots, Y_n$. So, we need first to find the derivative $\frac{\partial w_j(y)}{\partial Y_i}$. Recall that $\hat{r}_j = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_j(X_i))^2$. Hence,

$$\frac{\partial \hat{r}_j}{\partial Y_i} = \frac{2}{n}(Y_i - f_j(X_i))$$

and we have

$$\begin{aligned}
\frac{\partial w_j}{\partial Y_i} &= \frac{\exp(-\frac{n}{\beta}\hat{r}_j)\pi_j}{Z^2}\left[-\frac{2}{\beta}(Y_i - f_j(X_i))Z + \frac{2}{\beta}\sum_{k=1}^{M}(Y_i - f_k(X_i))\exp(-\frac{n}{\beta}\hat{r}_k)\pi_k\right] \qquad (58)\\
&= -\frac{2w_j}{\beta}\left[(Y_i - f_j(X_i)) + \sum_{k=1}^{M}(Y_i - f_k(X_i))w_k\right]\\
&= \frac{2}{\beta}(f_j(X_i) - \hat{f}^{EW}(X_i))w_j.
\end{aligned}$$

On the other hand, since $w_j \geq 0$, $\sum_{j=1}^{M} w_j = 1$, we have, by the "bias-variance" decomposition with respect to the distribution defined by $\{w_j\}$,

$$\begin{aligned}
\|\hat{f}^{EW} - y\|^2 &= \sum_{j=1}^{M} w_j\|f_j - y\|^2 - \sum_{j=1}^{M} w_j\|f_j - \hat{f}^{EW}\|^2 \qquad (59)\\
&= \underbrace{\sum_{j=1}^{M} w_j\hat{r}_j}_{=\hat{r}} - \sum_{j=1}^{M} w_j\|f_j - \hat{f}^{EW}\|^2.
\end{aligned}$$

Note also that, for all i,

$$\sum_{j=1}^{M}\left(\frac{\partial w_j}{\partial Y_i}\right)\hat{f}^{EW}(X_i) = \hat{f}^{EW}(X_i)\frac{\partial}{\partial Y_i}\underbrace{\sum_{j=1}^{M} w_j}_{=1} = 0. \tag{60}$$

Combining (58) – (60) we obtain

$$\begin{aligned}
\hat{R} &= \hat{r} - \sum_{j=1}^{M} w_j\|f_j - \hat{f}^{EW}\|^2 + \frac{2\sigma^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{M}\left(\frac{\partial w_j}{\partial Y_i}\right)f_j(X_i) - \sigma^2\\
&= \hat{r} - \sum_{j=1}^{M} w_j\|f_j - \hat{f}^{EW}\|^2 + \frac{4\sigma^2}{\beta n}\sum_{i=1}^{n}\sum_{j=1}^{M}(f_j(X_i) - \hat{f}^{EW}(X_i))^2 w_j - \sigma^2\\
&= \hat{r} - \sum_{j=1}^{M}\left(1 - \frac{4\sigma^2}{\beta}\right)w_j\|f_j - \hat{f}^{EW}\|^2 - \sigma^2.
\end{aligned}$$

Taking expectations of both sides of this inequality and using (57) we find

$$E\|\hat{f}^{EW} - f\|^2 = E(\hat{r}) - \sigma^2 - \left(1 - \frac{4\sigma^2}{\beta}\right) E\left(\sum_{j=1}^{M} w_j \|f_j - \hat{f}^{EW}\|^2\right)$$

which implies the proposition.

Theorem 15. *For $\beta \geq 4\sigma^2$ we have*

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

In particular, if $\pi_j = \frac{1}{M}$, $j = 1, \ldots, M$,

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

Proof. Propositions 13 and 14, and the fact that $E(\hat{r}_j) = E\|y - f_j\|^2 = \|f - f_j\|^2 + \sigma^2$ imply

$$\begin{aligned} E\|\hat{f}^{EW} - f\|^2 &\leq E(\hat{r}) - \sigma^2 \\ &\leq \min_{1\leq j\leq M}\left(E(\hat{r}_j) + \frac{\beta}{n}\log\frac{1}{\pi_j}\right) - \sigma^2 \\ &= \min_{1\leq j\leq M}\left(\|f - f_j\|^2 + \frac{\beta}{n}\log\frac{1}{\pi_j}\right). \end{aligned}$$

Remarks

1. Theorem 15 is proved in Dalalyan and Tsybakov (2007, 2008) where the result has a more general form:

$$E\|\hat{f}^{EW} - f\|^2 \leq \min_{\lambda\in\Lambda^M}\left(\sum_{j=1}^{M}\lambda_j\|f - f_j\|^2 + \frac{\beta}{n}\mathcal{K}(\lambda,\pi)\right). \tag{61}$$

   Indeed, the right-hand side of (61) does not exceed

$$\min_{\lambda\in\{e_1,\ldots,e_M\}}\left(\sum_{j=1}^{M}\lambda_j\|f - f_j\|^2 + \frac{\beta}{n}K(\lambda,\pi)\right) = \min_{1\leq j\leq M}\left(\|f - f_j\|^2 + \frac{\beta}{n}\log\frac{1}{\pi_j}\right).$$

2. The right-hand side of (61) is reminiscent of the variational interpretation of the exponential weighted estimator. If we replace $r_j = \|f - f_j\|^2$ by $\hat{r}_j = \|y - f_j\|^2$, $\hat{f}^{EW}$ is obtained by the minimization :

$$\min_{\lambda\in\Lambda^M}\left(\sum_{j=1}^{M}\lambda_j\hat{r}_j + \frac{\beta}{n}\mathcal{K}(\lambda,\pi)\right),$$

   which is the empirical analog of the right-hand side of (61).

3. Leung and Barron (2006) have proved a result analogous to Theorem 15 for the case where f_j are not any fixed functions but rather the least squares estimators on linear subspaces of $\mathbb{R}^M$. These estimators are constructed from the same sample y that is used to compute the weights. In their case, the exponential weights are slightly different. Namely, they take

$$w_j = \frac{\exp\left(-\frac{n\hat{r}_j}{\beta} - \frac{\dim(j)}{2}\right)\pi_j}{\sum_{k=1}^M \exp\left(-\frac{n\hat{r}_k}{\beta} - \frac{\dim(k)}{2}\right)\pi_k}$$

where $\dim(j)$ is the dimension of the space on which the jth least squares estimator projects.

