MODEL SELECTION FOR REGRESSION ON A FIXED DESIGN

YANNICK BARAUD

Abstract. We deal with the problem of estimating some unknown regression function involved in a regression framework with deterministic design points. For this end, we consider some collection of finite dimensional linear spaces (models) and the least-squares estimator built on a data driven selected model among this collection. This data driven choice is performed via the minimization of some penalized model selection criterion that generalizes on Mallows' C_p. We provide non asymptotic risk bounds for the so-defined estimator from which we deduce adaptivity properties. Our results hold under mild moment conditions on the errors. The statement and the use of a new moment inequality for empirical processes is at the heart of the techniques involved in our approach.

1. Introduction

Let ψ_j for $j = 1, ..., n$ be some basis of $\mathbb{R}^n$ which is endowed with the normalized Euclidean norm $\|.\|_n$ defined by $\|t\|_n^2 = n^{-1}\sum_{i=1}^n t_i^2$ for $t \in \mathbb{R}^n$. To start with, let us consider the problem of estimating the unknown $\mathbb{R}^n$-vector $s = \sum_{j=1}^n \beta_j\psi_j = \mathbb{E}[Y]$ from one realization of the random vector $Y \in \mathbb{R}^n$ deriving from

$$\text{Model 1} \quad Y = \sum_{j=1}^n \beta_j\psi_j + \varepsilon,$$

where ε denotes some centered random vector of $\mathbb{R}^n$ with i.i.d. components admitting a finite variance σ^2. For the sake of simplicity σ^2 is supposed to be known all along this section. The classical linear model relies on the assumption that s belongs to the linear span, S_{m_0}, of $\{\psi_j,\ j \in m_0\}$ where m_0 is a known subset of $[1, ..., n]$. One well-known method to estimate s is to use the least-squares estimator, $\hat{s}_{m_0}$, which minimizes the least-squares contrast function $\gamma_n(t) = n^{-1}\sum_{i=1}^n (Y_i - t_i)^2$ over the vectors t of S_{m_0}. Since the quadratic risk $\mathbb{E}[\|s - \hat{s}_{m_0}\|_n^2]$ of this estimator is equal to $\sigma^2|m_0|/n$, we see that $\hat{s}_{m_0}$ behaves poorly when $|m_0|$ is large. To improve the estimation of s let us consider the least-squares estimator $\hat{s}_m$ with respect to S_m

1991 *Mathematics Subject Classification.* Primary 62G07; Secondary 62J02, 60E15.
Key words and phrases. Nonparametric regression, Least-squares estimator, Model selection, Adaptive estimation, Moment inequality, Concentration of measure, Empirical processes.

for some $m \subset m_0$. We set $\mathrm{d}_n(s, S_m) = \inf_{t \in S_m} \|s - t\|_n$. Since the quadratic risk of $\hat{s}_m$ is equal to

$$R_n(m) = \mathbb{E}\left[\|s - \hat{s}_m\|_n^2\right] = \mathrm{d}_n^2(s, S_m) + \sigma^2|m|/n, \tag{1}$$

there exists some optimal choice m^* of m among the subsets of m_0 that realizes the best trade-off between the bias term $\mathrm{d}_n^2(s, S_m)$ and the variance term $\sigma^2|m|/n$. Unfortunately m^* depends on the unknown coefficients β_js. The problem of variable selection is to determine from the data some subset $\hat{m} \subset m_0$ for which the estimator $\tilde{s} = \hat{s}_{\hat{m}}$ admits a quadratic risk that is as close as possible to the infimum of the risks of the least-squares estimators $\hat{s}_m$ when m varies among the collection of all subsets of m_0. To solve this problem an heuristic approach was given by Mallows (1973). He suggested that $\hat{m}$ should be chosen to minimize the penalized criterion (Mallows' C_p) $\gamma_n(\hat{s}_m) + 2\sigma^2|m|/n$.

In a parametric framework, i.e. when m_0 is given by $[1, ..., N]$ for some integer N independent of n, Nishii (1984) considered the problem of variable selection under a Gaussian assumption on the ε_is and for n tending to infinity. In this asymptotic context the problem is then to determine $m^* = \{j/\ \beta_j \neq 0\}$. For this purpose Nishii studied different kind of penalized criteria and among them one that is similar to Mallows' C_p: $\hat{m}$ is obtained by minimizing over the subsets m of $[1, ..., N]$ $\gamma_n(\hat{s}_m) + a\sigma^2|m|/n$ for some arbitrary positive constant a. Nishii gave the exact asymptotics of $n\mathbb{E}[\|s - \tilde{s}\|_n^2]$ and showed that if this criterion fails to determine m^* asymptotically, one selects a model $\hat{m}$ that contains m^* with a probability that tends to one as n becomes large. Nevertheless, as N is fixed and n becomes large $\tilde{s}$ converges towards s at rate $1/\sqrt{n}$.

Now consider the problem of estimating the unknown function s mapping $\mathcal{X}$ into $\mathbb{R}$ from the sample of size n (Y_i, x_i) obtained from

$$\text{Model 2} \quad Y_i = s(x_i) + \varepsilon_i,\ i = 1, ..., n.$$

The design points x_is are deterministic points of $\mathcal{X}$ (not necessarily distinct) and the errors ε_is are unobservable i.i.d. centered random variables with common finite variance σ^2. When we deal with Model 2, we equip $\mathcal{X}$ with the measure μ_n given by $\mu_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$. By considering the mapping I from $\mathbb{L}^2(\mathcal{X}, \mu_n)$ to $\mathbb{R}^n$ defined by

$$f \ \mapsto\ I(f) = \ {}^t(f(x_1), ..., f(x_n)),$$

we have that the functional space $(\mathbb{L}^2(\mathcal{X}, \mu_n), \|\ \|_{\mu_n})$ is isometric to $\mathbb{R}^n$ (or to a subspace of $\mathbb{R}^n$ when the x_is are not distinct) endowed with the norm $\|\ \|_n$ defined previously. In the sequel we shall denote the same way $\|\ \|_{\mu_n}$ and $\|\ \|_n$, emphasizing thus the links between Model 1 and Model 2. To illustrate the relevance of our method of estimation, let us assume for a short time that for some $\alpha \in]0, 1[$ and $L > 0$, s belongs to the set of α-Hölderian functions $\mathcal{H}_\alpha(L)$ defined by

$$\mathcal{H}_\alpha(L) = \{s/\ |s(x) - s(y)| \leq L|x - y|^\alpha\ \forall x, y \in [0, 1]\}.$$

In addition, let us assume that the x_is are equidistant points of $[0,1]$, i.e. $x_i = i/n$ for all $i = 1, ..., n$. When s is known to belong this set, we describe two ways of estimating s:

