Hyper-sparse optimal aggregation

Stéphane Gaïffas\textsuperscript{1,3} and Guillaume Lecué\textsuperscript{2,3}

December 8, 2009

Abstract

In this paper, we consider the problem of hyper-sparse aggregation. Namely, given a dictionary $F = \{f_1, \ldots, f_M\}$ of functions, we look for an optimal aggregation algorithm that writes $\tilde{f} = \sum_{j=1}^{M} \theta_j f_j$ with as many zero coefficients $\theta_j$ as possible. This problem is of particular interest when $F$ contains many irrelevant functions that should not appear in $\tilde{f}$. We provide an exact oracle inequality for $\tilde{f}$, where only two coefficients are non-zero, that entails $\tilde{f}$ to be an optimal aggregation algorithm. Since selectors are suboptimal aggregation procedures, this proves that 2 is the minimal number of elements of $F$ required for the construction of an optimal aggregation procedures in every situations. A simulated example of this algorithm is proposed on a dictionary obtained using LARS, for the problem of selection of the regularization parameter of the LASSO. We also give an example of use of aggregation to achieve minimax adaptation over anisotropic Besov spaces, which was not previously known in minimax theory (in regression on a random design).

Keywords. Aggregation ; Exact oracle inequality ; Empirical risk minimization ; Empirical process theory ; Sparsity ; Minimax adaptation

1 Introduction

1.1 Motivations

In this paper, we consider the problem of sparse aggregation. Namely, given a dictionary $F = \{f_1, \ldots, f_M\}$ of functions, we look for an optimal aggregation algorithm that writes $\hat{f} = \sum_{j=1}^{M} \theta_j f_j$ with as many zero coefficients $\theta_j$ as possible. This question appears when one wants to use aggregation procedures to construct adaptive procedures. Indeed, in practice many elements of the dictionary appear to be irrelevant. We would like to remove completely these irrelevant elements of the dictionary from the final aggregate, while keeping the optimality of the procedure (“optimality” is used in reference to the definition of “optimal aggregation procedure” provided in Tsybakov (2003b) and Lecué and Mendelson (2009a)). Moreover, one could imagine large dictionaries containing many different types of estimators (kernel estimators, ...

\textsuperscript{1}Université Pierre et Marie Curie - Paris 6, Laboratoire de Statistique Théorique et Appliquée. email: stephane.gaiffas@upmc.fr

\textsuperscript{2}CNRS, Laboratoire d’Analyse et Mathématiques appliquées, Université Paris-Est - Marne-la-vallée email: guillaume.lecue@univ-mlv.fr

\textsuperscript{3}This work is supported by French Agence Nationale de la Recherche (ANR) ANR Grant “PROGnostic” ANR-09-JCJC-0101-01. (http://www.lsta.upmc.fr/prognostic/index.php)
projection estimators, etc.) with many different parameters (smoothing parameters, groups of variables, etc.). Some of the estimators are likely to be more adapted than the others, depending on the kind of models that fits well to the data. So, we would like to construct procedures that can adapt to different models combining only the estimators, contained in the dictionary, that are the more adapted for this model.

Up to now, optimal procedures are based on exponential weights (cf. Juditsky et al. (2008), Dalalyan and Tsybakov (2007)) providing aggregation procedures with no zero coefficients, even for the worse elements in the dictionary. An improvement going in the direction of sparse aggregation has been made using a preselection step in Lecué and Mendelson (2009a). This preselection step allows to remove all the estimators in $F$ which performs badly on a learning subsample.

In the present work, we prove that optimal aggregation algorithms with only two non-zero coefficients exists, see Section 2, Theorem 1. This means that the aggregate writes as a convex combination of only two elements of $F$. Then, we propose an original proof of an already known result, involving an explicit geometrical setup, of the fact that selecting a single element of $F$ using empirical risk minimization is a suboptimal aggregation procedure, see Theorem 2. Finally, we use our “hyper-sparse” aggregate on a dictionary “consisting” of penalized empirical risk minimizers (PERM). The aim is to construct, as an application of the previous analysis, an adaptive estimator over anisotropic Besov balls, namely an estimator that adapts to the unknown anisotropic smoothness of the regression function, in the sense that it achieves the optimal minmax rate without an a priori knowledge of the anisotropic smoothness parameters. This result was not, as far as we know, previously proposed in minimax theory. To do so, we use recent results by Mendelson and Neeman (2009) on regularized learning together with our oracle inequality for the hyper-sparse aggregate.

1.2 The model
Let $\Omega$ be a measurable space endowed with a probability measure $\mu$ and $\nu$ be a probability measure on $\Omega \times \mathbb{R}$ such that $\mu$ is its marginal on $\Omega$. Assume $(X,Y)$ and $D_n := (X_i,Y_i)_{i=1}^n$ to be $n + 1$ independent random variables distributed according to $\nu$. We work under the following assumption.

**Assumption 1.** We can write

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

where $\varepsilon$ is such that $\mathbb{E}(\varepsilon|X) = 0$ and $\mathbb{E}(\varepsilon^2|X) \leq \sigma^2_{\varepsilon}$ a.s. for some constant $\sigma_{\varepsilon} > 0$.

We will assume further that either $Y$ is bounded: $\|Y\|_{\infty} < +\infty$, or that $\varepsilon$ is subgaussian: $\|\varepsilon\|_{\psi_2} := \inf\{c > 0 : \mathbb{E}[\exp((\varepsilon/c)^2)] \leq 2\} < +\infty$, see below. We want to estimate the regression function $f_0$ using the observations $D_n$. If $f$ is a function, its error of prediction is given by the risk

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

and if $\hat{f}$ is a random function depending on the data $D_n$, the error of prediction is the conditional expectation

$$R(\hat{f}) = \mathbb{E}[(\hat{f}(X) - Y)^2|D_n].$$
Given a set of functions $F$, a natural way to approximate $f_0$ is to consider the empirical risk minimizer (ERM), that minimizes the functional

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

over $F$. This very basic principle is at the core of the procedures proposed in this paper. We will also commonly use the following notations. If $f \in \arg\min_{f \in F} R(f)$, we will consider the excess loss

$$L_f = L_F(f)(X,Y) := (Y - f(X))^2 - (Y - f_F(X))^2,$$

and use the notations

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

2 Hyper-sparse aggregation

2.1 The aggregation problem

Assume that we are given a finite set $F = \{f_1, \ldots, f_M\}$ of functions (usually called a dictionary), the aggregation problem is to construct procedures $\hat{f}$ (usually called an aggregate) satisfying inequalities of the form

$$R(\hat{f}) \leq c \min_{f \in F} R(f) + r(F,n),$$

where the result holds with high probability or in expectation. Inequalities of the form (2) are called oracle inequalities and $r(F,n)$ is called the residue. We want the residue to be as small as possible. A classical result (cf. Juditsky et al. (2008)) says that aggregates with values in $F$ cannot mimic exactly (that is for $c = 1$) the oracle faster than $r(F,n) \sim ((\log M)/n)^{1/2}$. Nevertheless, it is possible to mimic the oracle up to the residue $(\log M)/n$ (see Juditsky et al. (2008) and Lecué and Mendelson (2009a), among others).

An aggregate typically write as a convex combination of the elements of $F$, namely

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

where $\theta := (\theta_j(D_n,F))_{j=1}^{M}$ is a map $\{1, \ldots, M\} \rightarrow \Theta$, where

$$\Theta := \left\{ \lambda \in (\mathbb{R}^+)^M : \sum_{i=1}^{M} \lambda_j = 1 \right\}.$$

Popular examples of aggregation algorithms are the aggregate with cumulated exponential weights (ACEW), see Catoni (2001); Leung and Barron (2006); Juditsky et al. (2008, 2005); Audibert (2009), where the weights are given by

$$\theta_j^{(ACEW)} := \frac{1}{n} \sum_{k=1}^{n} \frac{\exp(-\sum_{i=1}^{k} (Y_i - f_j(X_i))^2/T)}{\sum_{l=1}^{M} \exp(-\sum_{i=1}^{k} (Y_i - f_l(X_i))^2/T)},$$
where $T$ is the so-called temperature parameter, and the aggregate with exponential weights (AEW), see Dalalyan and Tsybakov (2007) among others, where

$$
\theta_j^{(AEW)} := \frac{\exp(-\sum_{i=1}^{n}(Y_i - f_j(X_i))^2/T)}{\sum_{l=1}^{M}\exp(-\sum_{i=1}^{n}(Y_i - f_l(X_i))^2/T)}.
$$

The ACEW satisfies (2) with $c = 1$ and $r(F,n) \sim (\log M)/n$, see references above, hence it is optimal in the sense of Tsybakov (2003b). In these aggregates, no coefficient equals zero, although they can be very small, depending on the value of $R_n(f_j)$ and $T$ [this makes in particular the choice of $T$ of importance]. In this paper, we look for an aggregation algorithm that shares the same property of optimality, but with as few non-zero coefficients $\theta_j$ as possible, hence the name hyper-sparse aggregate. We ask for the following question:

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

It turns out that the answer to this question is two. Indeed, if every coefficient is zero, excepted for one, the aggregate coincides with an element of $F$, and we know that such a procedure can only achieve the rate $((\log M)/n)^{1/2}$ (see Juditsky et al. (2008) and Theorem 2 below where, in the particular case of the ERM, the suboptimality of this kind of procedure can be understood from a geometrical point of view (this differs from the statistical point of view from Juditsky et al. (2008) which involves “min-max” type theorem)). In Definition 1, we construct three procedures, where two of them (see (7) and (8)), only have two non-zero coefficients $\theta_j$, and we prove in Theorem 1 below that these procedures are optimal. We shall assume one of the following.

**Assumption 2.** One of the following holds.

- There is a constant $b > 0$ such that:
  $$
  \max(\|Y\|_\infty, \sup_{f \in F} \|f\|_\infty) \leq b. 
  $$
  (3)

- There is a constant $b > 0$ such that:
  $$
  \max(\|\varepsilon\|_{\psi_2}, \sup_{f \in F} |f(X) - f_0(X)|\|\psi_2\|) \leq b.
  $$
  (4)

Note that Assumption (4) allows an unbounded dictionary $F$. The results given below differ a bit depending on the considered assumption (there is an extra $\log n$ term in the subgaussian case given by (4)). To simplify the notations, we assume from now that we have $2n$ observations from a sample $D_{2n} = (X_i, Y_i)_{i=1}^{2n}$. Let us define our aggregation procedures.

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

(0. Initialization) Choose a confidence level $x > 0$. If (3) holds, define

$$
\phi = \phi_{n,M}(x) = b \sqrt{\frac{\log M + x}{n}}.
$$
If (4) holds, define
\[ \phi = \phi_n,M(x) = (\sigma_x + b)\sqrt{\frac{(\log M + x) \log n}{n}}. \]

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

(2. Preselection) Use \( D_{n,1} \) to define a random subset of \( F \):
\[ \hat{F}_1 = \left\{ f \in F : R_{n,1}(f) \leq R_{n,1}((\hat{f}_{n,1}) + c \max \{ \phi \| f_{n,1} - f \|_{n,1}, \phi^2 \} \right\}, \]
where \( \| f \|_{n,1}^2 = n^{-1} \sum_{i=1}^n f(X_i)^2, R_{n,1}(f) = n^{-1} \sum_{i=1}^n (f(X_i) - Y_i)^2, \hat{f}_{n,1} \in \arg\min_{f \in F} R_{n,1}(f) \).

(3. Aggregation) Choose \( \hat{F} \) as one of the following sets:
\[ \hat{F} = \text{conv}(\hat{F}_1) = \text{the convex hull of } \hat{F}_1 \]
\[ \hat{F} = \text{seg}(\hat{F}_1) = \text{the segments between the functions in } \hat{F}_1 \]
\[ \hat{F} = \text{star}(\hat{f}_{n,1}, \hat{F}_1) = \text{the segments between } \hat{f}_{n,1} \text{ with the elements of } \hat{F}_1, \]
and return the ERM relative to \( D_{n,2} \):
\[ \hat{f} \in \arg\min_{g \in \hat{F}} R_{n,2}(g), \]
where \( R_{n,2}(f) = n^{-1} \sum_{i=n+1}^{2n} (f(X_i) - Y_i)^2 \).