6 Sparsity pattern aggregation

In this section, we describe an aggregation procedure that will be shown to achieve universal aggregation. Let $\mathcal{P} = \{0,1\}^M$. We call a *sparsity pattern* any binary vector $p \in \mathcal{P}$. We denote by $|p| \stackrel{\text{def}}{=} |p|_0$ the number of ones in p. To each sparsity pattern $p \in \mathcal{P}$ we associate a linear subspace S^p of $\mathbb{R}^M$:

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

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 random errors $\mathcal{N}(0, 2\sigma^2)$, cf. 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 $y^{(1)}$:

$$\hat{\theta}_p = \underset{\theta \in S^p}{\text{argmin}} \|y^{(1)} - f_\theta\|^2.$$

Set $\hat{r}_p = \|y^{(2)} - f_{\hat{\theta}_p}\|^2$ 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, $\{\pi_p\}$ is a prior probability measure on $\mathcal{P}$ with $\pi_p \geq 0$ (not necessarily $\pi_p > 0$; $\pi_p = 0$ is possible, on the difference from priors in Section 5). Note that $\hat{\theta}^{SPA} \in \mathbb{R}^{2^M}$. The *Sparsity Pattern Aggregate* is defined by

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

From Theorem 15 we get: If $\beta = 8\sigma^2$ (because $\sigma^2 \to 2\sigma^2$ after sample cloning) then

$$\forall f: \quad E\|\hat{f}^{SPA} - f\|^2 \leq \min_{p \in \mathcal{P}, \pi_p \neq 0} \left[E\|f_{\hat{\theta}_p} - f\|^2 + \frac{8\sigma^2}{n} \log \frac{1}{\pi_p} \right]. \tag{62}$$

From Proposition 2,

$$E\|f_{\hat{\theta}_p} - f\|^2 \leq \min_{\theta \in S^p} \|f_\theta - f\|^2 + \frac{2\sigma^2|p|}{n}. \tag{63}$$

Combining (62) and (63), and choosing an appropriate prior π_p we obtain our main result that will be stated below. Namely, we will use the prior

$$\pi_p = \begin{cases} \left(\binom{M}{|p|} e^{|p|} H\right)^{-1} & \text{if } |p| \le R, \\ 1/2 & \text{if } |p| = M, \\ 0 & \text{otherwise,} \end{cases} \tag{64}$$

where $H = 2\sum_{k=0}^{R} e^{-k} \le 2\sum_{k=0}^{\infty} e^{-k} = 2e/(e-1)$. Clearly, $\sum_{p\in\mathcal{P}} \pi_p = 1$. Indeed,

$$\sum_{p\in\mathcal{P},|p|\le R} \pi_p = \sum_{k=0}^{R} \binom{M}{k} \frac{1}{\binom{M}{k} e^k H} = \frac{\sum_{k=0}^{R} e^{-k}}{H} = \frac{1}{2}\,.$$

Definition 16. Exponential Screening *(ES) estimator $\hat{f}^{ES}$ is defined as a sparsity pattern aggregate ($\hat{f}^{SPA}$) with the prior π_p given in (64). The corresponding vector of weights is denoted by $\hat{\theta}^{ES}$.*

Remark. The prior (64) can be called a *sparsity prior* because it downweights exponentially the non-sparse vectors. The only exception is done for the most non-sparse vector (the one with all non-zero components) for which we keep the global least squares estimator with weight 1/2. This point is technical; we introduce it for mathematical convenience in order to simplify the proofs.

From (62) with $|p| = M$ we obtain

$$E\|\hat{f}^{ES} - f\|^2 \le E\|f_{\hat{\theta}^{LS}} - f\|^2 + \frac{8\sigma^2}{n}\log 2$$

where we have used that $\hat{\theta}_p$ for $|p| = M$ coincides with the global least squares estimator $\hat{\theta}^{LS}$. This inequality and Proposition 2 imply:

$$E\|\hat{f}^{ES} - f\|^2 \le \min_{\theta\in\mathbb{R}^M} \|f_\theta - f\|^2 + \frac{2\sigma^2 R}{n} + \frac{8\sigma^2}{n}\log 2. \tag{65}$$

Let $p(\theta) \in \mathcal{P}$ be the sparsity pattern of $\theta \in \mathbb{R}^M$, i.e., a vector with components $p_j(\theta) = 1$ if $\theta_j \neq 0$, and $p_j(\theta) = 0$ otherwise. Note that $|p(\theta)| = |\theta|_0$. Using (62) and (63), we get

$$\begin{aligned} E\|\hat{f}^{ES} - f\|^2 &\le \min_{p\in\mathcal{P}:|p|\le R} \left[\min_{\theta\in S^p} \|f_\theta - f\|^2 + \frac{2\sigma^2|p|}{n} + \frac{8\sigma^2}{n}\log\frac{1}{\pi_p}\right] \\ &\le \min_{\{\theta:p(\theta)=p\}\subset S^p} \min_{p\in\mathcal{P}:|p|\le R} \min_{\theta:p(\theta)=p} \left[\|f_\theta - f\|^2 + \frac{2\sigma^2|p(\theta)|}{n} + \frac{8\sigma^2}{n}\log\left(\frac{1}{\pi_{p(\theta)}}\right)\right] \\ &= \min_{|p(\theta)|=|\theta|_0} \min_{\theta\in\mathbb{R}^M:|\theta|_0\le R} \left[\|f_\theta - f\|^2 + \frac{2\sigma^2|\theta|_0}{n} + \frac{8\sigma^2}{n}\log\left(\frac{1}{\pi_{p(\theta)}}\right)\right]. \end{aligned}$$

Now, we need to bound $\log\frac{1}{\pi_{p(\theta)}}$. We use the following fact:

$$\binom{M}{k} \le \left(\frac{eM}{K}\right)^k.$$

Then

$$\begin{aligned}\log\left(\frac{1}{\pi_{p(\theta)}}\right) &= \log\left(\binom{M}{|p(\theta)|}e^{|p(\theta)|}H\right)\\ &\underset{|p(\theta)|=|\theta|_0}{\leq} |\theta|_0\log\left(\frac{eM}{|\theta|_0\vee 1}\right)+|\theta|_0+\log\left(\frac{2e}{e-1}\right)\\ &\leq 2|\theta|_0\log\left(\frac{eM}{|\theta|_0\vee 1}\right)+\log\left(\frac{2e}{e-1}\right).\end{aligned}$$

Hence,

$$\begin{aligned}E\|\hat{f}^{ES}-f\|^2 &\leq \min_{\theta\in\mathbb{R}^M:|\theta|_0\leq R}\left[\|f_\theta-f\|^2+\frac{2\sigma^2|\theta|_0}{n}+\frac{16\sigma^2}{n}|\theta|_0\log\left(\frac{eM}{|\theta|_0\vee 1}\right)\right]\\ &\quad+\frac{8\sigma^2}{n}\underbrace{\log\left(\frac{2e}{e-1}\right)}_{\leq\log 2}\\ &\leq \min_{\theta\in\mathbb{R}^M:|\theta|_0\leq R}\left[\|f_\theta-f\|^2+\frac{18\sigma^2}{n}|\theta|_0\log\left(\frac{eM}{|\theta|_0\vee 1}\right)\right]+\frac{8\sigma^2}{n}\log 2. \qquad (66)\end{aligned}$$

Combining (65) and (66) we get that, for all $\theta\in\mathbb{R}^M$,

$$E\|\hat{f}^{ES}-f\|^2\leq\|f_\theta-f\|^2+\frac{18\sigma^2}{n}\left(R\wedge|\theta|_0\log\left(\frac{eM}{|\theta|_0\vee 1}\right)\right)+\frac{8\sigma^2}{n}\log 2\,.$$

Therefore, we have proved the following theorem.