1. One can introduce the linear spaces S_ms for $m = 1, ..., n$, where S_m is defined as the linear span of the $\mathbf{1}_{[(j-1)/m, j/m[}$ for $j = 1, ..., m$. We denote by $\hat{s}_m$ the least-squares estimator of s in $S_m \subset \mathbb{L}^2(\mathcal{X}, \mu_n)$, which is defined by
$$\hat{s}_m = \operatorname*{argmin}_{t \in S_m} \sum_{i=1}^{n} (Y_i - t(x_i))^2 ,$$
and thanks to the isometry I we know from (1) that
$$\mathbb{E}\left[\|s - \hat{s}_m\|_n^2\right] = \mathrm{d}_n^2(s, S_m) + \frac{m}{n}\sigma^2. \tag{2}$$
As s belongs to $\mathcal{H}_\alpha(L)$ a simple computation shows that the bias term $\mathrm{d}_n^2(s, S_m)$ is bounded by $L^2 m^{-2\alpha}$ and therefore a possible choice of m in (2) to realize the best trade-off between the bias and the variance term is given by $m(\alpha, L) = [(L/\sigma)^{2/(1+2\alpha)} n^{1/(1+2\alpha)}]$ ($[x]$ denotes the integer part of x). We notice that this particular choice of m leads to a risk of $\hat{s}_{m(\alpha,L)}$ of order $\sigma^{4\alpha/(1+2\alpha)} L^{2/(1+2\alpha)} n^{-2\alpha/(1+2\alpha)}$ which is known to be (up to a constant) the minimax risk on $\mathcal{H}_\alpha(L)$.
2. One can consider the least-squares estimator of s in $\mathcal{H}_\alpha(L)$. This approach was considered by van de Geer (1990) in the case of Sobolev balls. By introducing empirical processes techniques, she related the size (in some sense) of this set to the rate of convergence of the estimator. By so doing, she proved that such an estimator reaches the minimax rate of convergence (up to a multiplicative constant).

The defect of both approaches lies in the fact that they heavily rely on a *prior* information on the regularity of s that is seldom available in practice. In the first example, a selection procedure of some $\hat{m}$ among $[1, ..., n]$ solely based on the data offers the advantage to free the estimator from any *prior* knowledge of α and L. Of course the selection procedure is relevant if for all $\alpha \in]0,1[$ and $L > 0$ $\tilde{s} = \hat{s}_{\hat{m}}$ is proved to perform almost as well as $\hat{s}_{m(\alpha,L)}$ under the *posterior* information that s belongs to $\mathcal{H}_\alpha(L)$. If so, the resulting estimator is said to be *adaptive* in the minimax sense with respect to the class of Hölderian functions (for a precise definition we refer to Barron, Birgé and Massart (1999)).

The aim of this paper is to propose a model selection procedure (the word "model" will be repeatedly used to name the S_ms) by penalized least-squares which is relevant for both variable selection and adaptive estimation (in the minimax sense). This is actually possible since our approach is not asymptotic. Unlike the parametric approach, it must be emphasized that we consider collections of models where the dimension and the number of models are both allowed to depend on n. This makes it possible to take collections of models that properly approximate Hölderian functions and consequently to derive properties of adaptivity in the minimax sense on $\tilde{s}$.

Let us now describe our estimation procedure in details. We start with some collection of linear subspaces $(S_m)_{m\in\mathcal{M}_n}$ of $\mathbb{R}^n$, in the case of Model 1, or of $\mathbb{L}^2(\mathcal{X},\mu_n)$ in the case of Model 2. Our selection strategy consists in estimating $m^*(n)$ that minimizes $R_n(m)$ over $\mathcal{M}_n$ by $\hat{m}$ that minimizes among $m\in\mathcal{M}_n$ the penalized least-squares criterion $\gamma_n(\hat{s}_m)+\mathrm{pen}_n(m)$ where pen_n is some positive function defined on $\mathcal{M}_n$. The main issue is to discuss whether it is possible to define the penalty function pen_n in such a way that the resulting estimator $\tilde{s}=\hat{s}_{\hat{m}}$ performs almost as well as $\hat{s}_{m^*(n)}$. Under adequate assumptions on the distribution of the ε_is and the collections of models, a choice of a penalty of the form $\mathrm{pen}_n(m)=C\sigma^2 D_m/n$ with $C>1$ leads to risk bounds of the form

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right]\leq C'\inf_{m\in\mathcal{M}_n}\mathbb{E}\left[\|s-\hat{s}_m\|_n^2\right]=C'\inf_{m\in\mathcal{M}_n}\left(\mathrm{d}_n^2(s,S_m)+\sigma^2\frac{D_m}{n}\right). \tag{3}$$

Our method generalizes the well-known model selection method introduced by Mallows and known as Mallows' C_p. It amounts to take $\mathrm{pen}_n(m)=2\sigma^2D_m/n$ ($C=2$) for all $m\in\mathcal{M}_n$. As a consequence of (3), when $\mathbb{E}[|\varepsilon_1|^p]<\infty$ for some $p>2$, we show that building adaptive estimators (in the minimax sense) with respect to classical smoothness classes is possible. Moreover in the context of Model 1 our results are also meaningful for the ordered variable selection problem. To solve this problem we consider the collection of models given by the S_ms defined as the linear span of the ψ_js for $j\in m$ and m varying among all the subsets of $\{1,...,n\}$ of the form $\{1,...,J\}$ ($1\leq J\leq n$). The original variable selection problem mentioned at the beginning requires that m should vary among all the subsets of $[1,...,n]$ (i.e. $|\mathcal{M}_n|=2^n-1$), but we believe that very restrictive integrability conditions on the ε_is are then necessary. Since we are interested in weak integrability conditions on the ε_is, the method that is presented here only covers the case of collections of models with a polynomial restriction (relatively to n) on the number of models.

As far as we know, the first results about adaptation in the minimax sense are due to Efroimovich and Pinsker (1984) in the white noise model. They got very precise asymptotic results using a method which is more sophisticated than a model selection procedure. Considering the estimator $\tilde{s}$ built using the Mallows' C_p method, in the context of Model 2, Li (1987) (and earlier Shibata (1981) when the ε_is are Gaussian, see also Kneip (1994) for extension of the work of Li) showed that

$$\frac{\|s-\tilde{s}\|_n^2}{\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_m\|_n^2}\to 1 \quad\text{in probability}$$

assuming that the ε_is have moments of order 8. Assuming moments of order 4 only, Polyak and Tsybakov (1990) gave further results by using the Fourier expansion of s. Unfortunately, all the above mentioned results of Shibata, Li and Polyak and Tsybakov are of an asymptotic nature and only hold under the unpleasant assumption that s does not belong to any of the models S_m which excludes the classical parametric case.

In contrast, as announced in (3) we shall give non asymptotic risk bounds for the penalized least-squares estimators which are valid for all s (belonging to some S_m or not). As a consequence if s belongs to one of the models the method allows to recover the usual parametric rate of convergence $1/\sqrt{n}$ of $\tilde{s}$ to s when n tends to infinity.