These algorithms are illustrated in Figures 1 and 2. In Figure 1 we summarize the aggregation steps in the three cases. In Figure 2 we give a simulated illustration of the preselection step, and we show the value of the weights of the AEW for a comparison. As mentioned above, the Step 3 of the algorithm returns, when \( \hat{F} \) is given by (7) or (8), a function which is a convex combination of only two functions in \( F \), among the ones remaining after the preselection step. The preselection step was introduced in Lecué and Mendelson (2009a), with the use of (6) in the aggregation step.

Each of the three procedures proposed in Definition 1 are optimal in view of Theorem 1 below. From the computational point of view, procedure (8) is the most appealing: an ERM in \( \text{star}(\hat{f}_{n,1}, \hat{F}) \) can be computed in a fast and explicit way, see Algorithm 1 below. The next Theorem proves that each of these aggregation procedures are optimal.

**Theorem 1.** Let \( x > 0 \) be a confidence level, \( F \) be a dictionary with cardinality \( M \) and \( \hat{f} \) be one of the aggregation procedure given in Definition 1. If
\[ \max(\|Y\|_{\infty}, \sup_{f \in F} \|f\|_{\infty}) \leq b, \]
we have, with \( \nu^{2n} \)-probability at least \( 1 - 2e^{-x} \):
\[ R(\hat{f}) \leq \min_{f \in F} R(f) + c_b \frac{(1 + x) \log M}{n}, \]
Figure 1: Aggregation algorithms: ERM over \( \text{conv}(\tilde{F}_1) \), seg(\( \tilde{F}_1 \)), or star(\( \tilde{f}_{n,1}, \tilde{F}_1 \)).

where \( c_0 \) is a constant depending on \( b \), and where we recall that \( R(\tilde{f}) = \mathbb{E}[(Y - \tilde{f}(X))^2] \) if

\[
\max(\|\varepsilon\|_{\psi_2}, \|f - f_0\|_{\psi_2}) \leq b,
\]

we have, with \( \nu^n \)-probability at least \( 1 - 4e^{-x} \):

\[
R(\tilde{f}) \leq \min_{f \in F} R(f) + c_{\sigma, \varepsilon}(1 + x) \log M \log n / n.
\]

Remark 1. Note that the definition of the set \( \tilde{F}_1 \), and thus \( \tilde{f} \), depends on the confidence \( x \) through the factor \( \phi_{n,M}(x) \).

Remark 2. To simplify the proofs, we don’t give the explicit values of the constants. However, when (3) holds, one can choose \( c = 4(1 + 9b) \) in (5) and \( c = c_1(1 + b) \) when (4) holds (where \( c_1 \) is the absolute constant appearing in Theorem 5). Of course, this is not likely to be the optimal choice.

2.2 The star-shaped aggregate

In this section we give details for the computation of the star-shaped aggregate, namely the aggregate \( \tilde{f} \) given by Definition 1 when \( \tilde{F} \) is (8). Indeed, if \( \lambda \in [0, 1] \), we have

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

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

\[
\lambda_{n,2}(f, g) = a \lor \frac{1}{2} \left( \frac{R_{n,2}(g) - R_{n,2}(f)}{\|f - g\|_{n,2}^2} + 1 \right) \land 1,
\]

where \( a \lor b = \max(a, b) \), \( a \land b = \min(a, b) \) and \( \min_{\lambda \in [0, 1]} R_{n,2}(\lambda f + (1 - \lambda)g) \) is thus equal to \( R_{n,2}(\lambda_{n,2}(f, g), f + (1 - \lambda_{n,2}(f, g))g) \) given by

\[
\left\{ \begin{array}{ll}
\frac{R_{n,2}(f)}{2} + \frac{R_{n,2}(g)}{2} - \frac{(R_{n,2}(f) - R_{n,2}(g))^2}{4\|f - g\|_{n,2}^2} - \frac{\|f - g\|_{n,2}^2}{4} & \text{if } R_{n,2}(f) - R_{n,2}(g) \geq \|f - g\|_{n,2}^2 \\
\frac{R_{n,2}(f)}{2} + \frac{R_{n,2}(g)}{2} - \frac{(R_{n,2}(f) - R_{n,2}(g))^2}{4\|f - g\|_{n,2}^2} & \text{if } |R_{n,2}(f) - R_{n,2}(g)| \leq \|f - g\|_{n,2}^2 \\
\end{array} \right.
\]
This leads to the following algorithm for the computation of \( \hat{f} \).

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

**Input:** dictionary \( F \), data \((X_i, Y_i)_{i=1}^{2n}\), and a confidence level \( x > 0 \)

**Output:** star-shaped aggregate \( \tilde{f} \)

Split \( D_{2n} \) into two samples \( D_{n,1} \) and \( D_{n,2} \)

**foreach** \( j \in \{1, \ldots, M\} \) **do**

Compute \( R_{n,1}(f_j) \) and \( R_{n,2}(f_j) \), and use this loop to find \( \hat{f}_{n,1} \in \arg\min_{f \in F} R_{n,1}(f) \)

**end**

**foreach** \( j \in \{1, \ldots, M\} \) **do**

Compute \( \|f_j - \hat{f}_{n,1}\|_{n,1} \) and \( \|f_j - \hat{f}_{n,1}\|_{n,2} \)

**end**

Construct the set of preselected elements

\[
\hat{F}_1 = \left\{ f \in F : R_{n,1}(f) \leq R_{n,1}(\hat{f}_{n,1}) + c \max (\phi \|\hat{f}_{n,1} - f\|_{n,1}, \phi^2) \right\},
\]

where \( \phi \) is given in Definition 1.

**foreach** \( f \in \hat{F}_1 \) **do**

compute \( R_{n,2}(\lambda_{n,2}(\hat{f}_{n,1}, f)\hat{f}_{n,1} + (1 - \lambda_{n,2}(\hat{f}_{n,1}, f))f) \)

and keep the element \( f_j \in \hat{F}_1 \) that minimizes this quantity

**end**

return \( \tilde{f} = \lambda_{n,2}(\hat{f}_{n,1}, f_j)\hat{f}_{n,1} + (1 - \lambda_{n,2}(\hat{f}_{n,1}, f_j))f_j \),
2.3 Suboptimality of Penalized ERM

In this section, we prove that minimizing the empirical risk $R_n(\cdot)$ (or a penalized version, called PERM from now on) on $F(\Lambda)$ is a suboptimal aggregation procedure both in expectation and deviation. According to Tsybakov (2003b), the optimal rate of aggregation in the gaussian regression model is $(\log M)/n$. This means that it is the minimum price one has to pay in order to mimic the best function among a class of $M$ functions with $n$ observations. This rate is achieved by the aggregate with cumulative exponential weights, see Catoni (2001), Yang (2000) and Juditsky et al. (2008). In Theorem 2 below, we prove that the usual PERM procedure cannot achieve this rate and thus, that it is suboptimal compared to the aggregation methods with exponential weights. The lower bounds for aggregation methods appearing in the literature (see Tsybakov (2003b); Juditsky et al. (2008); Lecué (2006)) are usually based on minimax theory arguments. In particular, in Tsybakov (2003b), it is proved that a selector (that is an aggregation procedure taking its values in the dictionary itself) cannot mimic the oracle faster than $\sqrt{(\log M)/n}$. This result implies the one that we have here, but, it doesn’t provide an explicit setup for which a given selector performs poorly. The result in Juditsky et al. (2008) says that whatever the selector is, there exists a probability measure and a dictionary for which it cannot mimic the oracle faster than $\sqrt{(\log M)/n}$. The proof of this result does not tell explicitly which probabilistic setup is bad for this selector. In the present result, we are interested in a particular type of selector: the PERM for some penalty. We can provide an explicit framework (dictionary+probabilistic setup) because the argument considered here is based on some geometric considerations (in the same spirit as the lower bound obtained in Lee et al. (1996) and Mendelson (2008)). The explicit example that makes the PERM fail is the following Gaussian regression model with uniform design:

**Assumption 3 (G).** Assume that $\varepsilon$ is standard Gaussian and that $X$ is univariate and uniformly distributed on $[0, 1]$.

The dictionary is constructed as follow:

![Figure 3: Example of a setup in which ERM performs badly. The set $F(\Lambda) = \{f_1, \ldots, f_M\}$ is the dictionary from which we want to mimic the best element and $f_0$ is the regression function.](image)

For the regression function we take

$$f_0(x) = \begin{cases} 2h & \text{if } x^{(M)} = 1 \\ h & \text{if } x^{(M)} = 0, \end{cases} \quad (9)$$
where $x$ has the dyadic decomposition $x = \sum_{k \geq 1} x^{(k)}2^{-k}$ where $x^{(k)} \in \{0, 1\}$ and

$$h = C \sqrt{\log M \over n}.$$  

We consider the dictionary of functions $F_M = \{f_1, \ldots, f_M\}$

$$f_j(x) = 2x^{(j)} - 1, \quad \forall j \in \{1, \ldots, M\},$$

where again $(x^{(j)} : j \geq 1)$ is the dyadic decomposition of $x \in [0, 1]$.

**Theorem 2.** There exists an absolute constant $c_0 > 0$ such that the following holds. Let $M \geq 2$ be an integer and assume that (G) holds. We can find a regression function $f_0$ and a family $F(\Lambda)$ of cardinality $M$ such that, if one considers a penalization satisfying $|\text{pen}(f)| \leq C \sqrt{\log M / n}, \forall f \in F(\Lambda)$ with $0 \leq C < \sigma (24 \sqrt{2c^*})^{-1}$ ($c^*$ is an absolute constant from the Sudakov minorization, see Theorem 7 in Appendix A.2), the PERM procedure defined by

$$\hat{f}_n \in \arg\min_{f \in F(\Lambda)} \left( R_n(f) + \text{pen}(f) \right)$$

satisfies, with probability greater than $c_0$,

$$\|\hat{f}_n - f_0\|^2 \geq \min_{f \in F(\Lambda)} \|f - f_0\|^2 + C_3 \sqrt{\log M \over n}$$

for any integer $n \geq 1$ and $M \geq M_0(\sigma)$ such that $n^{-1} \log[(M - 1)(M - 2)] \leq 1/4$ where $C_3$ is an absolute constant.

This result tells that, in some particular cases, the PERM cannot mimic the best element in a class of cardinality $M$ faster than $(\log M / n)^{1/2}$. This rate is very far from the optimal one $(\log M / n)^{1/2}$. Of course, one can say that the PERM fails to achieve the optimal rate only in the very particular framework that we have constructed here. Nevertheless, this approach can be generalized (we refer the reader to Lecué and Mendelson [2009b] for instance). Finally, remark that classical penalty functions are of the order $[\text{Complexity of the class}]$ divided by $n$, which is in our aggregation setup of the order of $(\log M / n)$. Thus, the restriction that we have on the penalty function covers the classical cases that one can meet in the literature on penalization methods.

Let $F(\Lambda)$ be the set that we consider in the proof of Theorem 2 (see Section 5 below), and take $\text{pen}(f) = 0$. Using Monte-Carlo (we do 5000 loops), we compute the excess risk $E\|\hat{f}_n - f_0\|^2 - \min_{f \in F(\Lambda)} \|f - f_0\|^2$ of the ERM. In Figure 4 below, we compare the excess risk and the bound $(\log M / n)^{1/2}$ for several values of $M$ and $n$. It turns out that, for this set $F(\Lambda)$, the lower bound $(\log M / n)^{1/2}$ is indeed accurate for the excess risk. Actually, by using the classical symmetrization argument and the Dudley’s entropy integral (or Pisier’s inequality), it is easy to obtain an upper bound for the excess risk of the ERM of the order of $(\log M / n)^{1/2}$ for any class $F(\Lambda)$ of cardinality $M$.

As an application of the aggregation algorithm, we consider the problem of adaptation to the regularization parameter of a penalized empirical risk minimization procedure, denoted for short PERM in what follows.
3 An example of dictionary: Penalized ERM

3.1 Definition and tools

Let us fix a function space \( \mathcal{F} \), endowed with a seminorm \( |\cdot|_\mathcal{F} \). The set \( \mathcal{F} \) is a space of functions, such as a Sobolev, Besov or Reproducing Kernel Hilbert Space (RHKS), the latter being a common example in regularized learning, see [Cucker and Smale] (2002). A simple example (in the one dimensional case) is the Sobolev space \( W^s_2 \) of functions such that \( |f|^2_2 = \int f^{(s)}(t)^2 dt < +\infty \), which corresponds to the so-called smoothing splines estimator, see [Wahba] (1990). A PERM (which stands for penalized empirical risk minimization) minimizes the functional

\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \text{pen}(f)
\]

over \( \mathcal{F} \), where \( \text{pen}(f) \) is a quantity measuring the smoothness (or “roughness”) of \( f \in \mathcal{F} \). Typically, the penalization term writes \( \text{pen}(f) = h^2 |f|^2_\mathcal{F} \) (see [van de Geer] (2000) and [Györfi et al.] (2002) among others), where \( h > 0 \) is a regularization parameter.