Theorem 17. *Let $\hat{f}^{ES}$ be the exponential screening estimator with $\beta=8\sigma^2$. Then, for all f, n, M and $R\leq M\wedge n$ we have*

$$E\|\hat{f}^{ES}-f\|^2\leq\min_{\theta\in\mathbb{R}^M}\left[\|f_\theta-f\|^2+\frac{C\sigma^2}{n}\left(R\wedge|\theta|_0\log\left(\frac{eM}{|\theta|_0}\right)\right)\right]+\frac{C'\sigma^2}{n} \qquad (67)$$

where C, C' are absolute constants and by convention $0\cdot\log\infty=0$.

Remarks.

1. Recall that in Theorem 17 we need σ^2 to be known and we deal throughout these lectures with the case of Gaussian noise.

2. The result of Leung and Barron (2006) can be applied to a get similar oracle inequality for a slightly modified estimator without sample cloning.

3. Theorem 17 immediately implies the optimal rates of L-aggregation and MS-aggregation given in Table 1, as well as the optimal rate $R/n\wedge s\log(eM/s)$ for s-sparse aggregation.

We now show that Theorem 17 also implies the optimal rate for C-aggregation given in Table 1. The proof is based on the following result, which yields a special example of the "Maurey argument".

Lemma 18 ("Maurey Lemma"). *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*

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

Proof. See Section 11.

We now apply Lemma 18 to show that the exponential screening estimator attains the optimal rate of C-aggregation. From Theorem 17 we obtain, for some constant $C > 0$ and any $1 \le m \le M$,

$$\begin{aligned} E\|\hat{f}^{ES} - f\|^2 &\le \min_{\theta \in \mathbb{R}^M} \left[\|f_\theta - f\|^2 + \frac{C\sigma^2}{n} |\theta|_0 \log\Big(\frac{eM}{|\theta|_0}\Big) \right] \\ &\le \min_{\theta : |\theta|_0 \le m} \left[\|f_\theta - f\|^2 + \frac{C\sigma^2}{n} |\theta|_0 \log\Big(\frac{eM}{|\theta|_0}\Big) \right] \\ &\le \min_{\theta : |\theta|_0 \le m} \|f_\theta - f\|^2 + \frac{C\sigma^2 m}{n} \log\Big(\frac{eM}{m}\Big) \end{aligned} \tag{68}$$

where we have used the monotonicity of $x \mapsto x \log(eM/x)$. This and Lemma 18 imply

$$E\|\hat{f}^{ES} - f\|^2 \le \min_{\theta \in \Lambda^M} \|f_\theta - f\|^2 + \frac{L^2}{m} + \frac{C\sigma^2 m}{n} \log\Big(\frac{eM}{m}\Big)\,.$$

Choosing here

$$m = \left\lceil \frac{c' L}{\sigma} \sqrt{\frac{n}{\log(1 + c''(\sigma/L)M/\sqrt{n})}} \right\rceil$$

with suitable constants $c' > 0, c'' > 0$ we obtain the following result.

Corollary 19. *If $\|f_j\| \le 1$, $j = 1, \dots, M$, then there exists a constant $C > 0$ such that*

$$E\|\hat{f}^{ES} - f\|^2 \le \min_{\theta \in \Lambda^M} \|f_\theta - f\|^2 + C \left[\frac{\sigma^2 R}{n} \wedge \sigma \sqrt{\frac{1}{n} \log\Big(1 + \frac{M\sigma}{\sqrt{n}}\Big)} \right].$$

The remainder term in this oracle inequality is the optimal rate of C-aggregation given in Table 1.

7 Sparsity via Exponential Screening

The exponential screening estimator also solves the questions raised in Scenario 1 (sparse linear regression). To illustrate this, we use the following theorem proved in Rigollet and Tsybakov (2011); it is obtained as a corollary of Theorem 17 using Lemma 18 as explained above and refining the constants (we also note that $\hat{f}^{ES}$ is defined in a slightly different way in Rigollet and Tsybakov (2011), cf. Remark 3 at the end of Section 5, but we still use the same notation for this estimator in this section).

Theorem 20. *If $\|f_j\| \le 1$, $j = 1, \dots, M$, then, for any $M \ge 1, n \ge 1$,*

$$E\|\hat{f}^{ES} - f\|^2 \le \min_{\theta \in \mathbb{R}^M} \left(\|f_\theta - f\|^2 + \varphi_{n,M}(\theta) \right) + \frac{\sigma^2}{n}(9 \log(1 + eM) + 4 \log 2)$$

where the rate term $\varphi_{n,M}(\theta)$ is equal to

$$\frac{9\sigma^2|\theta|_0}{n} \log\left(\frac{M}{|\theta|_0 \vee 1}\right) \wedge \frac{11\sigma|\theta|_1}{\sqrt{n}} \sqrt{\log\left(1 + \frac{3eM\sigma}{|\theta|_1\sqrt{n}}\right)} \wedge \frac{\sigma^2 R}{n}$$

with $R = \operatorname{rank}(X)$.