To our knowledge, the first non asymptotic risk bounds for estimators built via model selection are due to Birgé and Massart (1997) in a density estimation context. They also showed how such bounds imply adaptive properties in the minimax sense for the corresponding estimators. Related results are to be found in Barron, Birgé and Massart (1999) for various statistical frameworks including regression. While our approach is inspired by their work, the techniques that are used here differ from those used by Barron, Birgé and Massart in their treatment of the regression framework. Their results concerning regression have indeed the following weaknesses:

- the ε_is are supposed to have exponential moments;
- the models and the regression function s are assumed to be bounded by some known constant B;
- the value of the penalty involves unpleasant quantities such as B and irrelevant numerical constants.

In order to relax the boundedness assumptions and to weaken the integrability condition on the ε_is, we introduce some new probabilistic tools which allow to control the supremum of unbounded empirical processes. More precisely we state a moment inequality for the supremum of an empirical process over some class of functions $\mathcal{G}$, controlling its fluctuations around its mean, which does not require that the class $\mathcal{G}$ should be uniformly bounded. This inequality can be seen as an analogue of Talagrand's Inequality (1996) (Theorem 1.4) for unbounded empirical processes. Moreover, this technique allows us to provide an explicit value for the constant involved in the penalty term.

The paper is organized as follows. The description of the statistical framework including the definition of the estimator is given in Section 2. The main statistical result is to be found in Section 3. This result assumes the variance σ^2 of the errors to be known. The case of an unknown variance will be treated in Section 6.

The properties of adaptation of the estimator is the subject of Section 4. Section 5 is devoted to the statement of a moment inequality on suprema of empirical processes and its consequences. The most technical proofs are given in Section 7. Some moment inequalities on sum of independent and centered random variables are recalled in the Appendix.

Throughout this paper C, C', C'' ... denote constants that may vary from line to line. The notation $C(.)$ specifies the dependency of C on some quantities.

2. The statistical framework

We observe a sample of pairs (Y_i, x_i), $i = 1, \cdots, n$ where the Y_is are independent real valued random variables and the x_is are deterministic design points with values in some measurable space $\mathcal{X}$. The Y_is and the x_is are constrained by the relation

$$Y_i = s(x_i) + \varepsilon_i, \tag{4}$$

where the real variables $\varepsilon_1, ..., \varepsilon_n$ are unobservable i.i.d. centered random variables with common variance σ^2. The object of this paper is to estimate the unknown function s from $\mathcal{X}$ to $\mathbb{R}$ thanks to the observation of those pairs. Note that Model 1 is a particular case of (4) by taking $\mathcal{X} = \{1, ..., n\}$ and $x_i = i$ for $i = 1, ..., n$.

For any $t \in \mathbb{L}^2(\mathcal{X}, \mu_n)$ we define the least-squares loss function by

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - t(x_i))^2. \tag{5}$$

Let us now consider a finite collection of linear subspaces of $\mathbb{L}^2(\mathcal{X}, \mu_n)$ denoted by $(S_m)_{m\in\mathcal{M}_n}$. We allow one of the S_ms to be reduced to $\{0\}$. The minimum of γ_n over S_m is acheived at a single point $\hat{s}_m$ called the Least-Squares Estimator of s in $S_m \subset \mathbb{L}^2(\mathcal{X}, \mu_n)$. Given some positive penalty function pen mapping $\mathcal{M}_n$ into $\mathbb{R}_+$, we define the Penalized Least-Squares Estimator (PLSE for short), $\tilde{s}$, by

$$\tilde{s} = \hat{s}_{\hat{m}} \tag{6}$$

where $\hat{m}$ is chosen to minimize over $\mathcal{M}_n$ the penalized criterion

$$\gamma_n(\hat{s}_m) \ + \ \mathrm{pen}(m). \tag{7}$$

Thus,

$$\tilde{s} = \arg\min_{\substack{m\in\mathcal{M}_n \\ t\in S_m}} \left[\gamma_n(t) \ + \ \mathrm{pen}(m)\right]$$

and therefore $\tilde{s}$ satisfies

$$\tilde{s} \in S_{\hat{m}} \text{ and } \gamma_n(\tilde{s}) + \mathrm{pen}(\hat{m}) \le \gamma_n(t) + \mathrm{pen}(m) \tag{8}$$

for all $m \in \mathcal{M}_n$ and $t \in S_m$.

Notations: for $t \in \mathbb{L}^2(\mathcal{X}, \mu_n)$ we set

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} t^2(x_i)$$

and $\mathrm{d}_n^2(s, S_m) = \inf\{\|s-t\|_n^2/\ t \in S_m\}$. We denote by D_m the dimension of the linear space S_m.

3. Main Results

Throughout this section the variance σ^2 of the ε_is is assumed to be known. For the case of an unknown σ^2, which is of practical relevance, we refer to Section 6.

3.1. Risk bounds in $(\mathbb{L}^2(\mathcal{X}, \mu_n), \|.\|_n)$.

As already mentioned in the introduction, an ideal selection procedure would select some $m^* \in \mathcal{M}_n$ that minimizes the function $m \mapsto \mathbb{E}[\|s - \hat{s}_m\|_n^2]$ over $m \in \mathcal{M}_n$, the minimum, denoted by M_n^*, representing the minimal risk that can be achieved with the collection of estimators $(\hat{s}_m)_{m\in\mathcal{M}_n}$. We show that for suitable choices of the penalty function the values of $\|s - \tilde{s}\|_n^2$ are closed to M_n^*. More precisely, the following result holds:

Theorem 3.1. *Consider the regression framework (4) and let $(S_m)_{m\in\mathcal{M}_n}$ be some finite collection of finite dimensional linear subspaces of $\mathbb{L}^2(\mathcal{X}, \mu_n)$. For each $m \in \mathcal{M}_n$, let $\hat{s}_m$ be the least-squares estimator of s in S_m. Set*

$$M_n^* = \inf_{m\in\mathcal{M}_n} \mathbb{E}\left[\|s - \hat{s}_m\|_n^2\right].$$

For any positive number θ let us define $\mathrm{pen}:\ \mathcal{M}_n \to \mathbb{R}_+$ *by*

$$\mathrm{pen}(m) = (1+\theta)\frac{D_m}{n}\sigma^2.$$

Let $q > 0$ be given such that there exists $p > 2(1+q)$ satisfying $\mathbb{E}[|\varepsilon_1|^p] < +\infty$. Then for some constant $\kappa(\theta) > 1$, the PLSE, $\tilde{s}$, defined by (6) satisfies

$$\mathbb{E}\left[\left(\|s - \tilde{s}\|_n^2 - \kappa(\theta)M_n^*\right)_+^q\right] \le \Delta_p^q \frac{\sigma^{2q}}{n^q}, \tag{9}$$

where

$$\Delta_p^q = C'(\theta,p,q)\frac{\mathbb{E}[|\varepsilon_1|^p]}{\sigma^p}\left(1 + \sum_{\substack{m\in\mathcal{M}_n\\ D_m\ge 1}} D_m^{-(p/2-1-q)}\right). \tag{10}$$