In [Mendelson and Neeman] (2009), sharp error bounds for the PERM are established, in the general context of a so-called ordered and parametrized hierarchy \( \{\mathcal{F}_r : r > 0\} \). An example of such an ordered and parametrized hierarchy is

\[ \mathcal{F}_r = r \mathcal{F}_1, \quad \text{where } \mathcal{F}_1 = \{f \in \mathcal{F} : |f|_\mathcal{F} \leq 1\}. \]

In the latter paper, a very sharp analysis is conducted when \( \mathcal{F} \) is a RKHS, allowing for penalizations less than quadratic in the RKHS norm. In this section, we use the tools proposed in [Mendelson and Neeman] (2009) to derive an error bound for the PERM using the standard penalty \( \text{pen}(f) = h^2 |f|^2_\mathcal{F} \), but when \( \mathcal{F} \) is a Besov space. In nonparametric estimation literature, Besov spaces are of particular interest since they include functions with inhomogeneous smoothness, for instance functions with rapid oscillations or bumps. Moreover, since the design random variable \( X \) is eventually multivariate, the question of anisotropic smoothness naturally arises. Anisotropy means that the smoothness of the regression function \( f_0 \) differs in each direction. As far as we know, adaptive estimation of a multivariate curve with anisotropic smoothness was previously considered only in Gaussian white noise or density models,
Assumption 4. \( \text{Assume that } \mathcal{F} \text{ embeds continuously in } C(\mathbb{R}^d), \text{ and that there is a number } \beta \in (0, 2) \text{ such that for any } \delta > 0, \text{ the unit ball of } \mathcal{F} \text{ satisfies:} \)

\[ H_\infty(\delta, \mathcal{F}_1) \leq c \delta^{-\beta}, \tag{12} \]

where \( c > 0 \) is independent of \( \delta \).

This assumption entails \( H_\infty(\delta, \mathcal{F}_r) \leq c(r/\delta)^\beta \) for any \( r > 0 \). Moreover, the continuous embedding gives that \( \|f\|_\infty \leq c |f|_\mathcal{F} \) for any \( f \in \mathcal{F} \). Assumption 4 is satisfied by barely all the smoothness spaces considered in nonparametric literature (at least when the smoothness of the space is large enough compared to the dimension, see below). Let us give an example. Let \( B^{s}_{p,q} \) be the anisotropic Besov space with smoothness \( s = (s_1, \ldots, s_d) \). This space is precisely defined in Appendix B. Each \( s_i \) corresponds to the smoothness in the \( i \)-th coordinate. The computation of the entropy of \( \mathcal{F}_1 \) when \( \mathcal{F} = B^{s}_{p,q} \) is done in Theorem 5.30 from Triebel (2006). Namely, if \( \bar{s} \) is the harmonic mean of \( s \), given by

\[ \frac{1}{\bar{s}} := \frac{1}{d} \sum_{i=1}^{d} \frac{1}{s_i}, \tag{13} \]

then the unit ball \( \mathcal{F}_1 \) of \( \mathcal{F} = B^{s}_{p,q} \) satisfies Assumption 4 with \( \beta = d/\bar{s} \), given that \( \bar{s} > d/p \), which is the usual condition to have the embedding in \( C(\mathbb{R}^d) \).

3.2 Local complexity using entropy

Now, we have in mind to use Theorem 2.5 from Mendelson and Neeman (2009), in order to derive a risk bound for the PERM. For this, we need a control on the local complexity of \( \mathcal{F}_r \), for any \( r > 0 \). The complexity is measured in this paper by the expectation \( \mathbb{E}||P - P_n||_{V_{r,\lambda}} \), where for \( r, \lambda > 0 \), \( V_{r,\lambda} \) is the class of excess losses

\[ V_{r,\lambda} := \{ \alpha \mathcal{L}_{r,f} : 0 \leq \alpha \leq 1, f \in \mathcal{F}_r, \mathbb{E}(\alpha \mathcal{L}_{r,f}) \leq \lambda \}, \]

where

\[ \mathcal{L}_{r,f} := (Y - f(X))^2 - (Y - f_r^*(X))^2 \]

and \( f_r^* \in \text{argmin}_{f \in \mathcal{F}_r} \mathbb{E}(Y - f(X))^2 \). The next Lemma (a proof is given in Appendix A.3) gives a bound on this measure of the complexity under Assumption 4.

**Lemma 1.** Assume that \( ||Y||_\infty < +\infty \) and grant Assumption 4. One has, for any \( r, \lambda > 0 \):

\[ \mathbb{E}||P - P_n||_{V_{r,\lambda}} \leq c \max \left[ r^2 n^{-1/3(1+\beta)/2}, \frac{r^{1+\beta/2} \lambda^{1-\beta/2}/2}{\sqrt{n}} \right], \]

11
where \( c = c_\beta \|Y\|_\infty \).

This Lemma, although probably not optimal, is sufficient to provide a satisfactory risk bound for the PERM with a penalization of the form \( \text{pen}(f) = h^2 |f|^2_F \). It is close in spirit to a bound proposed in Loustau (2009) (see Theorem 1) for the problem of classification framework using a Besov penalization, with an extra assumption on the inputs \( X_i \), since the proof involves a decomposition on a wavelet basis. Here we use only the entropy condition, together with some basic tools from empirical process theory, see the proof in Appendix A.3.

### 3.3 A risk bound for the PERM using entropy

Now, we can derive a risk bound for the PERM using Lemma 1 and the results from Mendelson and Neeman (2009). First, note that

\[
\lambda / 8 \geq c \max \left( r^2 n^{-1/(1+\beta/2)}, \frac{r^{1+\beta/2} \lambda (1-\beta/2)}{\sqrt{n}} \right)
\]

if and only if \( \lambda \geq cr^2 n^{-1/(1+\beta/2)} \). So, any \( \lambda \geq cr^2 n^{-1/(1+\beta/2)} \) satisfies, using Lemma 1, that

\[
\lambda / 8 \geq \mathbb{E} \|P - P_n\|_{\psi,\lambda},
\]

and consequently, using the “isomorphic coordinate projection” [see Theorem 2.2 in Mendelson and Neeman (2009)], we have that for any \( f \in \mathcal{F} \), the following holds w.p. larger than \( 1 - 2e^{-x} \):

\[
\frac{1}{2} P_n \mathcal{L}_{r,f} - \rho_n(r, x) \leq P \mathcal{L}_{r,f} \leq 2P_n \mathcal{L}_{r,f} + \rho_n(r, x),
\]

where

\[
\rho_n(r, x) := c \left( r^2 n^{-1/(1+\beta/2)} + \frac{(1 + r^2) x}{n} \right).
\]  \( (14) \)

This explains the shape of the usual quadratic penalization \( \text{pen}(f) = h^2 |f|^2_F \), where \( h = cn^{-1/(2+\beta)} \) (up to the other term, which is of smaller order \( 1/n \)), and this entails the following.

**Theorem 3.** Assume that \( \|Y\|_\infty \leq b \), and grant Assumption 4. Let \( \rho_n(r, x) \) be given by \( (14) \) and define for \( r, y > 0 \):

\[
\theta(r, y) = y + \log(\pi^2/6) + 2\log(1 + cn + \log r),
\]

where \( c = c_{\beta,b} \). Then, for any \( x > 0 \), with probability at least \( 1 - 2 \exp(-x) \), any \( f \in \mathcal{F} \) that minimizes the functional

\[
P_n \ell_f + c_1 \rho_n(2|f|_F, \theta(|f|_F, x))
\]

over \( \mathcal{F} \) also satisfies

\[
P \ell_f \leq \inf_{f \in \mathcal{F}} \left( P \ell_f + c_2 \rho_n(2|f|_F, \theta(|f|_F, x)) \right).
\]

**Proof.** The conditions of Theorem 2.5 in Mendelson and Neeman (2009) are satisfied with \( \rho_n(r, x) \) given by \( (14) \). The statement of the Theorem easily follows from it, using the same arguments as in the proof of Theorem 3.7 herein. \( \square \)
Let us rewrite the result of Theorem 3. For any $x > 0$, if $\bar{f}$ is the PERM at level $x$, one has, with $\nu^n$-probability larger than $1 - 2e^{-x}$:

$$P\ell_{\bar{f}} \leq \inf_{r > 0} \left\{ P\ell_{f_r} + c_1 r^2 n^{-1/3\beta/2} + c_2 \left(1 + r^2\right) \left(\frac{\pi^2}{6} + \log(1 + c_3 n + \log r)\right) \right\},$$

where we recall that $P\ell_{\bar{f}} = \mathbb{E}\left[(Y - \bar{f}(X))^2|X_1, \ldots, X_n\right]$ and $f_r \in \arg\min_{f \in \mathcal{F}} R(f)$. This inequality proves that $\bar{f}$ adapts to the radius $|f|_x$ of $f$ in $\mathcal{F}$. The leading term in the right hand side of this inequality is $r^2 n^{-2/\left(2 + \beta\right)}$. If $\mathcal{F} = B_{p, \infty}^s$, it becomes $r^2 n^{-2s/(2s + d)}$, which is the minimax optimal rate of convergence over anisotropic Besov space, see Kerkyacharian et al. (2007) for instance.

### 3.4 Adaptive estimation over anisotropic Besov space

What we have in mind now is the application of Theorems 1 and 3 to the problem of adaptive estimation over a collection of anisotropic Besov space. Consider two vectors $s^{\min}$ and $s^{\max}$ in $\mathbb{R}_+^d$ with positive coordinates and harmonic means $\bar{s}^{\min}$ and $\bar{s}^{\max}$ respectively, satisfying $s^{\min} \leq s^{\max}$ ($s^{\min}_i \leq s^{\max}_i$ for any $i \in \{1, \ldots, d\}$) and $s^{\min} > d/(\min(p, 2))$. Consider the collection of anisotropic Besov space

$$(B_{p, \infty}^s : s \in \mathcal{S}),$$

where $\mathcal{S} := \prod_{i=1}^d [s^{\min}_i, s^{\max}_i].$ (15)

The strategy is to aggregate a dictionary of PERM, corresponding to a discretization of $\mathcal{S}$, in order to adapt to the anisotropic smoothness of $f_0$. The steps are the following. We shall assume to simplify that we have $2n$ observations.

**Definition 2 (Adaptive estimator).**

1. **Split (at random) the whole sample** $(X_i, Y_i)_{i=1}^{2n}$ **into a training sample** $(X_i, Y_i)_{i=1}^{n}$ and a **learning sample** $(X_i, Y_i)_{i=n+1}^{2n}$. **Fix a confidence level** $x > 0$.

2. **Compute the uniform discretization of** $\mathcal{S}$ **with step** $(\log n)^{-1}$:

$$\mathcal{S}_n := \prod_{i=1}^d \left\{ s^{\min}_i + k(\log n)^{-1} : 1 \leq k \leq \floor{(s^{\max}_i - s^{\min}_i) \log n} \right\}. \quad (16)$$

Then, for each $s \in \mathcal{S}_n$, take $\bar{f}_s$ as a minimizer of the functional

$$\frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 + \text{pen}_s(f, x),$$

where

$$\text{pen}_s(f, x) = c_1 n^{-2s/(2s + d)} |f|_{B_{p, \infty}^s}^2 + c_2 \left(1 + |f|_{B_{p, \infty}^s}^2\right) \left(\frac{\pi^2}{6} + \log(1 + c_3 n + \log |f|_{B_{p, \infty}^s})\right).$$

If $b$ is such that $\|Y\|_\infty \leq b$, consider the dictionary of truncated PERM

$$F^{\text{perm}} = \{-b \vee \bar{f}_s \wedge b : s \in \mathcal{S}_n\}.$$
3. Using the learning sample \((X_i, Y_i)^{2n+1}_i\), compute one of the aggregates \(\tilde{f}\) given in Definition 4 using the dictionary \(F^{\text{PERM}}\).