In particular, if the model is linear, i.e., there exists $\theta^* \in \mathbb{R}^M$ such that $f = f_{\theta^*}$, we have

$$E|X(\hat{\theta}^{ES} - \theta^*)|_2^2/n \le \psi_{n,M}(\theta^*) + \frac{8\sigma^2 \log 2}{n}$$

where $\psi_{n,M}(\theta)$ is the minimum of 4 terms

$$\frac{9\sigma^2|\theta|_0}{n} \log\left(\frac{eM}{|\theta|_0 \vee 1}\right) \wedge \frac{11\sigma|\theta|_1}{\sqrt{n}} \sqrt{\log\left(1 + \frac{3eM\sigma}{|\theta|_1\sqrt{n}}\right)} \wedge 4|\theta|_1^2 \wedge \frac{\sigma^2 R}{n}.$$

If the model is linear we have

$$E|X(\hat{\theta}^{ES} - \theta^*)|_2^2/n \quad \le \quad C\delta_{n,M}(\theta^*)$$

where

$$\delta_{n,M}(\theta) = \frac{\sigma^2|\theta|_0}{n} \log\left(\frac{eM}{|\theta|_0 \vee 1}\right) \wedge \frac{\sigma|\theta|_1}{\sqrt{n}} \sqrt{\log\left(1 + \frac{\sigma M}{|\theta|_1\sqrt{n}}\right)} \wedge \frac{\sigma^2 R}{n} \wedge |\theta|_1^2. \tag{69}$$

This implies the following upper bound on the rate of sparse estimation:

$$\inf_{\hat{\theta}} \sup_{\theta \in B_0(s)} E_\theta |X(\hat{\theta}^{ES} - \theta)|_2^2/n \quad \le \quad C\sigma^2 \left(\frac{s}{n} \log\left(\frac{eM}{s}\right) \wedge \frac{R}{n} \right). \tag{70}$$

Here, 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 $\xi_i \sim \mathcal{N}(0, \sigma^2)$.

Additionally, we get a bound for the minimax risk on the ℓ_1-ball $B_1(\delta) = \{\theta : |\theta|_1 \le \delta\}$:

$$\inf_{\hat{\theta}} \sup_{\theta \in B_1(\delta)} E_\theta |X(\hat{\theta}^{ES} - \theta)|_2^2/n \quad \le \quad C \left(\frac{\sigma\delta}{\sqrt{n}} \sqrt{\log\left(1 + \frac{\sigma M}{\delta\sqrt{n}}\right)} \wedge \frac{\sigma^2 R}{n} \wedge \delta^2 \right). \tag{71}$$

More generally, (69) yields the following rate of estimation over the intersection of ℓ_0 and ℓ_1 balls $B_0(s) \cap B_1(\delta)$:

$$\inf_{\hat{\theta}} \sup_{\theta \in B_0(s) \cap B_1(\delta)} E_\theta |X(\hat{\theta}^{ES} - \theta)|_2^2/n \le C \left(\frac{\sigma^2 s}{n} \log\left(\frac{eM}{s}\right) \wedge \frac{\sigma\delta}{\sqrt{n}} \sqrt{\log\left(1 + \frac{\sigma M}{\delta\sqrt{n}}\right)} \wedge \frac{\sigma^2 R}{n} \wedge \delta^2 \right). \tag{72}$$

It is easy to see that if the assumption $\max_{1 \le j \le M} \|f_j\| \le 1$ is replaced by $\max_{1 \le j \le M} \|f_j\| \le L$, this bound remains valid with δ replaced by $L\delta$. As shown in Rigollet and Tsybakov (2011) (cf. also Section 10 below) the rate in (72) is optimal in a minimax sense.

Remarks.

1. On the difference from the Lasso and related techniques, the ES estimator satisfies the Sparsity Oracle Inequality **under no assumption on the dictionary** $\{f_1, \dots, f_M\}$.

2. The ES estimator simultaneously takes advantage of three types of sparsity:

   - small number of non-zero entries of θ (ℓ_0 norm),
   - small global weight (ℓ_1 norm),
   - small rank of matrix X.

3. Donoho and Johnstone (1992) studied *asymptotics* of the minimax risk, as $n \to \infty$, over $B_0(s)$ for the Gaussian sequence model, i.e., under the assumption that $M = n$, $X^T X/n = I_n$. They showed that, as $M = n \to \infty$ and s stays fixed,

$$\inf_{\hat\theta} \sup_{\theta \in B_0(s)} E_\theta |\hat\theta - \theta|_2^2/n \quad \sim \quad 2\sigma^2 \frac{s}{n} \log\left(\frac{n}{s}\right). \tag{73}$$

   We extend this result to the general linear model with $M \neq n$ and show that non-asymptotically (but without an exact constant):

$$\inf_{\hat\theta} \sup_{\theta \in B_0(s)} E_\theta |X(\hat\theta - \theta)|_2^2/n \quad \le \quad C\sigma^2 \left(\frac{s}{n} \log\left(\frac{eM}{s}\right) \wedge \frac{R}{n}\right). \tag{74}$$

   The difference of (74) from (73) is that, for the general case, the rank R of matrix X and the dimension M appear in the rate. The rate is substantially better for low-rank matrices X.

4. Donoho and Johnstone (1994), Abramovich et al. (2006) studied *asymptotics* of the minimax risk, as $n \to \infty$, over $B_q(\delta)$ for $q > 0$, again for the Gaussian sequence model, as in the previous item. We can compare (71) with their result for $q = 1$; they obtain

$$\inf_{\hat\theta} \sup_{\theta \in B_1(\delta)} E_\theta |\hat\theta - \theta|_2^2/n \quad \sim \quad \frac{\delta\sigma}{\sqrt{n}} \sqrt{2 \log\left(\frac{\sigma\sqrt{n}}{\delta}\right)} \quad \text{as } M = n \to \infty\,.$$

   Our non-asymptotic result for the Gaussian sequence model is of the form

$$\inf_{\hat\theta} \sup_{\theta \in B_1(\delta)} E_\theta |\hat\theta - \theta|_2^2/n \quad \asymp \quad \frac{\delta\sigma}{\sqrt{n}} \sqrt{\log\left(1 + \frac{\sigma\sqrt{n}}{\delta}\right) \wedge \delta^2}.$$

   So, in this case again, we find an additional effect that does not appear in the asymptotic result.

5. For the regression model with $R = n$, $M \ge n$, Raskutti et al. (2011) studied *asymptotics* of the minimax risk on $B_0(s)$ and $B_1(\delta)$ when M, n, s tend to ∞ in a specified way. They obtain that the rates $\frac{s}{n} \log\left(\frac{M}{s}\right)$ and $\delta\sqrt{\frac{\log M}{n}}$ respectively are optimal under the considered asymptotic regime for M, n and s. We see that several effects appearing in our bounds are shadowed by the specific asymptotics. Note also that Raskutti et al. (2011) use different estimators depending on s and δ to achieve the asymptotic rates separately on $B_0(s)$ and on $B_1(\delta)$ but not on their intersection. Conversely, our estimator is adaptive to s and δ and is (non-asymptotically) minimax optimal simultaneously for all s and δ.

8 Minimax estimation over intersection of ℓ_0 and ℓ_q balls

In this section, we show that Theorem 17 is powerful enough to guarantee the minimax optimality of the exponential screening estimator not only on the intersection of ℓ_0 and ℓ_1 balls but also on the intersection of ℓ_0 and ℓ_q balls with $0 < q \le 2$. We will consider separately the cases $0 < q < 1$ and $1 < q \le 2$. For $0 < q < 1$, we have the following ℓ_q analog of Lemma 18.