Comments:

- The Mallow's C_p criterion corresponds to the choice $\theta = 1$.
- Δ_p is a quantity that does not depend on s but rather on the collection of models and the moments on the ε_is. We say that a collection $(S_m)_{m\in\mathcal{M}_n}$ is more complex than $(S'_m)_{m\in\mathcal{M}'_n}$ if $\mathcal{M}'_n \subset \mathcal{M}_n$. It should be noticed that Δ_p increases with the "complexity" of the collection of models. In some sense Δ_p evaluates the complexity of the collection of models with respect to q and the integrability properties of the ε_is.
- It comes from the proof of Theorem 3.1 that one can take $\kappa(\theta) = 2(1+4/\theta)(1+\theta)$. Since $\kappa(\theta)$ increases toward infinity as θ tends to 0, in practice it does not seem reasonable to choose θ small. As it is hard to determine an optimal choice of θ from theoretical computations, simulations should be carried out to determine it.

We straightforwardly derive the following corollary about the risk of $\tilde{s}$:

Corollary 3.1. *Under the assumptions of Theorem 3.1, the PLSE, $\tilde{s}$, defined by (6) satisfies*

$$\left(\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\right]\right)^{1/q} \le 2^{(q^{-1}-1)_+}\left[\kappa(\theta)\inf_{m\in\mathcal{M}_n}\left(\mathrm{d}_n^2(s,S_m)+\frac{D_m}{n}\sigma^2\right)+\frac{\Delta_p}{n}\sigma^2\right], \tag{11}$$

where Δ_p is defined by (10).

Proof. We recall from (1) that

$$M_n^* = \inf_{m\in\mathcal{M}_n}\left(\mathrm{d}_n^2(s,S_m)+\frac{D_m}{n}\sigma^2\right).$$

Since

$$\left(\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\right]\right)^{1/q} \le \left(\mathbb{E}\left[\left(\kappa(\theta)M_n^* + \left(\|s-\tilde{s}\|_n^2-\kappa(\theta)M_n^*\right)_+\right)^q\right]\right)^{1/q}$$

it follows from Minkowski's inequality when $q \ge 1$ or convexity arguments when $0 \le q < 1$ that

$$\left(\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\right]\right)^{1/q} \le 2^{(q^{-1}-1)_+}\left[\kappa(\theta)M_n^* + \left(\mathbb{E}\left[\left(\|s-\tilde{s}\|_n^2-\kappa(\theta)M_n^*\right)_+^q\right]\right)^{1/q}\right].$$

The result then follows from (9). $\square$

3.2. **Risk bounds in $\mathbb{L}^2(\mathcal{X},\nu)$.** Let ν be some measure on $\mathcal{X}$ and let us denote by $\|\ \|_\nu^2$ the corresponding $\mathbb{L}^2$-norm. The results of this section are very similar to those given in Corollary 3.1, the difference lying in the fact that we no longer express the distance between s and S_m in terms of the discrete norm but rather in terms of the norm relative to $\mathbb{L}^2(\mathcal{X},\nu)$. This is actually possible thanks to the equivalence between norms over finite dimensional linear spaces.

Given a function t from $\mathcal{X}$ into $\mathbb{R}$ we denote by $\|t\|_\infty$, the quantity $\sup_{x\in\mathcal{X}}|t(x)|$ and we set $\mathcal{L}_\infty(\mathcal{X})$ for the space of functions t from $\mathcal{X}$ into $\mathbb{R}$ such that $\|t\|_\infty < +\infty$. Throughout this section we assume that s belongs to $\mathcal{L}_\infty(\mathcal{X})$. For any finite dimensional space S in $\mathbb{L}^2(\mathcal{X},\nu)\cap\mathcal{L}_\infty(\mathcal{X})$ we denote by $\mathrm{d}_\nu(s,S)=\inf\{\|s-t\|_\nu/\ t\in S\}$ and $\mathrm{d}_\infty(s,S)=\inf\{\|s-t\|_\infty/\ t\in S\}$.

Corollary 3.2. *Consider the regression framework (4) and assume that s belongs to $\mathbb{L}^2(\mathcal{X},\nu)\cap\mathcal{L}_\infty(\mathcal{X})$. Let $\mathcal{S}_n$ be some finite dimensional space of $\mathbb{L}^2(\mathcal{X},\nu)\cap\mathcal{L}_\infty(\mathcal{X})$ satisfying the condition that there exists some positive number R_n such that*

$$\sup_{\substack{t\in\mathcal{S}_n\\ t\neq 0}}\frac{\|t\|_n}{\|t\|_\nu}\leq R_n. \tag{12}$$

Let $(S_m)_{m\in\mathcal{M}_n}$ be some finite family of linear subspaces of Sn and $q>0$ such that there exists $p>2(1+q)$ satisfying $\mathbb{E}[|\varepsilon_1|^p]<+\infty$. Then the estimator $\tilde{s}$ defined by (6) satisfies

$$\begin{aligned}&\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\right]\\ &\leq\ C_1\left[\inf_{m\in\mathcal{M}_n}\left(\mathrm{d}_\nu^2(s,S_m)\ +\ \frac{D_m}{n}\sigma^2\right)+\mathrm{d}_\infty^2(s,\mathcal{S}_n)+\frac{\Delta_p}{n}\sigma^2\right]^q,\end{aligned}\tag{13}$$

where $C_1=C(\theta,q,R_n)$ and Δ_p is defined by (10).

If there exists a positive constant r_n such that

$$\sup_{\substack{t\in\mathcal{S}_n\\ t\neq 0}}\frac{\|t\|_\nu}{\|t\|_n}\leq r_n, \tag{14}$$

then we can control $\|s-\tilde{s}\|_\nu$ in the following way

$$\begin{aligned}&\mathbb{E}\left[\|s-\tilde{s}\|_\nu^{2q}\right]\\ &\leq\ C_2\left[\inf_{m\in\mathcal{M}_n}\left(\mathrm{d}_\nu^2(s,S_m)\ +\ \frac{D_m}{n}\sigma^2\right)+\mathrm{d}_\infty^2(s,\mathcal{S}_n)+\frac{\Delta_p}{n}\sigma^2\right]^q,\end{aligned}\tag{15}$$

where $C_2=C(\theta,q,R_n,r_n)$.