The next Theorem, which is an immediate consequence of Theorems 1 and 3, proves that the aggregate \(\tilde{f}\) is minimax adaptative over the collection of anisotropic Besov spaces (15).

**Theorem 4.** Let \(\tilde{f}\) be the aggregated estimator given in Definition 4. Assume that 
\[
\max(\|Y\|_\infty, \|f_0\|_\infty) \leq b
\]
and that \(f_0 \in B_{p,\infty}^{s_0}\) for some \(s_0 \in S\), where \((B_{p,\infty}^s : s \in S)\) is the collection given by (15) that satisfies \(s_{\min} > d/p\). Then, with \(\nu^{2n}\)-probability larger than \(1 - 4e^{-x}\), we have:
\[
\|\tilde{f} - f_0\|_{L^2(\mu)}^2 
\leq c_1 r_0^2 n^{-\frac{2s_0}{2s_0+d}} + c_2 \frac{1 + r_0^2 + \log \log n}{n} \left( x + \log(\pi^2/6) + c \log(1 + c n + \log r_0) \right),
\]
where \(r_0 = \|f_0\|_{B_{p,\infty}^{s_0}}\) and
\[
\frac{1}{s_0} = \frac{1}{d} \sum_{i=1}^{d} \frac{1}{s_{0,i}}.
\]

The dominating term in the right hand side is of order \(n^{-\frac{2s_0}{2s_0+d}}\), which is the minimax optimal rate of convergence over anisotropic Besov space (a minimax lower bound over \(B_{p,\infty}^{s_0}\) can be easily obtained using standard arguments, such as the ones from Tsybakov (2003a), together with Bernstein estimates over \(B_{p,\infty}^s\) (that can be found in Triebel (2006) for instance). Note that there is no regular or sparse zone here, since the error of estimation is measured with \(L^2(\mu)\) norm. The result obtained here is stronger than the ones usually obtained in minimax theory, where one only gives an upper bound for \(\mathbb{E}\|\tilde{f} - f_0\|_{L^2(\mu)}^2\), while here is given a concentration inequality for \(\|\tilde{f} - f_0\|_{L^2(\mu)}^2\).

4 Simulation study

In this section, we propose a simulation study for the problem of selection of the smoothing parameter of the LASSO, see Tibshirani (1996); Efron et al. (2004). We simulate i.i.d. data
\[
Y_i = \beta_0^\top X_i + \epsilon_i,
\]
where \(\beta_0\) is a vector of size \(p = 91\) given by
\[
\beta_0 = (3, 1.5, 0^{30}, 2, -6, 4, 0^{25}, -4, 0^{15}, 2.5, 3, 0^{10}, 3, 1, -2)
\]
where \(0^n\) is the vector in \(\mathbb{R}^n\) with each coordinate set to zero. The noise \(\epsilon_i\) is centered Gaussian with variance \(\sigma^2\). The vector \(X = (X^1, \ldots, X^d)\) is a centered Gaussian vector such that the correlation between \(X^i\) and \(X^j\) is \(2^{-|i-j|}\) (following the examples from Tibshirani (1996)). Using the lars routine from R\(^1\), we construct a dictionary \(F\) made of the entire sequence of LASSO type estimators for various regularization parameters coming out of the LARS algorithm. Then we compare the prediction error \(\|X(\bar{\beta} - \beta_0)\|_2\) and the estimation error \(\|\bar{\beta} - \beta_0\|_2\) where \(\bar{\beta}\) is:

\(^1\text{www.r-project.org}\)
• $\hat{\beta}(C_p)$ = the LASSO with regularization parameter selected using Mallows-$C_p$ selection rule, see Efron et al. (2004).

• $\hat{\beta}(\text{AEW})$ = The aggregate with exponential weights computed on $F$ with temperature parameter $4\sigma^2$, see for instance Dalalyan and Tsybakov (2007).

• $\hat{\beta}(\text{star})$ = the star-shaped aggregate, see Algorithm 1, with constant $c = 2$.

We compute the errors $|X(\hat{\beta} - \beta_0)|_2$ and $|\hat{\beta} - \beta_0|_2$ using 100 simulations for several values of $n$ and $\sigma^2$. The splits taken are chosen at random with size $n/2$ for training and $n/2$ for learning for both the AEW and star-shaped aggregate (we don’t split the learning sample). For both aggregates we do some jackknife: instead of using a single aggregate, we compute a mean of 10 aggregates obtained with several splits chosen at random. This makes the final aggregates less dependent on the split. In order to make the oracle and the Mallows-$C_p$ errors comparable to the error of the aggregates (that need to split the data, while Mallows-$C_p$ doesn’t), we compute the weights of aggregation using splitting, then we compute the aggregate using a dictionary $F$ computed using the whole sample.

The conclusion is that, for this example, the star-shaped does a better job than both the AEW and the $C_p$ in most cases. When the noise level is not too high ($\sigma = 2$, which corresponds to a RSNR of 5), see the errors given in Figure 5, the star-shaped is always the best. When the noise level is high ($\sigma = 5$, RSNR=2) and $n$ is small, see Figure 6 the story is different: the AEW is better than the star-shaped. In such an extreme situation the AEW takes advantage of the averaging (recall that no coefficient is zero in the AEW). However, when $n$ becomes larger than $p$, the star-shaped improves again upon AEW.

5 Proofs of the main results

5.1 Proof of Theorem 1

Proof of Theorem 1. Let us prove the result in the $\psi_2$ case, the other case is similar. Fix $x > 0$ and let $\bar{F}$ be either (6), (7) or (8). Set $d := \text{diam}(\bar{F}, L^2(\mu))$. Consider the second half of the sample $D_{n,2} = (X_i, Y_i)_{i=n+1}^{2n}$. By Corollary 1 (see Appendix A below), with probability at least $1 - 4\exp(-x)$ (relative to $D_{n,2}$), we have for every $f \in \bar{F}$

$$\left| \frac{1}{n} \sum_{i=1+n}^{2n} \mathcal{L}_\hat{\beta}(f)(X_i, Y_i) - \mathbb{E}(\mathcal{L}_\hat{\beta}(f)(X, Y)|D_{n,1}) \right| \leq c(\sigma_x + b) \max(d\phi, b\phi^2),$$

where $\mathcal{L}_\hat{\beta}(f)(X, Y) := (f(X) - Y)^2 - (f^*_{\hat{\beta}}(X) - Y)^2$ is the excess loss function relative to $\hat{\beta}$, $f^*_{\hat{\beta}} \in \text{Arg min}_{f \in \hat{\beta}} R(f)$ and where $\phi = \sqrt{(\log M + x) \log n}/n$. By definition
Figure 5: Errors $|\hat{\beta} - \beta_0|_2$ (first row) and $|X(\hat{\beta} - \beta_0)|_2$ (second row) for $\sigma = 2$ and $n = 70, 100, 150$. 
Figure 6: Errors $|\hat{\beta} - \beta_0|_2$ (first row) and $|X(\hat{\beta} - \beta_0)|_2$ (second row) for $\sigma = 5$ and $n = 70, 100, 150$. 
of \( \hat{f} \), we have \( \frac{1}{n} \sum_{i=n+1}^{2n} \mathcal{L}_{\hat{f}}(\hat{f})(X_i, Y_i) \leq 0 \), so, on this event (relative to \( D_{n,2} \))

\[
R(\hat{f}) \leq R(f^\hat{F}) + \mathbb{E}(\mathcal{L}_{\hat{f}}(\hat{f})|D_{n,1}) - \frac{1}{n} \sum_{i=n+1}^{2n} \mathcal{L}_{\hat{f}}(\hat{f})(X_i, Y_i)
\]

(17)

\[
\leq R(f^\hat{F}) + c(\sigma + b) \max(d\phi, b\phi^2)
\]

\[
= R(f^F) + \left( c(\sigma + b) \max(d\phi, b\phi^2) - (R(f^F) - R(f^\delta)) \right)
\]

\[
=: R(f^F) + \beta,
\]

and it remains to show that

\[
\beta \leq c_{b,\sigma} \left( 1 + x \right) \log M \log n.
\]

When \( \hat{F} \) is given by (6) or (7), the geometrical configuration is the same as in \( \hat{F} \), so we skip the proof. Let us turn out to the situation where \( \hat{F} \) is given by (8). Recall that \( \hat{F}_n,1 \) is the ERM on \( \hat{F}_1 \) using \( D_{n,1} \). Consider \( f_1 \) such that \( \|\hat{f}_n,1 - f_1\|_{L^2(\mu)} = \max_{f \in \hat{F}_1} \|\hat{f}_n,1 - f\|_{L^2(\mu)} \), and note that \( \|\hat{f}_n,1 - f_1\|_{L^2(\mu)} \leq d \leq 2\|\hat{f}_n,1 - f_1\|_{L^2(\mu)}. \) The mid-point \( f_2 := (\hat{f}_n,1 + f_1)/2 \) belongs to star(\( \hat{f}_n,1, \hat{F}_1 \)). Using the parallelogram identity, we have for any \( u, v \in L_2(\nu) \):

\[
\mathbb{E}_\nu\left(\frac{u + v}{2}\right)^2 \leq \mathbb{E}_\nu(u^2) + \mathbb{E}_\nu(v^2) - \frac{\|u - v\|^2_{L_2(\nu)}}{4},
\]

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

\[
R(f_2) = \mathbb{E}(f_2(X) - Y)^2 = \mathbb{E}\left(\frac{\hat{f}_n,1(X) + f_1(X)}{2} - Y\right)^2
\]

\[
\leq \frac{1}{2} \mathbb{E}(\hat{f}_n,1(X) - Y)^2 + \frac{1}{2} \mathbb{E}(f_1(X) - Y)^2 - \frac{1}{4} \|\hat{f}_n,1 - f_1\|^2_{L^2(\mu)}
\]

\[
\leq \frac{1}{2} R(\hat{f}_n,1) + \frac{1}{2} R(f_1) - \frac{d^2}{16},
\]

where the expectations are taken conditioned on \( D_{n,1} \). By Lemma \( 4 \) (see Appendix A below), since \( \hat{f}_n,1, f_1 \in \hat{F}_1 \), we have

\[
\frac{1}{2} R(\hat{f}_n,1) + \frac{1}{2} R(f_1) \leq R(f^F) + c(\sigma + b) \max(\phi d, b\phi^2),
\]

and thus, since \( f_2 \in \hat{F} \)

\[
R(f^\hat{F}) \leq R(f_2) \leq R(f^F) + c(\sigma + b) \max(\phi d, b\phi^2) - cd^2.
\]

Therefore,

\[
\beta = c(\sigma + b) \max(\phi d, b\phi^2) - (R(f^F) - R(f^\hat{F}))
\]

\[
\leq c(\sigma + b) \max(\phi d, b\phi^2) - cd^2.
\]

Finally, if \( d \geq c_{\sigma, \phi} \) then \( \beta \leq 0 \), otherwise \( \beta \leq c_{\sigma, \phi} \).