Lemma 21. *Let $\|f_j\| \le L$, $j = 1, \dots, M$, and $1 \le m \le M$. Then, for any f, any $0 < q < 1$ and any $\theta \in \mathbb{R}^M$ there exists $\bar\theta \in \mathbb{R}^M$ such that $|\bar\theta|_0 \le 2m$ and*

$$\|f_{\bar\theta} - f\|^2 \le \|f_\theta - f\|^2 + \left(\frac{q}{1-q}\right)^2 L^2 |\theta|_q^2 m^{1-2/q}\,.$$

Proof. By Lemma 18, 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

$$\|f_{\theta'} - h\|^2 \le \|f_{\theta''} - h\|^2 + \frac{|\theta''|_1^2 L^2}{m}\,. \tag{75}$$

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 (75) with $\theta'' = \theta_{J^c}$, $h = f - f_{\theta_J}$ where $\theta_J = (\theta_j I(j \in J),\ j = 1, \dots, M)$. Then (75) takes the form

$$\|f_{\theta' + \theta_J} - f\|^2 \le \|f_\theta - f\|^2 + \frac{|\theta_{J^c}|_1^2 L^2}{m}\,. \tag{76}$$

Set $\bar\theta = \theta' + \theta_J$. By construction, $|\bar\theta|_0 \le 2m$. To complete the proof, we evaluate the norm $|\theta_{J^c}|_1$.

Let $|\theta|_{(j)}$ denote the jth largest absolute value of the components of θ. We have

$$|\theta|_{(j)} \le \frac{|\theta|_q}{j^{1/q}} \tag{77}$$

since otherwise $\sum_{k=1}^j |\theta|_{(k)}^q \ge j|\theta|_{(j)}^q > |\theta|_q^q$, which is impossible. Using (77) we find

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

Plugging this bound in (76) yields the lemma.

Lemmas 18 and 21 combined with Theorem 17 imply the following result.

Theorem 22. *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 there exists a constant $C_q > 0$ such that, for any $\delta > 0$,*

$$E\|\hat f^{ES} - f\|^2 \le \min_{\theta \in B_q(\delta)} \|f_\theta - f\|^2 + C_q \psi_{n,M}(B_q(\delta)) \tag{78}$$

where

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

Furthermore, if the model is linear, there exists a constant $C_q' > 0$ such that, for any $s \in \{1, \dots, M\}$ and $\delta > 0$,

$$\inf_{\hat{\theta}} \sup_{\theta \in B_0(s) \cap B_q(\delta)} E_\theta |X(\hat{\theta}^{ES} - \theta)|_2^2 / n \le C_q' \left(\psi_{n,M}(B_q(\delta)) \wedge \frac{\sigma^2 s}{n} \log\left(\frac{eM}{s}\right) \right). \tag{79}$$

Here, 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 $\xi_i \sim \mathcal{N}(0, \sigma^2)$.

Proof. We consider only the case $0 < q < 1$; the case $q = 1$ is treated similarly to Corollary 19. Arguing as in (68), we find from Theorem 17 that, for some constant $C > 0$ and any $1 \le m \le M$,

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

where we have used the monotonicity of the mapping $x \mapsto x \log(eM/x)$. This and Lemma 21 imply

$$E\|\hat{f}^{ES} - f\|^2 \le \min_{\theta \in B_q(\delta)} \|f_\theta - f\|^2 + C\left(\delta^2 m^{1-2/q} + \frac{\sigma^2 m}{n} \log\left(\frac{eM}{m}\right) \right). \tag{80}$$

The remainder terms on the right hand side of (80) are balanced by choosing

$$m = \left\lceil c' \left(\frac{\delta}{\sigma}\right)^q \left(\frac{n}{\log\left(1 + c'' \left(\frac{\sigma}{\delta}\right)^q \frac{M}{n^{q/2}}\right)} \right)^{q/2} \right\rceil$$

with suitable constants $c' > 0, c'' > 0$. Plugging this m in (80) and using again Theorem 17 to include the minimum with remainder term of the order $\sigma^2 R/n$ we obtain (78). Finally, (79) is straightforward in view of (79) and (70).

It can be shown that the remainder terms in the upper bounds of Theorem 22 are optimal in a minimax sense. Theorem 22 allows not only to prove that the estimator $\hat{\theta}^{ES}$ adaptively attains minimax rates over ℓ_q-balls with $0 < q \le 1$ (cf. (79)) but also to show that it achieves optimal rates of aggregation on such balls (cf. (78)). For $1 < q \le 2$ we are only able to show that $\hat{\theta}^{ES}$ accomplishes the first task – minimax rates over ℓ_q-balls and for this we will need an additional assumption on the dictionary. Let $\lambda_{\max}(X^T X/n)$ denote the maximal eigenvalue of matrix $X^T X/n$. We will assume that $\lambda_{\max}(X^T X/n) \le L^2$, which is a more restrictive condition than $\|f_j\| \le L$, $j = 1, \dots, M$.

Theorem 23. *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 $\beta = 8\sigma^2$. Let the model be linear, $y = X\theta + \xi$ where ξ is a Gaussian vector in $\mathbb{R}^n$ with i.i.d. components $\xi_i \sim \mathcal{N}(0, \sigma^2)$. Then there exists a constant $C_q' > 0$ such that, for any $s \in \{1, \dots, M\}$ and $\delta > 0$,*

$$\inf_{\hat{\theta}} \sup_{\theta \in B_0(s) \cap B_q(\delta)} E_\theta |X(\hat{\theta}^{ES} - \theta)|_2^2 / n \le C_q' \left(\psi_{n,M}(B_q(\delta)) \wedge \frac{\sigma^2 s}{n} \log\left(\frac{eM}{s}\right) \right). \tag{81}$$

where $\psi_{n,M}(B_q(\delta))$ is defined in Theorem 22.

Proof. It is enough 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

$$|X(\bar{\theta} - \theta)|_2^2/n = \|f_{\bar{\theta}} - f_\theta\|^2 \le c_q L^2 |\theta|_q^2 m^{1-2/q} \tag{82}$$

for a constant $c_q > 0$ depending only on q. This replaces Lemma 21 when the model is linear, i.e., $f = f_\theta$. The rest of the proof follows the same lines as that of (79).

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 $\|f_{\bar{\theta}} - f_\theta\|^2 \le \lambda_{\max}(X^T X/n)|\bar{\theta} - \theta|_2^2 \le L^2 |\theta_{J^c}|_2^2$. Using (77) we obtain

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

Thus, (82) holds with $c_q = q(2-q)^{-1}$.