Proof. Let us denote by π_m the $\mathbb{L}^2(\mathcal{X},\nu)$-orthogonal projector onto S_m. For any $s_n\in\mathcal{S}_n$, we have that $\mathrm{d}_n(s,S_m)\leq\|s-\pi_m s\|_n\leq\|s-s_n\|_n+\|s_n-\pi_m s\|_n$. Since $s_n-\pi_m s\in\mathcal{S}_n$ we know from (12) that $\|s_n-\pi_m s\|_n\leq R_n\|s_n-\pi_m s\|_\nu\leq R_n(\|s-s_n\|_\nu+$

$\|s-\pi_m s\|_\nu)$. Thus for any $s_n \in \mathcal{S}_n$, $\mathrm{d}_n(s,S_m) \leq (1+R_n)\|s-s_n\|_\infty + R_n\|s-\pi_m s\|_\nu$, therefore $\mathrm{d}_n(s,S_m) \leq (1+R_n)\mathrm{d}_\infty(s,\mathcal{S}_n) + R_n\mathrm{d}_\nu(s,S_m)$ and (13) follows from (11). Thanks to (14), similar arguments lead to $\|s-\tilde{s}\|_\nu \leq (1+r_n)\|s-s_n\|_\infty + r_n\|s-\tilde{s}\|_n$ and (15) follows from (13). □

Comment: In practice, the right-hand sides of (12) and (14) can be computed via the evaluation of spectral radii of Gramm matrices. More precisely, let A be some matrix of $\mathbb{M}_k(\mathbb{R})$ ($k \in \mathbb{N}^*$) and let us denote by $\rho(A)$ its spectral radius, which is defined by

$$\rho(A) = \sup_{x \neq 0} \frac{\|Ax\|_k}{\|x\|_k}. \tag{16}$$

Then the following result holds:

Lemma 3.1. *Let $(\varphi_\lambda)_{\lambda\in\Lambda_n}$ be an orthonormal basis of $\mathcal{S}_n \subset \mathbb{L}^2(\mathcal{X},\nu)$. Set Φ_n, the Gramm matrix*

$$\Phi_n = \left(\frac{1}{n}\sum_{i=1}^{n}\varphi_\lambda(x_i)\varphi_{\lambda'}(x_i)\right)_{\lambda,\lambda'\in\Lambda_n}.$$

Then we have

$$\sup_{t\in\mathcal{S}_n\setminus\{0\}} \frac{\|t\|_n^2}{\|t\|_\nu^2} = \rho(\Phi_n). \tag{17}$$

If Φ_n is positive definite, we also have

$$\sup_{t\in\mathcal{S}_n\setminus\{0\}} \frac{\|t\|_\nu^2}{\|t\|_n^2} = \rho(\Phi_n^{-1}). \tag{18}$$

The proof is deferred to Section 7.5.

4. Adaptivity properties of the PLSE in the minimax sense

In the sequel we assume that $\mathbb{L}^2(\mathcal{X},\nu) = \mathbb{L}^2([0,1],dx)$ and that $x_i = i/n$, $i = 1,...,n$. This section is devoted to asymptotic properties of the PLSE, i.e. n is no longer fixed but allowed to increase towards infinity. As a consequence of Corollary 3.2, we show that under the assumption that $\mathbb{E}[|\varepsilon_1|^p] < +\infty$ for some $p > 2$ and for a suitable choice of the family of models, $\tilde{s}$ is adaptive in the minimax sense simultaneously over Besov spaces of the form $\mathcal{B}_{\alpha,l,\infty}$ with $\alpha > 1/l$ and $l \geq 2$ (for a precise definition of Besov spaces we refer to DeVore and G.Lorentz (1993)). In order to prove such a result we introduce collections of models for which both the cardinality and the dimension of the models are allowed to depend on n. Moreover each collection is chosen in order to satisfy three important properties:

- The collection is not too "complex" with respect to the integrability properties of the errors ε_is i.e. there exists some positive q and some $p > 2(q+1)$ such that $\mathbb{E}[|\varepsilon_1|^p] < +\infty$ and such that $\Delta_p = \Delta_p(n)$ remains bounded as n becomes large.

- On each model of the collection $\| \ \|_n$ remains controlled by $\| \ \|_\nu$, i.e the quantity R_n involved in equation (12) remains bounded as n becomes large.

- Each model of the collection provides a good linear approximation of functions belonging to Besov spaces $\mathcal{B}_{\alpha,l,\infty}$ ($\alpha > 1/l$ and $l \geq 2$) with respect to both norms $\| \ \|_\nu$ and $\| \ \|_\infty$. Namely, for each $s \in \mathcal{B}_{\alpha,l,\infty}$

  $$\mathrm{d}_\nu(s, S_m) \leq C(\alpha)|s|_{\alpha,2} D_m^{-\alpha} \text{ and } \mathrm{d}_\infty(s, S_m) \leq C'(\alpha)|s|_{\alpha,l} D_m^{-\alpha+1/l}$$

  where $|.|_{\alpha,l}$ denotes the semi-norm associated to the Besov space $\mathcal{B}_{\alpha,l,\infty}$.

In the sequel $[x]$ denotes the integer part of x and we set $m_n = \max\{m \in \mathbb{N}, 2^m \leq n\}$. Two collections of models satisfying the required properties are described below:

(a) $\mathcal{M}_n = [0, ..., m_n]$, S_m (resp. $\mathcal{S}_n$) is the space of piecewise polynomials of degree less or equal r based on the dyadic grid $\{j2^{-m}/0 \leq j \leq 2^m\}$ (resp. $\{j2^{-m_n}/0 \leq j \leq 2^{m_n}\}$).

(b) $\mathcal{M}_n = [0, ..., m_n - 2]$, S_m (resp. $\mathcal{S}_n$) is the space of trigonometric polynomials of degree less or equal 2^m (resp. $2^{m_n - 2}$).

For those collections the following holds

Proposition 4.1. *For each $\alpha > 0$ ($\alpha < r+1$ in the case of the collection (a)) and $L > 0$ we set $l = (2\alpha+1)/(2\alpha^2)$. Consider either the collection of models (a) or (b) and assume that $\mathbb{E}[|\varepsilon_1|^p] < \infty$ for some $p > 2$. Then the PLSE $\tilde{s}$ defined by (6) satisfies for each positive $q < p/2 - 1$*

$$\left(\sup_{|s|_{\alpha,l} \leq L} \mathbb{E}\left[\|s - \tilde{s}\|_n^{2q} \right] \right)^{1/(2q)} \leq C(L, \alpha) n^{-\frac{\alpha}{2\alpha+1}}. \tag{19}$$

In the case of the collection (b), $\|s - \tilde{s}\|_n$ can be replaced by $\|s - \tilde{s}\|_\nu$.

The proof is deferred to Section 7.4.

Comment: Note that for each $l' \geq l$, $\mathcal{B}_{\alpha,l',\infty} \subset \mathcal{B}_{\alpha,l,\infty}$ and for all $s \in \mathcal{B}_{\alpha,l,\infty}$, $|s|_{\alpha,l} \leq |s|_{\alpha,l'}$. Thus, (19) also holds by replacing l by l'.