\( \square \)
Proof of Theorem 2. The dictionary $F_M$ is chosen so that we have, for any $j \in \{1, \ldots, M-1\}$

$$\|f_j - f_0\|_{L^2([0,1])}^2 = \frac{5h^2}{2} + 1$$

and

$$\|f_M - f_0\|_{L^2([0,1])}^2 = \frac{5h^2}{2} - h + 1.$$ 

Thus, we have

$$\min_{j=1,\ldots,M} \|f_j - f_0\|_{L^2([0,1])}^2 = \|f_M - f_0\|_{L^2([0,1])}^2 = \frac{5h^2}{2} - h + 1.$$ 

This geometrical setup for $F(\Lambda)$, which is an unfavourable setup for the ERM, is represented in Figure 3. For

$$\hat{f}_n := \tilde{f}_{n,\text{PERM}} \in \arg\min_{f \in F_M} (R_n(f) + \text{pen}(f)),$$

where we take $R_n(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 = \|Y - f\|_n^2$, we have

$$\mathbb{E} \|\hat{f}_n - f_0\|_{L^2([0,1])}^2 = \min_{j=1,\ldots,M} \|f_j - f_0\|_{L^2([0,1])}^2 + h\mathbb{P}[\hat{f}_n \neq f_M].$$

(18)

Now, we upper bound $\mathbb{P}[\hat{f}_n = f_M]$. We consider the dyadic decomposition of the design variable $X$:

$$X = \sum_{k=1}^{+\infty} X^{(k)} 2^{-k},$$

(19)

where $(X^{(k)} : k \geq 1)$ is a sequence of i.i.d. random variables following a Bernoulli $\mathcal{B}(1/2, 1)$ with parameter 1/2 (because $X$ is uniformly distributed on $[0,1]$). If we define

$$N_j := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_i^{(j)} \varepsilon_i$$

and $\zeta_i^{(j)} := 2X_i^{(j)} - 1$, we have by the definition of $h$ and since $\zeta_i^{(j)} \in \{-1, 1\}$:

$$\frac{\sqrt{n}}{2\sigma} (\|Y - f_M\|_n^2 - \|Y - f_j\|_n^2)

= N_j - N_M + \frac{h}{2\sigma \sqrt{n}} \sum_{i=1}^{n} (\zeta_i^{(j)})^{(M)} \zeta_i + 3(\zeta_i^{(j)} - \zeta_i^{(M)}) - 1

\geq N_j - N_M - \frac{4C}{\sigma} \sqrt{\log M}.$$ 

This entails, for $\bar{N}_{M-1} := \max_{1 \leq j \leq M-1} N_j$, that

$$\mathbb{P}[\hat{f}_n = f_M] = P \left[ \bigcap_{j=1}^{M-1} \left\{ \|Y - f_M\|_n^2 - \|Y - f_j\|_n^2 \leq \text{pen}(f_j) - \text{pen}(f_M) \right\} \right]

\leq P \left[ N_M \geq \bar{N}_{M-1} - \frac{6C}{\sigma} \sqrt{\log M} \right].$$

It is easy to check that $N_1, \ldots, N_M$ are $M$ normalized standard gaussian random variables uncorrelated (but dependent). We denote by $\zeta$ the family of Rademacher
variables $(\zeta^{(j)}_i : i = 1, \ldots, n; j = 1, \ldots, M)$. We have for any $6C/\sigma < \gamma < (2\sqrt{2}\epsilon^*)^{-1}$ ($\epsilon^*$ is the “Sudakov constant”, see Theorem A.7,

$$\mathbb{P}[\hat{f}_n = f_M] \leq \mathbb{E}\left[ \mathbb{P}\left( \mathcal{N}_M \geq \mathcal{N}_{M-1} - \frac{6C}{\sigma} \sqrt{\log M} \mathbb{E}[\zeta] \right) \right]$$

$$\leq \mathbb{P}\left[ \mathcal{N}_M \geq -\gamma \sqrt{\log M} + \mathbb{E}(\mathcal{N}_{M-1}\zeta) \right] + \mathbb{E}\left[ \mathbb{E}(\mathcal{N}_{M-1}\zeta) - \mathcal{N}_{M-1} \geq (\gamma - \frac{6C}{\sigma}) \sqrt{\log M} \mathbb{E}[\zeta] \right].$$

Conditionally to $\zeta$, the vector $(\mathcal{N}_1, \ldots, \mathcal{N}_{M-1})$ is a linear transform of the Gaussian vector $(\varepsilon_1, \ldots, \varepsilon_n)$. Hence, conditionally to $\zeta$, $(\mathcal{N}_1, \ldots, \mathcal{N}_{M-1})$ is a Gaussian vector. Thus, we can use a standard deviation result for the supremum of Gaussian random vectors (see for instance [Massart (2007), Chapter 3.2.4]), which leads to the following inequality for the second term of the RHS in (20):

$$\mathbb{P}\left\{ \mathbb{E}(\mathcal{N}_{M-1}\zeta) - \mathcal{N}_{M-1} \geq (\gamma - \frac{6C}{\sigma}) \sqrt{\log M} \mathbb{E}[\zeta] \right\} \leq \exp(-\frac{1}{2}(\mathbb{C}/\mathbb{\sigma})^2 \log \mathbb{M}).$$

Remark that we used $\mathbb{E}[\mathcal{N}_j^2|\zeta] = 1$ for any $j = 1, \ldots, M - 1$. For the first term in the RHS of (20), we have

$$\mathbb{P}\left[ \mathcal{N}_M \geq -\gamma \sqrt{\log M} + \mathbb{E}(\mathcal{N}_{M-1}\zeta) \right]$$

$$\leq \mathbb{P}\left[ \mathcal{N}_M \geq -2\gamma \sqrt{\log M} + \mathbb{E}(\mathcal{N}_{M-1}\zeta) \right] + \mathbb{P}\left[ -\gamma \sqrt{\log M} + \mathbb{E}(\mathcal{N}_{M-1}) \geq \mathbb{E}(\mathcal{N}_{M-1}\zeta) \right].$$

Next, we use Sudakov’s Theorem (cf. Theorem A.7 in Appendix A.2) to lower bound $\mathbb{E}(\mathcal{N}_{M-1})$. Since $(\mathcal{N}_1, \ldots, \mathcal{N}_{M-1})$ is, conditionally to $\zeta$, a Gaussian vector and since for any $1 \leq j \neq k \leq M$ we have

$$\mathbb{E}[(\mathcal{N}_k - \mathcal{N}_j)^2|\zeta] = \frac{1}{n} \sum_{i=1}^{n} (\zeta^{(k)}_i - \zeta^{(j)}_i)^2$$

then, according to Sudakov’s minoration (cf. Theorem A.7 in the Appendix), there exits an absolute constant $c^* > 0$ such that

$$c^* \mathbb{E}[\mathcal{N}_{M-1}|\zeta] \geq \min_{1 \leq j \neq k \leq M-1} \left( \frac{1}{n} \sum_{i=1}^{n} (\zeta^{(k)}_i - \zeta^{(j)}_i)^2 \right)^{1/2} \sqrt{\log M}.$$

Thus, we have

$$c^* \mathbb{E}[\mathcal{N}_{M-1}] \geq \mathbb{E}\left[ \min_{j \neq k} \left( \frac{1}{n} \sum_{i=1}^{n} (\zeta^{(k)}_i - \zeta^{(j)}_i)^2 \right)^{1/2} \right] \sqrt{\log M}$$

$$\geq \sqrt{2} \left( 1 - \mathbb{E}\left[ \max_{j \neq k} \frac{1}{n} \sum_{i=1}^{n} \zeta^{(k)}_i \zeta^{(j)}_i \right] \right) \sqrt{\log M},$$

where we used the fact that $\sqrt{x} \geq x/\sqrt{2}, \forall x \in [0,2]$. Besides, using Hoeffding’s inequality we have $\mathbb{E}[\exp(s(\zeta^{(k)}_i))] \leq \exp(s^2/(2n))$ for any $s > 0$, where $\zeta^{(j,k)} :=$
\[ n^{-1} \sum_{i=1}^{n} \zeta_{\ell_i}^{(j)}, \] Then, using a maximal inequality (cf. Theorem 3 in Appendix A.2), and since \( n^{-1} \log((M-1)(M-2)) \leq 1/4, \)
\[ \mathbb{E}\left[ \max_{j \neq k} \frac{1}{n} \sum_{i=1}^{n} \zeta_{\ell_i}^{(k)}, \zeta_{\ell_i}^{(j)} \right] \leq \left( \frac{1}{n} \log((M-1)(M-2)) \right)^{1/2} \leq \frac{1}{2}. \quad (22) \]
This entails
\[ c^* \mathbb{E}[N_{M-1}] \geq \left( \frac{\log M}{2} \right)^{1/2}. \]

Thus, using this inequality in the first RHS of (21), and the usual inequality on the tail of a Gaussian random variable (\( N_M \) is standard Gaussian), we obtain:
\[ \mathbb{P}\left[ N_M \geq 2\gamma \sqrt{\log M} + \mathbb{E}(N_{M-1}) \right] \leq \mathbb{P}\left[ N_M \geq ((c^* \sqrt{2})^{-1} - 2\gamma) \sqrt{\log M} \right] \]
\[ \leq \mathbb{P}\left[ N_M \geq ((c^* \sqrt{2})^{-1} - 2\gamma) \sqrt{\log M} \right] \]
\[ \leq \exp \left( -((c^* \sqrt{2})^{-1} - 2\gamma)^2 (\log M)/2 \right). \quad (23) \]

Remark that we used \( 2\sqrt{2}c^* \gamma < 1. \) For the second term in (21), we apply the concentration inequality of Theorem 6 to the non-negative random variable \( \mathbb{E}[\bar{N}_{M-1}\zeta] \). We first have to control the second moment of this variable. We know that, conditionally to \( \zeta \), \( N_j|\zeta \sim \mathcal{N}(0,1) \) thus, \( N_j|\zeta \in L_{\psi_2} \) (for more details on Orlicz norm, we refer the reader to van der Vaart and Wellner (1996)). Thus,
\[ \max_{1 \leq j \leq M-1} \mathbb{E}[\bar{N}_{M-1}] \leq \mathbb{E}\left[ \sum_{j=1}^{M-1} \mathbb{E}^{\max_{1 \leq j \leq M-1} \| N_j \|_{\psi_2}} \right] \]
(cf. Lemma 2.2.2 in van der Vaart and Wellner (1996)). Since \( \| N_j \|_{\psi_2}^2 = 1 \), we have
\[ \max_{1 \leq j \leq M-1} \mathbb{E}[\bar{N}_{M-1}] \leq K \log M. \] In particular, we have \( \mathbb{E}[\bar{N}_{M-1}] \leq K \log M \) and so \( \mathbb{E}[\mathbb{E}[\bar{N}_{M-1}]^2] \leq K \log M. \) Then, Theorem 6 provides
\[ \mathbb{P}\left[ -\gamma \sqrt{\log M} + \mathbb{E}(N_{M-1}) \geq \mathbb{E}[\bar{N}_{M-1}] \right] \leq \exp(-\gamma^2/c_0), \quad (24) \]
where \( c_0 \) is an absolute constant.

Finally, combining (20), (23), (21), (24) in the initial inequality (20), we obtain
\[ \mathbb{P}[\hat{f}_n = f_M] \leq \exp(-3C/\sigma - \gamma)^2 \log M \]
\[ + \exp \left( -((c^* \sqrt{2})^{-1} - 2\gamma)^2 (\log M)/2 \right) + \exp(-\gamma^2/c_0). \]

Take \( \gamma = (12\sqrt{2}c^*)^{-1} \). It is easy to find an integer \( M_0(\sigma) \) depending only on \( \sigma \) such that for any \( M \geq M_0 \), we have \( \mathbb{P}[\hat{f}_n = f_M] \leq c_1 < 1 \), where \( c_1 \) is an absolute constant.