9 Universal aggregation

Along with the three main types of aggregation (MS, C, L), two other choices of Θ have been proposed and analyzed in the literature (Bunea, Tsybakov and Wegkamp (2007), Lounici (2007)) corresponding to s-sparse aggregation (L_s) and to convex s-sparse aggregation (C_s). Again, these classes Θ can be understood in terms of intersection of ℓ_0 and ℓ_1 balls as summarized in the next table.

Problem	Θ	Description
(MS)	$\Theta_{(\mathrm{MS})} \subset B_0(1) \cap B_1(1)$	Best in dictionary
(C)	$\Theta_{(\mathrm{C})} \subset B_1(1)$	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)	$\Theta_{(\mathrm{C}_s)} \subset B_0(s) \cap B_1(1)$	Best s-sparse convex combination

Table 2.

The next theorem (Rigollet and Tsybakov, 2011) easily follows from the argument in Sections 6 and 7 and shows that the exponential screening estimator enjoys the property of universal aggregation, i.e., it attains optimal rates of aggregation simultaneously on all these classes Θ.

Theorem 24. *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*

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

Here, $C > 0$ *is a numerical constant and the remainder term is the optimal rate of aggregation on* Θ *which has the form*

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

where $\psi^*_{n,M}(\Theta)$ *is given in Table 3.*

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

Table 3.

Note that if the assumption $\max_{1\le j\le M} \|f_j\| \le 1$ in Theorem 24 is replaced by $\max_{1\le j\le M} \|f_j\| \le L$, the rates in this table remain valid with $\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{eM\sigma}{\sqrt{n}}\right)}$ replaced by $L\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{eM\sigma}{L\sqrt{n}}\right)}$.

Remark a connection between aggregation and minimax estimation on the intersection of ℓ_0 and ℓ_1 balls. Indeed, for the both problems upper bounds for the risk or excess risk are attained on one and the same estimator, which is the exponential screening estimator. Also the rates of convergence for aggregation in Theorem 24 are similar to rates of minimax estimation on the intersection of the corresponding ℓ_0 and ℓ_1 balls (cf. Section 7).

10 Minimax lower bounds

In this section, we show that the rates in the bounds for minimax risk on the intersection of ℓ_0 and ℓ_1 balls (cf. Section 7) and the rates of aggregation obtained in Theorem 24 are optimal in a minimax sense. Note first that it suffices to prove the lower bounds for the minimax risk only, since obviously

$$\inf_{\hat f} \sup_{f} \left(E\|\hat f - f\|^2 - \min_{\theta\in\Theta} \|f_\theta - f\|^2 \right)$$

$$\ge \inf_{\hat f} \sup_{\theta\in\Theta} E_\theta \|\hat f - f_\theta\|^2$$

while the sets Θ used in aggregation settings are either ℓ_0 balls or intersections of ℓ_0 balls with the ℓ_1 simplex, which is undistinguishable from the ℓ_1 ball in what concerns the minimax rates. In fact, it is enough for our purposes to prove minimax lower bounds on the sets

$$\Theta_{\delta,s} = \delta\Lambda_M \cap B_0(s), \quad s = 1,\dots,M, \quad \delta > 0.$$

Theorem 25. *Let $M \ge 2, n \ge 1, 1 \le s \le M$, $M \le n$, and $\delta \ge c_*\sigma/\sqrt{n}$ for some constant $c_* > 0$. Then there exists a dictionary $f_1,\dots,f_M$ with $\max_{1\le j\le M} \|f_j\| \le 1$ such that*

$$\inf_{\hat f} \sup_{\theta\in\Theta_{\delta,s}} E_\theta \|\hat f - f_\theta\|^2 \ge C\psi_{n,M}(\delta, s)$$

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

$$\psi_{n,M}(\delta, s) = \frac{\sigma^2 R}{n} \wedge \delta\sqrt{\frac{\sigma^2}{n}\log\left(1 + \frac{eM\sigma}{\delta\sqrt{n}}\right)} \wedge \frac{\sigma^2 s}{n}\log\left(\frac{eM}{s}\right) \wedge \delta^2.$$

Proof. (To be inserted.) A more general lower bound is given in Rigollet and Tsybakov (2011).

Note that the lower bound of Theorem 25 matches the upper bound (72) while in the particular cases corresponding to the five aggregation problems it implies that the upper bounds of Theorem 24 cannot be improved. If we assume that $\max_{1\le j\le M}\|f_j\| \le L$ instead of $\max_{1\le j\le M}\|f_j\| \le 1$, the lower bound of Theorem 25 remains valid with δ replaced by $L\delta$.

11 Proof of "Maurey Lemma"

If $\theta = 0$, the result of the lemma is obvious. So, it is enough to consider $\theta \neq 0$. Set $p_j = |\theta_j|/|\theta|_1, j = 1, \dots, M$. Let $\eta_1, \dots, \eta_m$ be i.i.d. random variables such that η_m takes value $j \in \{1, \dots, M\}$ with probability p_j. Set

$$\chi_j = \sum_{s=1}^{m} I(\eta_s = j).$$

Denote $\mathbb{E}$ the expectation with respect to the joint distribution of $\eta_1, \dots, \eta_m$. We have $\mathbb{E}(\chi_j) = mp_j$. Define a random vector $\bar\theta \in \mathbb{R}^M$ with components $\bar\theta_j = \chi_j \mathrm{sign}(\theta_j)|\theta|_1/m$ (with $\mathrm{sign}(0) \stackrel{\mathrm{def}}{=} 0$). Then

$$\mathbb{E}(\bar\theta_j) = \theta_j,$$

so that, for any x,

$$\mathbb{E}(f_{\bar\theta}(x)) = \mathbb{E}\left(\sum_{j=1}^{M} \bar\theta_j f_j(x)\right) = f_\theta(x).$$

Now, for any x, the variance of $f_{\bar\theta}(x)$ is given by

$$\begin{aligned}
\mathrm{Var}(f_{\bar\theta}(x)) &= \mathrm{Var}\left(\sum_{j=1}^{M} \bar\theta_j f_j(x)\right) \\
&= \mathrm{Var}\left(\sum_{j=1}^{M} \chi_j \mathrm{sign}(\theta_j)|\theta|_1 f_j(x)/m\right) \\
&= \frac{|\theta|_1^2}{m^2}\mathrm{Var}\Big(\sum_{s=1}^{m} \zeta_s\Big)
\end{aligned}$$

where $\zeta_s = \sum_{j=1}^{M} I(\eta_s = j)\mathrm{sign}(\theta_j) f_j(x)$, $s = 1, \dots, m$, are independent random variables. Thus,

$$\mathrm{Var}(f_{\bar\theta}(x)) = \frac{|\theta|_1^2}{m}\mathrm{Var}(\zeta_1) \le \frac{|\theta|_1^2}{m}\mathbb{E}(\zeta_1^2).$$