5. Moment inequalities for empirical processes

The proof of our main theorem relies on a sharp control of the fluctuation of a supremum, over some class of functions, of an empirical process. In the recent years, Talagrand (1996) stated a very powerful theorem on the concentration of such a supremum around its expectation. This theorem is recalled below

Theorem 5.1 (Talagrand's Theorem). *Let $U_1, \cdots, U_n$ be independent random variables with values in some measurable space $\mathcal{E}$. Let $\mathcal{G}$ be some countable class of real valued measurable functions on $\mathcal{E}$. Assume that there exists a constant $b > 0$ such that for all $g \in \mathcal{G}$, $\|g\|_\infty \leq b$. Let us set*

$$\text{either } Z = \sup_{g\in\mathcal{G}} \sum_{i=1}^n g(U_i) \text{ , or } Z = \sup_{g\in\mathcal{G}} \left| \sum_{i=1}^n g(U_i) \right| \text{ and } V^2 = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^n g^2(U_i) \right].$$

Then for each $t > 0$, we have in both cases

$$\mathbb{P}\left[|Z - \mathbb{E}[Z]| \geq t\right] \leq C_0 \exp\left(-\frac{t}{C_0 b} \ln\left(1 + \frac{tb}{V^2}\right)\right), \tag{20}$$

where C_0 denotes some universal constant.

Unfortunately, the class of functions that is relevant for proving our main result is not uniformly bounded. In order to get a suitable result about the concentration of our empirical process around its mean, we use the following moment inequality (the proof is deferred to Section 7.2):

Theorem 5.2. *Let $U_1, \cdots, U_n$ be independent random variables with values in some measurable space $\mathcal{E}$. Let $\mathcal{G}$ be some countable class of real valued measurable functions on $\mathcal{E}$. Let us set*

$$\text{either } Z = \sup_{g\in\mathcal{G}} \sum_{i=1}^n g(U_i) \text{ , or } Z = \sup_{g\in\mathcal{G}} \left| \sum_{i=1}^n g(U_i) \right|.$$

Then in both cases we have for all $p \geq 2$,

$$C(p)^{-1}\mathbb{E}\left[|Z - \mathbb{E}[Z]|^p\right] \leq \mathbb{E}\left[\max_{i=1,\cdots,n} \sup_{g\in\mathcal{G}} |g(U_i)|^p \right] + \left(\mathbb{E}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^n g^2(U_i) \right] \right)^{p/2}, \tag{21}$$

where $C(p)$ is a constant that depends on p only.

Comment: By a density argument, the previous result extends to classes of functions $\mathcal{G}$ that are not countable but for which there exists some countable subset $\mathcal{G}' \subset \mathcal{G}$ that is dense in $\mathcal{G} \subset \mathbb{L}^\infty$.

As a consequence of Theorem 5.2, we give a deviation inequality for random variables ζ^2 of the form

$$\zeta^2 = \sum_{i=1}^{n} (A\varepsilon)_i^2 = {}^t\varepsilon\ {}^tAA\varepsilon,$$

where A denotes some matrix of $\mathbb{M}_n(\mathbb{R})$. This deviation bound is used to prove Theorem 3.1. When the ε_is are i.i.d. standard Gaussian random variables and A a projector onto some linear space of dimension D then ζ^2 is known to be distributed like a khi-square with D degrees of freedom. In this case, a deviation inequality was established by Laurent and Massart (1998). An analogue is given below which holds for more general symmetric matrices and under weak moment condition on the ε_is. We recall that $\rho(A)$ denotes the spectral radius of A defined by (16).

Corollary 5.1. *Let $\tilde{A}$ denote some symmetric nonnegative matrix in $\mathbb{M}_n(\mathbb{R}) \setminus \{0\}$ and $\varepsilon = {}^t(\varepsilon_1, ..., \varepsilon_n)$ a random vector in $\mathbb{R}^n$ with i.i.d centered components. Assume that $\sigma^2 = \mathbb{E}[\varepsilon_1^2] < +\infty$ and set*

$$\zeta(\varepsilon) = \sqrt{{}^t\varepsilon \tilde{A} \varepsilon}.$$

For all $p \geq 2$ such that $\mathbb{E}[|\varepsilon_1|^p] < +\infty$ we have

$$\mathbb{P}\left[\zeta^2(\varepsilon) \geq \mathrm{tr}(\tilde{A})\sigma^2 + 2\sigma^2\sqrt{\rho(\tilde{A})\mathrm{tr}(\tilde{A})x} + \sigma^2\rho(\tilde{A})x\right] \leq C(p)\tau_p \frac{\mathrm{tr}(\tilde{A})}{\rho(\tilde{A})x^{p/2}}, \tag{22}$$

where $\tau_p = \mathbb{E}[|\varepsilon_1|^p]/\sigma^p$.

Proof. By homogeneity we assume that $\sigma^2 = 1$. Let $\mathcal{B}_n$ be the unit ball of $\mathbb{R}^n$ (with respect to the Euclidean norm denoted by $\|\ \|$) and $\{e_1, ..., e_n\}$ be the canonical basis. Since $\tilde{A}$ is nonnegative and symmetric there exists $A \in \mathbb{M}_n(\mathbb{R}) \setminus \{0\}$ such that $\tilde{A} = {}^tAA$. By Cauchy-Schwarz we have

$$\zeta^2(\varepsilon) = \|A\varepsilon\|^2 = \left[\sup_{u \in \mathcal{B}_n} \sum_{i=1}^{n} (A\varepsilon)_i u_i\right]^2 = \left[\sup_{u \in \mathcal{B}_n} \sum_{i=1}^{n} \varepsilon_i ({}^tAu)_i\right]^2.$$

By defining

$$\mathcal{G} = \left\{ g_u /\ g_u(x) = \sum_{j=1}^{n} x_j ({}^tAu)_j,\ u \in \mathcal{B}_n \right\}$$

we see that

$$\zeta(\varepsilon) = \sup_{u \in \mathcal{B}_n} \sum_{i=1}^{n} g_u(\varepsilon_i e_i),$$

the supremum being nonnegative since $0 \in \mathcal{B}_n$. By applying Theorem 5.2 with $U_i = \varepsilon_i e_i$, $\mathcal{E} = \mathbb{R}^n$ we obtain for each positive number t

$$\begin{aligned}
&\mathbb{P}\left[\zeta(\varepsilon) \geq \mathbb{E}[\zeta(\varepsilon)] + t\right] \\
&\quad\leq \; t^{-p}\mathbb{E}\left[|\zeta(\varepsilon) - \mathbb{E}[\zeta(\varepsilon)]|^p\right] \\
&\quad\leq \; C(p)t^{-p}\left(\mathbb{E}\left[\max_{i=1,\dots,n}\sup_{u\in\mathcal{B}_n}|\varepsilon_i|^p({}^tAu)_i^p\right] + \left(\mathbb{E}\left[\sup_{u\in\mathcal{B}_n}\sum_{i=1}^n \varepsilon_i^2({}^tAu)_i^2\right]\right)^{p/2}\right) \\
&\quad= \; C(p)t^{-p}\left(\mathbb{E}_1 + \mathbb{E}_2^{p/2}\right). \qquad (23)
\end{aligned}$$