We complete the proof by using this last result in (18).

\[ \square \]

**Proof of Theorem 4.** Recall that we use the sample \( D_{n,1} \) to compute the family \( F = \{ -b \vee \hat{f}_k \wedge b : s \in S_n \} \) of \( \text{PERM} \), which has cardinality \( c(\log n)^d \), and the sample \( D_{n,2} \) to compute the weights of the aggregate \( \hat{f} \), see Definition 2. Recall also that there is \( s_0 = (s_{0,1}, \ldots, s_{0,d}) \in S \) such that \( f_0 \in B_{p,\infty}^{s_0} \), and denote \( r_0 = |f_0|_{B_{p,\infty}^{s_0}}. \)
Take \( s_* = (s_{*,1}, \ldots, s_{*,d}) \in S_n \) such that \( s_{*,j} \leq s_{0,j} \leq s_{*,j} + (\log n)^{-1} \) for all \( j = 1, \ldots, d. \) Remark that for this choice, one has \( B_{p,\infty}^{s_*} \subset B_{p,\infty}^{s_0} \) and \( n^{-2s_*/(2s_*/d)} \leq \sqrt{n} \). \( \square \)
Theorem 4, since

\[ | \tilde{f} - f_0 |^2 I (\nu) \leq \mathbb{E}(\ell f_* - \ell f_0 | D_{n,1}) + c \frac{(1 + x) \log \log n}{n}. \]

Now, using Theorem 3, we can find an event \( A_{x,n,1} \) satisfying \( \nu^{2n}(A_{x,n,1}) \geq 1 - 2e^{-x} \) on which:

\[
\mathbb{E}(\ell f_* - \ell f_0 | D_{n,1}) \leq \inf_{f : f \leq r_0} \mathbb{E}(\ell f - \ell f_0) + c_1 r_0^2 n^{-2\kappa / (2\kappa + d)} \\
+ \frac{c_2(1 + r_0^2)}{n} (x + \log(\frac{\pi^2}{6}) + \log(1 + c_3 n + \log r_0)) \leq c_1' r_0^2 n^{-2\kappa / (2\kappa + d)} + \frac{c_2(1 + r_0^2)}{n} (x + \log(\frac{\pi^2}{6}) + \log(1 + c_3 n + \log r_0)),
\]

where we used the fact that \( |f_0|_{p_n} \leq |f_0|_{p_n^*} \leq r_0 \). This concludes the proof of Theorem 4 since \( \nu^{2n}(A_{x,n,1} \cap A_{x,n,2}) \geq 1 - 4e^{-x}. \) □

A Tools from empirical process theory

A.1 Useful results from literature

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

**Theorem 5** (Theorem 4, [Adamczak (2008)]). Assume that \( X, X_1, \ldots, X_n \) are independent random variables and \( F \) is a countable set of functions such that \( \mathbb{E}f(X) = 0, \forall f \in F \) and, for some \( \alpha \in (0, 1) \), \( \| \sup_{f \in F} f(X) \|_{\psi_\alpha} < +\infty \). Define

\[
Z := \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right|
\]

and

\[
\sigma^2 = \sup_{f \in F} \mathbb{E}f(X)^2 \text{ and } b := \| \max_{i=1, \ldots, n} \sup_{f \in F} |f(X_i)| \|_{\psi_\alpha}.
\]

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

\[
\mathbb{P}[Z \geq (1 + \eta)EZ + \sigma \sqrt{2(1 + \delta) \frac{x}{n} + cb \left( \frac{x}{n} \right)^{1/\alpha}}] \leq 4e^{-x}
\]

\[
\mathbb{P}[Z \leq (1 - \eta)EZ - \sigma \sqrt{2(1 + \delta) \frac{x}{n} - cb \left( \frac{x}{n} \right)^{1/\alpha}}] \leq 4e^{-x}.
\]

A.2 Some probabilistic tools

For the first Theorem, we refer to [Einmahl and Mason (1996)]. The two following Theorems can be found, for instance, in [Massart (2007); van der Vaart and Wellner (1996); Ledoux and Talagrand (1991)].
symmetrization argument, see Giné and Zinn (1984), we have
\[
\mathbb{P} \left[ \sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i] \leq -n\delta \right] \leq \exp \left( -\frac{n\delta^2}{2\sigma^2} \right).
\]

**Theorem 6** (Einmahl and Masson (1996)). Let \( Z_1, \ldots, Z_n \) be \( n \) independent non-negative random variables such that \( \mathbb{E}[Z_i^2] \leq \sigma^2, \forall i = 1, \ldots, n \). Then, we have, for any \( \delta > 0 \),
\[
\mathbb{P} \left[ Z_i - \mathbb{E}[Z_i] \leq -n\delta \right] \leq \exp \left( -\frac{n\delta^2}{2\sigma^2} \right).
\]

**Theorem 7** (Sudakov). There exists an absolute constant \( c^* > 0 \) such that for any integer \( M \), any centered gaussian vector \( X = (X_1, \ldots, X_M) \) in \( \mathbb{R}^M \), we have,
\[
c^* \mathbb{E} \left[ \max_{1 \leq j \leq M} X_j \right] \geq \varepsilon \sqrt{\log M},
\]
where \( \varepsilon := \min \left\{ \sqrt{\mathbb{E}[(X_i - X_j)^2]} : i \neq j \in \{1, \ldots, M\} \right\} \).

**Theorem 8** (Maximal inequality). Let \( Y_1, \ldots, Y_M \) be \( M \) random variables satisfying \( \mathbb{E}[\exp(sY_j)] \leq \exp((s^2\sigma^2)/2) \) for any integer \( j \) and any \( s > 0 \). Then, we have
\[
\mathbb{E} \left[ \max_{1 \leq j \leq M} Y_j \right] \leq \sigma \sqrt{\log M}.
\]

### A.3 Some technical lemmas

In this section we state some technical Lemmas, used in the proof of Theorem 1.

**Notations**

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

**Proof of Lemma 1** We first start with lemma 4.6 of Mendelson and Neeman (2009) to obtain
\[
\mathbb{E} \| P - P_n \|_{V,\lambda} \leq 2 \sum_{i=0}^{\infty} 2^{-i} \mathbb{E} \| P - P_n \|_{L,2^{i+1}\lambda}
\]
(25)
where \( L,2^{i+1}\lambda := \{L_r,f : f \in F_r, \mathbb{E}[L_r,f] \leq 2^{i+1}\lambda \} \). Let \( i \in \mathbb{N} \). Using the Giné-Zinn symmetrization argument, see Giné and Zinn (1984), we have
\[
\mathbb{E} \| P - P_n \|_{L,2^{i+1}\lambda} \leq 2 \mathbb{E}_{X,Y} \mathbb{E}_{\epsilon} \left[ \sup_{\epsilon \in L,2^{i+1}\lambda} \left| \sum_{i=1}^{n} \epsilon_i L(X_i,Y_i) \right| \right],
\]
where \( (\epsilon_i) \) is a sequence of i.i.d. Rademacher variables. Recall that there is (see Ledoux and Talagrand (1991)) an absolute constant \( c_g \) such that for any \( T \subset \mathbb{R}^n \), we have
\[
\mathbb{E}_{\epsilon} \left[ \sup_{t \in T} \left| \sum_{i=1}^{n} \epsilon_i t_i \right| \right] \leq c_g \mathbb{E}_{\epsilon} \left[ \sup_{t \in T} \left| \sum_{i=1}^{n} g_i t_i \right| \right],
\]
(26)
where \( (g_i) \) are i.i.d. standard normal. So, we have
\[
\mathbb{E} \| P - P_n \|_{L,2^{i+1}\lambda} \leq \frac{2c_g}{n} \mathbb{E}_{X,Y} \mathbb{E}_{\epsilon} \left[ \sup_{\epsilon \in L,2^{i+1}\lambda} \left| \sum_{i=1}^{n} g_i L(X_i,Y_i) \right| \right].
\]
Consider the Gaussian process \( f \rightarrow Z_f := \sum_{i=1}^{n} g_{i} L_{f} (X_{i}, Y_{i}) \) indexed by \( \mathcal{F}_{r,2^{i+1} \lambda} := \{ f \in \mathcal{F}_{r} : E L_{r,f} \leq 2^{i+1} \lambda \} \). For every \( f, f' \in \mathcal{F}_{r,2^{i+1} \lambda} \), we have (conditionally to the observations)

\[
E_{\theta}|Z_f - Z_{f'}|^2 \leq 4(\|Y\|_{\infty} + r)^2 E_{\theta}|Z_f - Z_{f'}|^2,
\]

where \( Z_{f'} := \sum_{i=1}^{n} g_{i} (f(X_{i}) - f^*_r(X_{i})) \) and \( f^*_r \in \arg\min_{f \in \mathcal{F}_{r}} R(f) \). Using the convexity of \( \mathcal{F}_{r} \), it is easy to get \( E[\mathcal{L}_{r,f}] \geq \|f - f^*_r\|^2, \forall f \in \mathcal{F}_{r} \). Define

\[
B_{r,2^{i+1} \lambda} := \{ f - f^*_r : f \in \mathcal{F}_{r}, \|f - f^*_r\| \leq \sqrt{2^{i+1} \lambda} \}.
\]

Using Slepian’s Lemma (see, e.g. Ledoux and Talagrand (1991); Dudley (1999)), we have:

\[
E\|P - P_n\|_{\mathcal{L}_{r,2^{i+1} \lambda}} \leq c_{\gamma}(r + 1) E,
\]

where we put

\[
E := \frac{1}{n} E_{X} E_{\theta} \left[ \sup_{f \in B_{r,2^{i+1} \lambda}} \left( \sum_{i=1}^{n} g_{i} f(X_{i}) \right) \right].
\]

Moreover, using Dudley’s entropy integral argument (again, see Ledoux and Talagrand (1991); Dudley (1999), Massart (2007)), and Assumption 4, we have

\[
E \leq \frac{12}{\sqrt{n}} E_{X} \int_{0}^{\Delta} \sqrt{N(B_{r,2^{i+1} \lambda}, \|\cdot\|_{n}, t)} dt
\]

\[
\leq \frac{12 \sqrt{r}}{\sqrt{n}} E_{X} \int_{0}^{\Delta} \left( \frac{r}{t} \right)^{\beta/2} dt = \frac{c_{\beta}}{\sqrt{n}} r^{\beta/2} E_{X} |\Delta^{1-\beta/2}|
\]

\[
\leq \frac{c_{\beta}}{\sqrt{n}} r^{\beta/2} (E_{X} |\Delta^{2}|)^{(1-\beta/2)/2},
\]

where \( c_{\beta} = 12 \sqrt{r}/(1-\beta/2) \) and \( \Delta := \text{diam}(B_{r,2^{i+1} \lambda}, \|\cdot\|_{n}) \). But, one has, if \( B_{r,2^{i+1} \lambda} := \{ f^2 : f \in B_{r,2^{i+1} \lambda} \} \), using a contraction argument (see Ledoux and Talagrand (1991), Chapter 4), with again a Giné-Zinn symmetrization,

\[
E_{X} |\Delta^{2}| \leq E_{X} \|P - P_n\|_{B_{r,2^{i+1} \lambda}} + 2^{i+1} \lambda
\]

\[
\leq 4c_{\gamma}(r + 1) E + 2^{i+1} \lambda.
\]

Hence, \( E \) satisfies

\[
E \leq \frac{c}{\sqrt{n}} r^{\beta/2}((r + 1) E + 2^{i+1} \lambda)^{(1-\beta/2)/2},
\]

thus

\[
E\|P - P_n\|_{\mathcal{L}_{r,2^{i+1} \lambda}} \leq \max \left( \frac{r^2}{n^{3/(2+\beta)}}, \frac{r^{1+\beta/2}(2^{i+1} \lambda)^{1/2-\beta/4}}{\sqrt{n}} \right).
\]

Plugging the last result in the sum of Equation (25) entails the result.

\[
\square
\]

**Lemma 2.** Define

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

\[
\partial F := \{ g \in L^\infty : \langle g, f \rangle = 0, \quad \forall f \in F \}.
\]
and \( \mathcal{L}_c(C) = \{(Y - f(X))^2 - (Y - f^c(X))^2 : f \in C\} \), where \( f^c \in \text{argmin}_{g \in C} R(g) \). If 
\[
\text{max}(\|Y\|_\infty, \sup_{f \in F} \|f\|_\infty) \leq b, \text{ we have}
\]
\[
\mathbb{E} \left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^n f^2(X_i) \right] \leq c \max \left( \sigma^2(F), \frac{b^2 \log M}{n} \right), \text{ and}
\]
\[
\mathbb{E} \|P_n - P\|_{\mathcal{L}_c(C)} \leq cb \sqrt{\frac{\log M \log n}{n}} \max \left( b \sqrt{\frac{\log M}{n}}, d(F) \right).
\]
If \( \max(\|\epsilon\|_{\psi_2}, \sup_{f \in F} \|f(X) - f_0(X)\|_{\psi_2}) \leq b \), we have
\[
\mathbb{E} \left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^n f^2(X_i) \right] \leq c \max \left( \sigma^2(F), \frac{b^2 \log M \log n}{n} \right), \text{ and}
\]
\[
\mathbb{E} \|P_n - P\|_{\mathcal{L}_c(C)} \leq cb \sqrt{\frac{\log M \log n}{n}} \max \left( b \sqrt{\frac{\log M \log n}{n}}, d(F) \right).
\]