Here

$$\begin{aligned}
\mathbb{E}(\zeta_1^2) &= \mathbb{E}\left[\left(\sum_{j=1}^{M} I(\eta_1 = j)\mathrm{sign}(\theta_j) f_j(x)\right)^2\right] \\
&= \sum_{k=1}^{M} p_k(\mathrm{sign}(\theta_k) f_k(x))^2 = \frac{1}{|\theta|_1}\sum_{k=1}^{M} |\theta_k| f_k^2(x).
\end{aligned}$$

Combining the last two displays we find

$$\mathrm{Var}(f_{\bar\theta}(x)) \le \frac{|\theta|_1}{m}\sum_{k=1}^M |\theta_k| f_k^2(x).$$

By the bias-variance decomposition of the squared risk, for any x,

$$\begin{aligned}\mathbb{E}[(f_{\bar\theta}(x)-f(x))^2] &= \mathbb{E}[(f_{\bar\theta}(x)-f_\theta(x))^2] + (f_\theta(x)-f(x))^2 \\ &= \mathrm{Var}(f_{\bar\theta}(x)) + (f_\theta(x)-f(x))^2.\end{aligned}$$

Using this for $x = X_i, \quad i = 1,\dots,n$, and averaging out we obtain

$$\begin{aligned}\mathbb{E}\|f_{\bar\theta}-f\|^2 &\le \|f_\theta - f\|^2 + \frac{1}{n}\sum_{i=1}^n\left[\frac{|\theta|_1}{m}\sum_{k=1}^M |\theta_k| f_k^2(X_i)\right] \\ &= \|f_\theta - f\|^2 + \frac{|\theta|_1}{m}\sum_{k=1}^M |\theta_k|\frac{1}{n}\sum_{i=1}^n f_k^2(X_i) \\ &\le \|f_\theta - f\|^2 + \frac{|\theta|_1^2}{m}L^2.\end{aligned}$$

Since $\bar\theta$ takes values in a finite set of vectors (let us denote this set by $\bar\Theta$), we have

$$\mathbb{E}\|f_{\bar\theta}-f\|^2 \ge \min_{\theta\in\bar\Theta}\|f_\theta - f\|^2 = \|f_{\theta'}-f\|^2$$

for some $\theta' \in \bar\Theta$. Finally, note that $|\bar\theta|_0 \le m$, so that $|\theta'|_0 \le m$.

12 Some inequalities for Gaussian random variables

Lemma 26. *Let $\eta \sim \mathcal{N}(0,1)$. Then, for all $x > 0$,*

$$P(|\eta| > x) \le \sqrt{\frac{2}{\pi}}\frac{e^{-x^2/2}}{x}, \tag{83}$$

$$E\big[\eta^2 I(|\eta| > x)\big] \le \sqrt{\frac{2}{\pi}}\Big(x + \frac{2}{x}\Big)e^{-x^2/2}. \tag{84}$$

Proof. Inequality (83) follows from the fact that

$$\forall x > 0: \quad \int_x^\infty e^{-u^2/2}du \le \int_x^\infty \frac{u}{x}e^{-u^2/2}du = \frac{1}{x}e^{-x^2/2}.$$

The proof of (84) is similar :

$$E\big[\eta^2 I(|\eta| > x)\big] = \frac{2}{\sqrt{2\pi}}\int_x^\infty u^2 e^{-u^2/2}du \le \sqrt{\frac{2}{\pi}}\int_x^\infty \frac{u^3}{x}e^{-u^2/2}du = \sqrt{\frac{2}{\pi}}\Big(x+\frac{2}{x}\Big)e^{-x^2/2}.$$

Lemma 27. *Let $\eta_1, \dots, \eta_M, M \geq 2$, be $\mathcal{N}(0,1)$ random variables. Then*

$$\forall t > \sqrt{2}: \quad P\Big(\max_{1\leq j\leq M} |\eta_j| > t\sqrt{\log M}\Big) \leq M^{1-t^2/2}.$$

and

$$\forall u > 0: \quad P\Big(\max_{1\leq j\leq M} |\eta_j| > \sqrt{2\log M} + u\Big) \leq \frac{1}{\sqrt{\pi \log M}} e^{-u^2/2}.$$

Proof. Using the union bound we get

$$\begin{aligned} P\Big(\max_{1\leq j\leq M} |\eta_j| > t\sqrt{\log M}\Big) &\leq \sum_{j=1}^{M} P(|\eta_j| > t\sqrt{\log M}) \\ &= M P(|\eta_1| > t\sqrt{\log M}) \end{aligned}$$

We now apply Lemma 26 for $x = t\sqrt{\log M} \geq \sqrt{2\log M} \geq \sqrt{2\log 2}$. This gives the bound

$$P(|\eta_1| > t\sqrt{\log M}) \leq \frac{1}{\sqrt{\pi \log 2}}\, e^{-\frac{t^2 \log M}{2}} < M^{-t^2/2}.$$

Similarly, for any $u > 0$,

$$P\Big(\max_{1\leq j\leq M} |\eta_j| > \sqrt{2\log M} + u\Big) \leq \frac{M}{\sqrt{\pi \log M}}\, e^{-(\sqrt{2\log M}+u)^2/2} \leq \frac{e^{-u^2/2}}{\sqrt{\pi \log M}}\,.$$

Lemma 28. *Let $\sigma > 0, M \geq 2$ and $Y_1, \dots, Y_M$ be random variables such that*

$$\forall s > 0, \forall i, \quad E(e^{sY_i}) \leq e^{s^2\sigma^2/2}.$$

Then

$$E\left(\max_{1\leq i\leq M} Y_i\right) \leq \sigma\sqrt{2\log M}.$$

If, in addition, $E(e^{-sY_i}) \leq e^{s^2\sigma^2/2}$, then

$$E\left(\max_{1\leq i\leq M} |Y_i|\right) \leq \sigma\sqrt{2\log(2M)}.$$

Proof. We first use Jensen's inequality for the convex function $x \mapsto \exp(sx)$:

$$\begin{aligned} \forall s > 0, \quad \exp\left(sE\left(\max_{1\leq i\leq M} Y_i\right)\right) &\leq E\left(\exp\left(s \max_{1\leq i\leq M} Y_i\right)\right) \\ &= E\left(\max_{1\leq i\leq M} \exp(sY_i)\right) \\ &\leq \sum_{i=1}^{M} E(\exp(sY_i)) \\ &\leq M e^{s^2\sigma^2/2}. \end{aligned}$$

This implies

$$\forall s > 0, \quad E\left(\max_{1\leq i\leq M} Y_i\right) \leq \frac{\log M}{s} + \frac{s\sigma^2}{2}.$$

Putting here $s = \sqrt{\frac{2\log M}{\sigma^2}}$ we get the first inequality. The proof of the second one is analogous.