We start by bounding $\mathbb{E}_1$. For all $u \in \mathcal{B}_n$ and $i \in \{1, ..., n\}$,

$$({}^tAu)_i^2 \leq \|{}^tAu\|^2 \leq \rho^2({}^tA) \times 1 = \rho^2(A),$$

therefore

$$\mathbb{E}_1 \leq \rho^{p-2}(A)\mathbb{E}\left[\sup_{u\in\mathcal{B}_n}\sum_{i=1}^n ({}^tAu)_i^2|\varepsilon_i|^p\right].$$

Writing that $u = \sum_{j=1}^n u_j e_j$ with $\sum_{j=1}^n u_j^2 = 1$ gives

$$({}^tAu)_i^2 = \left(\sum_{j=1}^n u_j({}^tAe_j)_i\right)^2 \leq \sum_{j=1}^n ({}^tAe_j)_i^2 = \sum_{j=1}^n A_{ji}^2.$$

Thus,

$$\mathbb{E}_1 \leq \rho^{p-2}(A)\mathbb{E}[|\varepsilon_1|^p]\sum_{i=1}^n\sum_{j=1}^n A_{ji}^2 = \rho^{p-2}(A)\mathbb{E}[|\varepsilon_1|^p]\mathrm{tr}(\tilde{A}).$$

We now bound $\mathbb{E}_2$ via a truncation argument. Since for all $u \in \mathcal{B}_n$, $\|{}^tAu\|^2 \leq \rho^2(A)$, for any positive number c to be specified later we have

$$\begin{aligned}
\mathbb{E}_2 \;&\leq\; \mathbb{E}\left[\sup_{u\in\mathcal{B}_n}\sum_{i=1}^n({}^tAu)_i^2\varepsilon_i^2\mathbb{1}_{|\varepsilon_i|\leq c}\right] + \mathbb{E}\left[\sup_{u\in\mathcal{B}_n}\sum_{i=1}^n({}^tAu)_i^2\varepsilon_i^2\mathbb{1}_{|\varepsilon_i|> c}\right] \\
&\leq\; c^2\rho^2(A) + c^{2-p}\mathbb{E}\left[\sup_{u\in\mathcal{B}_n}\sum_{i=1}^n({}^tAu)_i^2|\varepsilon_i|^p\right] \\
&\leq\; c^2\rho^2(A) + c^{2-p}\mathbb{E}[|\varepsilon_1|^p]\mathrm{tr}(\tilde{A})
\end{aligned}$$

using the preceding result on $\mathbb{E}_1$. It remains to take $c^p = \mathbb{E}[|\varepsilon_1|^p]\mathrm{tr}(\tilde{A})/\rho^2(A)$ to get that

$$2^{-p/2}\mathbb{E}_2^{p/2} \leq \rho^{p-2}(A)\mathbb{E}[|\varepsilon_1|^p]\mathrm{tr}(\tilde{A}).$$

Since $(\mathbb{E}[\zeta(\varepsilon)])^2 \leq \mathbb{E}[\zeta^2(\varepsilon)]$, we straightforwardly derive from (23) that

$$\mathbb{P}\left[\zeta^2(\varepsilon) \geq \mathbb{E}[\zeta^2(\varepsilon)] + 2\sqrt{\mathbb{E}[\zeta^2(\varepsilon)]t^2} + t^2\right] \leq C'(p)t^{-p}\rho^{p-2}(A)\mathbb{E}[|\varepsilon_1|^p]\mathrm{tr}(\tilde{A}), \tag{24}$$

for all $t > 0$. Moreover $\mathbb{E}[\zeta^2(\varepsilon)] = \mathbb{E}[\,{}^t\!\varepsilon\tilde{A}\varepsilon] = \mathrm{tr}(\tilde{A})$ and the result follows by choosing $t^2 = \rho(\tilde{A})x > 0$. □

6. Estimation of s when the variance is unknown

In contrast with Section 3, in this section the variance σ^2 is assumed to be some unknown quantity. Since this quantity can no longer appear in the definition of the estimator (more precisely, in the definition of the penalty term) we introduce some estimator for it, namely a residual least-squares estimator, $\hat{\sigma}_n^2$ defined as follows

Let V_n be some linear subspace of $\mathbb{R}^n$ such that $\dim(V_n) = [n/2]$, we define

$$\hat{\sigma}_n^2 = \frac{n}{n - [n/2]}\mathrm{d}_n^2\left(Y, V_n\right), \tag{25}$$

then the following result holds.

Theorem 6.1. *Let $(S_m)_{m\in\mathcal{M}_n}$, be some collection of finite dimensional linear subspaces of $\mathbb{L}^2(\mathcal{X}, \mu_n)$. For any positive number θ, let us define* pen : $\mathcal{M}_n \to \mathbb{R}_+$ *by*

$$\mathrm{pen}(m) = (1+\theta)\frac{D_m}{n}\hat{\sigma}_n^2,$$

where $\hat{\sigma}_n^2$ is defined by (25). Let $q \in]0, 1]$ be given such that $\mathbb{E}[|\varepsilon_1|^p] < +\infty$ for some $p > 2(1 + 2q)$. Then the PLSE defined by (6) satisfies for some constant C that depends on θ and q only

$$\left(\mathbb{E}\left[\|s - \tilde{s}\|_n^{2q}\right]\right)^{1/q} \leq C\left[\inf_{m\in\mathcal{M}_n}\left(\mathrm{d}_n^2(s, S_m) \;+\; \frac{D_m}{n}\sigma^2\right) + \mathrm{d}_n^2(s, V_n) + \Delta'_p\frac{\sigma^2}{n}\right], \tag{26}$$

where

$$\left(\Delta_p'\right)^q = C'(p,q,\theta)\left[\frac{\mathbb{E}[|\varepsilon_1|^p]}{\sigma^p}\left(1 + \sum_{\substack{m\in\mathcal{M}_n\\ D_m\geq 1}} D_m^{-(p/2-1-q)}\right) + \frac{\|s\|_n^2}{\sigma^2}\right].$$