Proof. First, consider the case when \( \|\epsilon\|_{\psi_2}, \|f(X) - f_0(X)\|_{\psi_2} \leq b \). Define
\[
r^2 = \sup_{f \in F} \frac{1}{n} \sum_{i=1}^n f(X_i)^2,
\]
and note that \( \mathbb{E}_X(r^2) \leq \mathbb{E}_X \|P - P_n\|_{F_2} + \sigma(F)^2 \), where \( F = \{f^2 : f \in F\} \). Using the same argument as in the beginning of the proof of Lemma 1, see above, we have
\[
\mathbb{E}_X \|P - P_n\|_{F_2} \leq \frac{c}{n} \mathbb{E}_X \mathbb{E}_g \left[ \sup_{f \in F} \left| \sum_{i=1}^n g_i f^2(X_i) \right| \right].
\]
The process \( f \mapsto Z_{2,f} = \sum_{i=1}^n g_i f^2(X_i) \) is Gaussian, with intrinsic distance
\[
\mathbb{E}_g \|Z_{2,f} - Z_{2,f'}\| = \sum_{i=1}^n (f(X_i)^2 - f'(X_i)^2)^2 \leq d_{n,\infty}(f, f')^2 \times 4nr^2,
\]
where \( d_{n,\infty}(f, f') = \max_{i=1,\ldots,n} |f(X_i) - f'(X_i)| \). So, using Dudley’s entropy integral, we have
\[
\mathbb{E}_g \|Z_{2,f} - Z_{2,f'}\| \leq \frac{c}{n} \int_0^{\Delta_{n,\infty}(F)} \sqrt{\log N(F, d_{n,\infty}, t)} dt \leq c \Delta_{n,\infty}(F) \sqrt{\frac{\log M}{n}},
\]
where \( \Delta_{n,\infty} \) is the \( d_{n,\infty} \)-diameter of \( F \). So, we get
\[
\mathbb{E}_X \|P - P_n\|_{F_2} \leq \frac{c}{n} \sqrt{\frac{\log M}{n}} \mathbb{E}_X[\Delta_{n,\infty}(F)r] \leq c \sqrt{\frac{\log M}{n}} \sqrt{\mathbb{E}_X[\Delta_{n,\infty}^2(F)]} \sqrt{\mathbb{E}_X[r^2]},
\]
which entails that
\[
\mathbb{E}_X(r^2) \leq \frac{c \log M}{n} \mathbb{E}_X[\Delta_{n,\infty}^2(F)] + 4\sigma(F)^2.
\]
Since \( \mathbb{E}[Z^2] \leq 2\|Z\|_{\psi_2}^2 \) for a subgaussian variable \( Z \), we have, by using Pisier’s inequality,
\[
\mathbb{E}_X[\Delta_{n,\infty}^2(F)] \leq 4\| \max_{i=1,\ldots,n} \sup_{f \in F} |f(X_i) - f_0(X_i)|_{\psi_2}^2 \leq 4 \log(n+1) \| \sup_{f \in F} |f(X) - f_0(X)|_{\psi_2}^2 \leq 4b^2 \log(n+1),
\]

25
so we have proved that
\[ \mathbb{E}_X(r^2) \leq c \max \left( b^2 \frac{\log M \log n}{n}, \sigma(F)^2 \right). \]

When \( \|f\|_\infty \leq b \), the proof is easier, since we can use the contraction principle for Rademacher process after the symmetrization argument:
\[ \mathbb{E}_X \|P - P_n\|_{F^2} \leq \frac{2}{n} \mathbb{E}_X \mathbb{E}_\epsilon \left[ \sup_{f \in F} \left| \sum_{i=1}^n \epsilon_i f^2(X_i) \right| \right] \leq \frac{8b}{n} \mathbb{E}_X \mathbb{E}_\epsilon \left[ \sup_{f \in F} \left| \sum_{i=1}^n \epsilon_i f(X_i) \right| \right], \]
and one obtains from this, as previously, that
\[ \mathbb{E}_X(r^2) \leq c \max \left( b^2 \frac{\log M}{n}, \sigma(F)^2 \right). \]

Let us turn to the part of the Lemma concerning \( \mathbb{E}\|P - P_n\|_{\mathcal{L}_C(F)} \). Recall that \( \mathcal{C} = \text{conv}(F) \) and write for short \( \mathcal{L}_f(X,Y) = \mathcal{L}_C(f)(X,Y) = (Y - f(X))^2 - (Y - f^C(X))^2 \) for each \( f \in \mathcal{C} \), where we recall that \( f^C \in \text{argmin}_{g \in \mathcal{C}} R(g) \). Using the same argument as before we have
\[ \mathbb{E}\|P - P_n\|_{\mathcal{L}_C(F)} \leq \frac{c}{n} \mathbb{E}(X,Y) \mathbb{E}_\varphi \left[ \sup_{f \in \mathcal{C}} \left| \sum_{i=1}^n g_i \mathcal{L}_f(X_i,Y_i) \right| \right]. \]

Consider the Gaussian process \( f \to Z_f := \sum_{i=1}^n g_i \mathcal{L}_f(X_i,Y_i) \) indexed by \( \mathcal{C} \). For every \( f, f' \in \mathcal{C} \), the intrinsic distance of \( Z_f \) satisfies
\[ \mathbb{E}_g |Z_f - Z_{f'}|^2 = \sum_{i=1}^n (\mathcal{L}_f(X_i,Y_i) - \mathcal{L}_{f'}(X_i,Y_i))^2 \]
\[ \leq \max_{i=1,...,n} |2Y_i - f(X_i) - f'(X_i)|^2 \times \sum_{i=1}^n (f(X_i) - f'(X_i))^2 \]
\[ = \max_{i=1,...,n} |2Y_i - f(X_i) - f'(X_i)|^2 \times \mathbb{E}_g |Z_f - Z_{f'}|^2, \]
where \( Z_{f'} := \sum_{i=1}^n g_i (f(X_i) - f^C(X_i)) \). Therefore, by Slepian’s Lemma, we have for every \( (X_i,Y_i)_{i=1}^n \):
\[ \mathbb{E}_g \left[ \sup_{f \in \mathcal{C}} Z_f \right] \leq \max_{i=1,...,n} \sup_{f,f' \in \mathcal{C}} |2Y_i - f(X_i) - f'(X_i)| \times \mathbb{E}_g \left[ \sup_{f \in \mathcal{C}} Z_{f'} \right], \]
and since for every \( f = \sum_{j=1}^M \alpha_j f_j \in \mathcal{C} \), where \( \alpha_j \geq 0, \forall j = 1,...,M \) and \( \sum \alpha_j = 1 \), \( Z_{f'} = \sum_{j=1}^M \alpha_j Z_{f_j} \), we have
\[ \mathbb{E}_g \left[ \sup_{f \in \mathcal{C}} Z'_{f} \right] \leq \mathbb{E}_g \left[ \sup_{f \in \mathcal{C}} Z'_{f'} \right]. \]
Moreover, we have using Dudley’s entropy integral argument,
\[ \frac{1}{n} \mathbb{E}_g \left[ \sup_{f \in \mathcal{F}} Z'_{f} \right] \leq \frac{c}{\sqrt{n}} \int_0^{\Delta_n(F')} \sqrt{N(F',\|\cdot\|_n,t)} dt \leq c \sqrt{\frac{\log M}{n}} r', \]

26
where \( F' := \{ f - f^C : f \in F \} \) and \( \Delta_n(F') := \text{diam}(F', \| \cdot \|_n) \) and

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

On the other hand, we can prove, using Pisier’s inequality for \( \psi_1 \) random variables and the fact that \( \|U^2\|_{\psi_1} = \|U\|_{\psi_2}^2 \) for every random variable \( U \), that

\[
\sqrt{\mathbb{E}\left[ \max_{i=1,\ldots,n} |2Y_i - f(X_i) - f'(X_i)|^2 \right]} \leq 2\sqrt{2\log(n+1)}(\|\varepsilon\|_{\psi_1} + \sup_{f \in F} \|f(X) - f_0(X)\|_{\psi_2}).
\]

(27)

So, we finally obtain

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

and the conclusion follows from the first part of the Lemma, since \( \sigma(F') \leq d(F) \). The case \( \max(\|Y\|_{\infty}, \sup_{f \in F} \|f\|_{\infty}) \leq b \) is easier and follows from the fact that the left hand side of (27) is smaller than \( 4b \).

Theorem 5 combined with Lemma 2 leads to the following corollary.

**Corollary 1.** Let \( d(F) = \text{diam}(F, L^2(P_X)) \), \( \mathcal{C} := \text{conv}(F) \) and \( \mathcal{L}_f(X,Y) = (Y - f(X))^2 - (Y - f^C(X))^2 \). If \( \max(\|\varepsilon\|_{\psi_2}, \sup_{f \in F} \|f(X) - f_0(X)\|_{\psi_2}) \leq b \) we have, with probability larger than \( 1 - 4e^{-x} \), that for every \( f \in \mathcal{C} \):

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

If \( \max(\|Y\|_{\infty}, \sup_{f \in F} \|f\|_{\infty}) \leq b \), we have, with probability larger than \( 1 - 4e^{-x} \), that for every \( f \in \mathcal{C} \):

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

**Proof.** We apply Theorem 5 to the process

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

to obtain that with a probability larger than \( 1 - 4e^{-x} \):

\[
Z \leq c \left( \mathbb{E}Z + \sigma(\mathcal{C}) \sqrt{\frac{x}{n}} + b_n(\mathcal{C}) \frac{x}{n} \right),
\]

27
where
\[ \sigma(C)^2 = \sup_{f \in C} \mathbb{E}[\mathcal{L}_f(X,Y)^2], \]
and
\[ b_n(C) = \left\| \max_{i=1,\ldots,n} \sup_{f \in C} |\mathcal{L}_f(X_i, Y_i) - \mathbb{E}[\mathcal{L}_f(X,Y)]| \right\| \psi_1. \]

Since \( \mathcal{L}_f(X,Y) = 2\varepsilon(f^C(X) - f(X)) + (f^C(X) - f(X))(2f_0(X) - f(X) - f^C(X)) \), we have using Assumption 1
\[ \mathbb{E}[\mathcal{L}_f(X,Y)^2] \leq 4\sigma_e^2 \|f - f^C\|_{L^2(P_X)}^2 + 4 \sqrt{\mathbb{E}[(f^C(X) - f(X))^4]} \sqrt{\mathbb{E}[(2f_0(X) - f(X) - f^C(X))^4]}, \]
If \( U_f := (f^C(X) - f(X))^2 \) we have \( \|U_f\|_{\psi_1} = \|f^C - f\|_{\psi_1}^2 \leq (2b)^2 \) for any \( f \in C \), so using the \( \psi_1 \) version of Bernstein’s inequality (see van der Vaart and Wellner [1996]), we have that \( \mathbb{P}(|U_f - \mathbb{E}[U_f]| \geq m \|U_f\|_{\psi_1}) \leq 2 \exp(-c \min(m, m^2)) = 2 \exp(-cm) \) for any \( m \in \mathbb{N} - \{0\} \). But for such a random variable, one has \( \mathbb{E}(U_f)^{1/p} \leq c_p \mathbb{E}(U_f) \) for any \( p > 1 \) (cf. Mendelson [2004]). So, in particular for \( p = 2 \), we derive
\[ \sqrt{\mathbb{E}[(f^C(X) - f(X))^4]} \leq c\|f - f^C\|_{L^2(P_X)}^2. \]
Moreover, since \( \mathbb{E}(Z^4) \leq 16\|Z\|^4_{\psi_2} \), we have
\[ \sqrt{\mathbb{E}[(2f_0(X) - f(X) - f^C(X))^4]} \leq 8b^2. \]
So, we can conclude that
\[ \sigma(C)^2 \leq (4\sigma_e^2 + 8eb^2)d(F). \]
Since \( \mathbb{E}(Z) \leq \|Z\|_{\psi_1} \), we have \( b_n(C) \leq 2\log(n + 1)\|\sup_{f \in C} |\mathcal{L}_f(X,Y)|\|_{\psi_1} \). Moreover, a straightforward calculation gives \( \mathcal{L}_f(X,Y) \leq \varepsilon^2 + (f^C(X) - f_0(X))^2 + 3(f(X) - f_0(X))^2 \), so
\[ b_n(C) \leq 10\log(n + 1)b^2. \]
Putting all this together, and using Lemma 2 we arrive at
\[ Z \leq c(\sigma_e + b) \sqrt{\frac{(\log M + x) \log n}{n}} \max \left( b \sqrt{\frac{(\log M + x) \log n}{n}}, d(F) \right), \]
with probability larger than \( 1 - 4e^{-x} \) for any \( x > 0 \). In the bounded case where \( \max(\|Y\|_{\infty}, \sup_{f \in F} \|f\|_{\infty}) \leq b \), the proof is easier, and one can use the original Talagrand’s concentration inequality.