In particular, the results of Lemma 28 hold for $Y_i \sim \mathcal{N}(0, \sigma^2)$. Note that in Lemmas 27 and 28 we do not assume independence of $\eta_1, \dots, \eta_M$ or of $Y_1, \dots, Y_M$.

13 A lemma about hard thresholding

Lemma 29. *Let $r > 0$ and $y, \theta \in \mathbb{R}$ be such that $|y - \theta| \le r$. Then*

$$\Big| yI(|y| > 2r) - \theta \Big| \le 3\min(|\theta|, r).$$

Proof. Since $|y - \theta| \le r$,

$$I(|y| > 2r) = I(|y| > 2r)I(|\theta| > r).$$

Thus,

$$\begin{aligned} |yI(|y| > 2r) - \theta| &= |yI(|y| > 2r)I(|\theta| > r) - \theta| \\ &\le |(yI(|y| > 2r) - \theta)I(|\theta| > r)| + |\theta| I(|\theta| \le r). \end{aligned}$$

On the other hand, for $|y - \theta| \le r$,

$$|yI(|y| > 2r) - \theta| \le |y - \theta| + |yI(|y| \le 2r)| \le 3r.$$

This implies

$$\begin{aligned} |yI(|y| > 2r) - \theta| &\le 3rI(|\theta| > r) + |\theta| I(|\theta| \le r) \\ &\le 3\Big(rI(|\theta| > r) + |\theta| I(|\theta| \le r) \Big) \\ &= 3\min(|\theta|, r). \end{aligned}$$

References

[1] Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, **34** 584–653.

[2] Baraniuk, R., Davenport, M., DeVore, R. and Wakin, M. (2008). A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, **28** 253– 263.

[3] Birgé, L. and Massart, P. (2007) Minimal penalties for Gaussian model selection. *Probab. Theory Rel. Fields*, **138** 33–73.

[4] Bunea, F., Tsybakov, A. B. and Wegkamp, M. (2004). Aggregation for regression learning. Available from: `arXiv:math.ST/0410214` and `https://hal.ccsd.cnrs.fr/ccsd-00003205`.

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

[6] Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization.* Lecture Notes in Mathematics 1851. Springer-Verlag, Berlin. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, 2001.

[7] Dai, D, Rigollet, P, and Zhang, T. (2012) Deviation Optimal Learning using Greedy Q-aggregation. To appear in *Annals of Statistics*. `arXiv:1203.2507`.

[8] Dalalyan, A. S. and Salmon, J. (2011). Sharp oracle inequalities for aggregation of affine estimators. To appear in *Annals of Statistics*. `arXiv:1104.3969`.

[9] Dalalyan, A. and Tsybakov, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. *Proceedings of the 20th Annual Conference on Learning Theory (COLT 2007), Lecture Notes in Artificial Intelligence* v.4539 (N.H. Bshouty and C.Gentile, eds.), Springer-Verlag, Berlin-Heidelberg, 97–111. `http://www.crest.fr/ckfinder/userfiles/files/Pageperso/tsybakov/DTcolt2007.pdf`

[10] Dalalyan, A. and Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning* **72**, 39–61. `arxiv:0803.2839`

[11] Dalalyan, A. and Tsybakov, A. B. (2012a). Mirror averaging with sparsity priors. *Bernoulli* **18**, 914–944. `arxiv:1003.1189`.

[12] Dalalyan, A. and Tsybakov, A.B. (2012b) Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences* **78**, 1423–1443. `arXiv:0903.1223`

[13] Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54**, 41–81. With discussion and a reply by the authors.

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

[15] Gerchinovitz, S. (2011). Prediction of individual sequences and prediction in the statistical framework : some links around sparse regression and aggregation techniques. PhD thesis, Université Paris Sud. `http://www.math.ens.fr/~gerchino/docs/These-Gerchinovitz.pdf`

[16] Giraud, C. (2008). Mixing least-squares estimators when the variance is unknown. *Bernoulli* **14** 1089–1107.

[17] Johnstone, I. (2013) Gaussian estimation: Sequence and wavelet models. Draft of a book. `http://www-stat.stanford.edu/~imj/GE06-11-13.pdf`

[18] Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, **39** 2302–2329. `arXiv:1011.6256`

[19] Lecué, G. (2011). *Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis*. Habilitation à diriger des recherches. Université de Marne-la-Vallée. `http://perso-math.univ-mlv.fr/users/lecue.guillaume/`

[20] Lecué, G. (2012). Empirical risk minimization is optimal for the Convex aggregation problem. *Bernoulli*, to appear. `http://perso-math.univ-mlv.fr/users/lecue.guillaume/L1.pdf`

[21] Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, **52** 3396–3410.

[22] Littlestone, M. and Warmuth, M. (1994) The weighted majority algorithm. *Information and Computation*, **108** 212–261.

[23] Lounici, K. (2007). Generalized mirror averaging and D-convex aggregation. *Mathematical Methods of Statistics*, **16** 246–259.

[24] Nemirovski, A. (2000) *Topics in non-parametric statistics.* Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998. Lecture Notes in Mathematics, v. 1738. Springer, New York.

[25] Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over l_q -balls. *IEEE Trans. Inform. Th.* **57**, 6976–6994.

[26] Rigollet, P. (2006). Oracles inequalities, aggregation and adaptation. PhD thesis, Université Paris 6. `http://tel.archives-ouvertes.fr/docs/00/11/54/94/PDF/these.pdf`

[27] Rigollet, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models. *Annals of Statistics*, **40**, 639-665.

[28] Rigollet, P. and Tsybakov, A. B. (2007) Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, **16** 260–280. `http://arxiv.org/pdf/math/0605292.pdf`

[29] Rigollet, P. and Tsybakov, A. B. (2011). Exponential Screening and optimal rates of sparse estimation. *Annals of Statistics*, **39** 731-771.

[30] Rigollet, P. and Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. To appear in *Statistical Science*, **27**, 558–575.

[31] Tsybakov, A. B. (2003). Optimal rates of aggregation. In *COLT* (B. Schölkopf and M. K. Warmuth, eds.), vol. 2777 of *Lecture Notes in Computer Science*. Springer, 303–313.
`http://www.crest.fr/ckfinder/userfiles/files/Pageperso/tsybakov/colt.pdf`

[32] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer, New York.

[33] Verzelen, N. (2012) Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.*, **6**, 38–90.

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

[35] Wang, Z., Paterlini, S., Gao, F., and Yang, Y. (2011) Adaptive minimax estimation over sparse ℓ_q-hulls. `arXiv:1108.1961`

[36] Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli*, **10** 25–47.

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