\[ \Box \]

**Lemma 3.** Let \( \mathcal{L}_f(X,Y) = (Y - f(X))^2 - (Y - f^F(X))^2 \). If we have \( \max(\|\varepsilon\|_{\psi_1}, \|\sup_{f \in F} |f(X) - f_0(X)|\|_{\psi_2}) \leq b \) we have, with probability larger than \( 1 - 4e^{-x} \), that for every \( f \in F \):
\[ \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_f(X_i, Y_i) - \mathbb{E}\mathcal{L}_f(X,Y) \right| \leq c(\sigma_e + b) \sqrt{\frac{(\log M + x) \log n}{n}} \max \left( b \sqrt{\frac{(\log M + x) \log n}{n}}, \|f - f^F\|_{\infty} \right). \]
Also, with probability at least $1 - 4e^{-x}$, we have for every $f, g \in F$:

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

When $\max(\|Y\|_\infty, \sup_{f \in F} \|f\|_\infty) \leq b$, we have, with probability larger than $1 - 2e^{-x}$, that for every $f \in F$:

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

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

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

**Proof of Lemma 4** The proof uses exactly the same arguments as that of Lemma 2 and Corollary 1, and thus is omitted.

**Lemma 4.** Let $\hat{F}_1$ be given by (5) and recall that $f_F \in \arg\min_{f \in F} R(f)$ and let $d(\hat{F}_1) = \text{diam}(\hat{F}_1, L_2(P_X))$. Assume that

$$\max(\|\varepsilon\|_{\psi_2}, \sup_{f \in F} |f(X) - f_0(X)|\|\psi_2\|) \leq b.$$

Then, with probability at least $1 - 4\exp(-x)$, we have $f_F \in \hat{F}_1$, and any function $f \in \hat{F}_1$ satisfies

$$R(f) \leq R(f_F) + c(\sigma + b) \sqrt{\frac{(\log M + x) \log n}{n}} \max \left( b \sqrt{\frac{(\log M + x) \log n}{n}}, d(\hat{F}_1) \right).$$

If $\max(\|Y\|_\infty, \sup_{f \in F} \|f\|_\infty) \leq b$, we have with probability at least $1 - 2\exp(-x)$ that $f_F \in \hat{F}_1$, and any function $f \in \hat{F}_1$ satisfies

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

**Proof.** The proof follows the lines of the proof of Lemma 4.4 in Lecué and Mendelson (2009a), together with Lemma 3, so we don’t reproduce it here.

**B Function spaces**

In this section we give precise definitions of the spaces of functions considered in the paper, and give useful related results. The definitions and results presented here can be found in Triebel (2006), in particular in Chapter 5 which is about anisotropic spaces, anisotropic multiresolutions, and entropy numbers of the embeddings of such spaces (see Section 5.3.3) that we use in particular to derive condition ($C_{\beta}$), for the anisotropic Besov space, see Section 3.
Let \( \{e_1, \ldots, e_d\} \) be the canonical basis of \( \mathbb{R}^d \) and \( \mathbf{s} = (s_1, \ldots, s_d) \) with \( s_i > 0 \) be a vector of directional smoothness, where \( s_i \) corresponds to the smoothness in direction \( e_i \). Let us fix \( 1 \leq p, q \leq \infty \). If \( f : \mathbb{R}^d \to \mathbb{R} \), we define \( \Delta^k_h f \) as the difference of order \( k \geq 1 \) and step \( h \in \mathbb{R}^d \), given by \( \Delta^k_h f(x) = f(x + h) - f(x) \) and \( \Delta^k_h f(x) = \Delta^k_h (\Delta^{k-1}_h f)(x) \) for any \( x \in \mathbb{R}^d \).

**Definition 3.** We say that \( f \in L^p(\mathbb{R}^d) \) belongs to the anisotropic Besov space \( B^s_{p,q}(\mathbb{R}^d) \) if the semi-norm

\[
|f|_{B^s_{p,q}(\mathbb{R}^d)} := \sum_{i=1}^d \left( \int_0^1 (t^{-s_i} \|\Delta^{k_i}_{i\mathbf{e}_i} f\|_p)^q \frac{dt}{t} \right)^{1/q}
\]

is finite (with the usual modifications when \( p = \infty \) or \( q = \infty \)).

We know that the norms

\[
\|f\|_{B^s_{p,q}} := \|f\|_p + |f|_{B^s_{p,q}}
\]

are equivalent for any choice of \( k_i > s_i \). An equivalent definition of the seminorm can be given using the directional differences and the anisotropic distance, see Theorem 5.8 in Triebel (2006).

Several explicit particular cases for the space \( B^s_{p,q} \) are of interest. If \( s = (s, \ldots, s) \) for some \( s > 0 \), then \( B^s_{p,q} \) is the standard isotropic Besov space. When \( p = q = 2 \) and \( s = (s_1, \ldots, s_d) \) has integer coordinates, \( B^s_{2,2} \) is the anisotropic Sobolev space

\[
B^s_{2,2} = W^s_2 = \left\{ f \in L^2 : \sum_{i=1}^d \left\| \frac{\partial^{s_i} f}{\partial x_i^{s_i}} \right\|_2 < \infty \right\}.
\]

If \( s \) has non-integer coordinates, then \( B^s_{2,2} \) is the anisotropic Bessel-potential space

\[
H^s = \left\{ f \in L^2 : \sum_{i=1}^d \left\| (1 + |\xi_i|^2)^{s_i/2} \hat{f}(\xi) \right\|_2 < \infty \right\}.
\]

As we mentioned below, Assumption A is satisfied for barely all smoothness spaces considered in nonparametric literature. In particular, if \( \mathcal{F} = B^s_{p,q} \) is the anisotropic Besov space defined above, \((C_3)\) is satisfied: it is a consequence of a more general Theorem (see Theorem 5.30 in Triebel (2006)) concerning the entropy numbers of embeddings (see Definition 1.87 in Triebel (2006)). Here, we only give a simplified version of this Theorem, which is sufficient to derive \((C_3)\) for \( B^s_{p,q} \). Indeed, if one takes \( s_0 = s, p_0 = p, q_0 = q \) and \( s_1 = 0, p_0 = \infty, q_0 = \infty \) in Theorem 5.30 from Triebel (2006), we obtain the following

**Theorem 9.** Let \( 1 \leq p, q \leq \infty \) and \( s = (s_1, \ldots, s_d) \) where \( s_i > 0 \), and let \( \bar{s} \) be the harmonic mean of \( s \) (see (13)). Whenever \( \bar{s} > d/p \), we have

\[
B^s_{p,q} \subset C(\mathbb{R}^d),
\]

where \( C(\mathbb{R}^d) \) is the set of continuous functions on \( \mathbb{R}^d \), and for any \( \delta > 0 \), the sup-norm entropy of the unit ball of the anisotropic Besov space, namely the set

\[
U^s_{p,q} := \{ f \in B^s_{p,q} : |f|_{B^s_{p,q}} \leq 1 \}
\]

30
satisfies
\[ H_\infty(\delta, U_{p,q}^\delta) \leq D \delta^{-\bar{s}/d}, \tag{28} \]
where \( D > 0 \) is a constant independent of \( \delta \).

For the isotropic Sobolev space, Theorem 9 was obtained in the key paper [Birman and Solomjak (1967)] (see Theorem 5.2 herein), and for the isotropic Besov space, it can be found, among others, in [Birgé and Massart (2000)] and [Kerkyacharian and Picard (2003)].

**Remark 3.** A more constructive computation of the entropy of anisotropic Besov spaces can be done using the replicant coding approach, which is done for Besov bodies in [Kerkyacharian and Picard (2003)]. Using this approach together with an anisotropic multiresolution analysis based on compactly supported wavelets or atoms, see Section 5.2 in [Triebel (2006)], we can obtain a direct computation of the entropy. The idea is to do a quantization of the wavelet coefficients, and then to code them using a replication of their binary representation, and to use 01 as a separator (so that the coding is injective). A lower bound for the entropy can be obtained as an elegant consequence of Hoeffding’s deviation inequality for sums of i.i.d. variables and a combinatorial lemma.

**References**

**Adamczak, R.** (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, **13** no. 34, 1000–1034.

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

**Birgé, L.** and **Massart, P.** (2000). An adaptive compression algorithm in Besov spaces. *Constr. Approx.*, **16** 1–36.

**Birman, M. Š.** and **Solomjak, M. Z.** (1967). Piecewise polynomial approximations of functions of classes \( W_p^\alpha \). *Mat. Sb. (N.S.),* **73** (115) 331–355.

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

**Cucker, F.** and **Smale, S.** (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.),* **39** 1–49 (electronic).

**Dalalyan, A. S.** and **Tsybakov, A. B.** (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *COLT.* 97–111.

**Dudley, R. M.** (1999). *Uniform central limit theorems,* vol. 63 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge.

**Efron, B., Hastie, T., Johnstone, I.** and **Tibshirani, R.** (2004). Least angle regression. *Ann. Statist.*, **32** 407–499. With discussion, and a rejoinder by the authors.
Einmahl, U. and Mason, D. M. (1996). Some universal results on the behavior of increments of partial sums. *Ann. Probab.*, **24** 1388–1407.

Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.*, **12** 929–998.

Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics, Springer-Verlag, New York.

Hoffmann, M. and Lepski, O. V. (2002). Random rates in anisotropic regression. *The Annals of Statistics*, **30** 325–396.

Juditsky, A., Rigollet, P. and Tsybakov, A. B. (2008). Learning by mirror averaging. *Ann. Statist.*, **36** 2183–2206. URL https://acces-distant.upmc.fr:443/http/dx.doi.org/10.1214/07-AOS546.

Juditsky, A. B., Nazin, A. V., Tsybakov, A. B. and Vayatis, N. (2005). Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii*, **41** 78–96.

Kerkyacharian, G., Lepski, O. and Picard, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields*, **121** 137–170.

Kerkyacharian, G., Lepski, O. and Picard, D. (2007). Nonlinear estimation in anisotropic multi-index denoising. Sparse case. *Teor. Veroyatn. Primen.*, **52** 150–171.

Kerkyacharian, G. and Picard, D. (2003). Replicant compression coding in Besov spaces. *ESAIM Probab. Stat.*, **7** 239–250 (electronic).

Lecué, G. (2006). Lower bounds and aggregation in density estimation. *J. Mach. Learn. Res.*, **7** 971–981.

Lecué, G. and Mendelson, S. (2009a). Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, **145** 591–613. URL https://acces-distant.upmc.fr:443/http/dx.doi.org/10.1007/s00440-008-0180-8.

Lecué, G. and Mendelson, S. (2009b). Sharper lower bounds on the performance of the empirical risk minimization algorithm. To appear in Bernoulli.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces*, vol. 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. Isoperimetry and processes.

Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1996). The importance of convexity in learning with squared loss. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*. ACM Press, 140–146.

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

Loustau, S. (2009). Penalized empirical risk minimization over besov spaces. *Electronic Journal of Statistics*, **3** 824–850.

32
Massart, P. (2007). *Concentration inequalities and model selection*, vol. 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Mendelson, S. (2004). On the performance of kernel classes. *J. Mach. Learn. Res.*, 4 759–771. URL [http://dx.doi.org/10.1162/1532443041424337](http://dx.doi.org/10.1162/1532443041424337).

Mendelson, S. (2008). Lower bounds for the empirical minimization algorithm. *IEEE Trans. Inform. Theory*, 54 3797–3803.

Mendelson, S. and Neeman, J. (2009). Regularization in kernel learning. Tech. rep. To appear in Annals of Statistics, available at [http://www.imstat.org/aos/](http://www.imstat.org/aos/)

Neumann, M. H. (2000). Multivariate wavelet thresholding in anisotropic function spaces. *Statist. Sinica*, 10 399–431.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58 267–288.

Triebel, H. (2006). *Theory of function spaces. III*, vol. 100 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel.

Tsybakov, A. (2003a). *Introduction l’estimation non-paramtrique*. Springer.

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

Van de Geer, S. A. (2000). *Applications of empirical process theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics, Springer-Verlag, New York. With applications to statistics.

Wahba, G. (1990). *Spline models for observational data*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.*, 28 75–87. URL [http://dx.doi.org/10.1214/aos/1016120365](http://dx.doi.org/10.1214/aos/1016120365)