7. Proofs

7.1. **Proof of Theorem 3.1.** In this section, we actually show something stronger than Theorem 3.1, namely we show that for any $\eta > 0$ and any sequence of positive numbers L_m, if the penalty function is chosen to satisfy

$$\mathrm{pen}(m) \geq (1+\eta+L_m)\frac{D_m}{n}\sigma^2 \quad \text{for all } m \in \mathcal{M}_n, \tag{27}$$

then for each $x > 0$ and $p \geq 2$

$$\mathbb{P}\left[\mathcal{H}(s) \geq \left(1+2\eta^{-1}\right)\frac{x}{n}\sigma^2\right] \leq C(p,\eta)\tau_p \sum_{m\in\mathcal{M}_n} \frac{D_m \vee 1}{(L_m D_m + x)^{p/2}}, \tag{28}$$

where

$$\mathcal{H}(s) = \left(\|s-\tilde{s}\|_n^2 - \left(2+4\eta^{-1}\right)\inf_{m\in\mathcal{M}_n}\left(\mathrm{d}_n^2(s,S_m)+\mathrm{pen}(m)\right)\right)_+$$

and $\tau_p = \mathbb{E}[|\varepsilon_1|^p]/\sigma^p$. To obtain (9), take $\eta = \theta/2 = L_m$. As for each $m \in \mathcal{M}_n$,

$$\mathrm{d}_n^2(s,S_m) + \mathrm{pen}(m) \leq (1+\theta)\left(\mathrm{d}_n^2(s,S_m)+\frac{D_m}{n}\right) = (1+\theta)\mathbb{E}\left[\|s-\hat{s}_m\|_n^2\right],$$

we get that for all $q > 0$,

$$\begin{aligned} \mathcal{H}^q(s) &\geq \left(\|s-\tilde{s}\|_n^2 - (2+8\theta^{-1})(1+\theta)M_n^*\right)_+^q \\ &= \left(\|s-\tilde{s}\|_n^2 - \kappa(\theta)M_n^*\right)_+^q. \end{aligned} \tag{29}$$

Since

$$\mathbb{E}[\mathcal{H}^q(s)] = \int_0^{+\infty} qu^{q-1}\mathbb{P}[\mathcal{H}(s) > u]du,$$

we derive from (28) and (29) that for all $p > 2(1+q)$

$$\begin{aligned} &\mathbb{E}\left[\left(\|s-\tilde{s}\|_n^2 - \kappa(\theta)M_n^*\right)_+^q\right] \\ &\leq \mathbb{E}[\mathcal{H}^q(s)] \\ &\leq C(p,\theta)(1+\frac{4}{\theta})^q\tau_p\frac{\sigma^{2q}}{n^q}\sum_{m\in\mathcal{M}_n}\int_0^{+\infty} qx^{q-1}\left[\frac{D_m\vee 1}{(\theta D_m/2+x)^{p/2}}\wedge 1\right]dx \\ &\leq C'(p,q,\theta)\tau_p\frac{\sigma^{2q}}{n^q}\left(1+\sum_{\substack{m\in\mathcal{M}_n\\ D_m\geq 1}} D_m^{-(p/2-1-q)}\right). \end{aligned} \tag{30}$$

Indeed, if there exists some $m \in \mathcal{M}_n$ such that $S_m = \{0\}$, then for such an m,

$$\begin{aligned}\int_0^{+\infty} qx^{q-1}\left[\frac{D_m \vee 1}{(\theta D_m/2+x)^{p/2}} \wedge 1\right] dx &= \int_0^{+\infty} qx^{q-1}\left[\frac{1}{x^{p/2}} \wedge 1\right] dx \\ &= \int_0^1 qx^{q-1}dx \ + \ \int_1^{+\infty} qx^{q-1-p/2}dx \\ &= 1 + \frac{q}{p/2-q}.\end{aligned}$$

On the other hand, for $m \in \mathcal{M}_n$ such that $D_m \geq 1$,

$$\begin{aligned}&\int_0^{+\infty} qx^{q-1}\frac{D_m}{(\theta D_m/2+x)^{p/2}}dx \\ &\qquad \leq \ 2^{p/2}\theta^{-p/2}D_m^{1-p/2}\int_0^{D_m} qx^{q-1}dx \ + \ D_m\int_{D_m}^{+\infty} qx^{q-1-p/2}dx \\ &\qquad \leq \ D_m^{-(p/2-1-q)}\left(2^{p/2}\theta^{-p/2} + \frac{q}{p/2-q}\right).\end{aligned}$$

This proves (30) which leads to (9).

We now turn to the proof of (28). For the sake of simplicity, we identify the function f with the n-dimensional vector ${}^t(f(x_1),...,f(x_n))$ and we denote by $<,>_n$ the inner product of $\mathbb{R}^n$ associated to the norm $\| \ \|_n$. For each $m \in \mathcal{M}_n$ we denote by Π_m the orthogonal projector onto the linear space $\{\ {}^t(f(x_1),...,f(x_n))/\ f \in S_m\} \subset \mathbb{R}^n$. This linear space is also denoted by S_m. From now on, the subscript m denotes any minimizer of the function $m' \mapsto \|s-\Pi_{m'}s\|_n^2 + \mathrm{pen}(m')$, $m' \in \mathcal{M}_n$. Using the definition of γ_n we have that for all $f \in \mathbb{R}^n$, $\|s-f\|_n^2 = \gamma_n(f) + 2 < f-Y, \varepsilon >_n + \|\varepsilon\|_n^2$. We derive that

$$\|s-\tilde{s}\|_n^2 - \|s-\Pi_m s\|_n^2 = \gamma_n(\tilde{s}) - \gamma_n(\Pi_m s) + 2 < \tilde{s} - \Pi_m s, \varepsilon >_n . \tag{31}$$

By (8) we know that $\gamma_n(\tilde{s}) - \gamma_n(\Pi_m s) \leq \mathrm{pen}(m) - \mathrm{pen}(\hat{m})$ so we get from (31) that

$$\begin{aligned}\|s-\tilde{s}\|_n^2 \ &\leq \ \|s-\Pi_m s\|_n^2 + 2 < s - \Pi_m s, \varepsilon >_n + \ 2 < \Pi_{\hat{m}} s - s, \varepsilon >_n \\ &\quad + \ 2 < \tilde{s} - \Pi_{\hat{m}} s, \varepsilon >_n \ + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}).\end{aligned} \tag{32}$$

In the following we set for each $m' \in \mathcal{M}_n$,

$$\mathcal{B}_{m'} = \{t \in S_{m'}/\ \|t\|_n \leq 1\},\ G_{m'} = \sup_{t \in \mathcal{B}_{m'}} < t, \varepsilon >_n = \|\Pi_{m'}\varepsilon\|_n$$

and

$$u_{m'} = \begin{cases} (\Pi_{m'}s - s)/\|\Pi_{m'}s - s\|_n & \text{if } \Pi_{m'}s \neq s \\ 0 & \text{otherwise.}\end{cases}$$

Since $\tilde{s} = \Pi_{\hat{m}} s + \Pi_{\hat{m}} \varepsilon$, (32) gives

$$\begin{aligned} \|s-\tilde{s}\|_n^2 &\leq \|s-\Pi_m s\|_n^2 + 2\|s-\Pi_m s\|_n| < u_m, \varepsilon >_n | \\ &\quad + 2\|s-\Pi_{\hat{m}} s\|_n| < u_{\hat{m}}, \varepsilon >_n | + \ 2G_{\hat{m}}^2 \ + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}). \end{aligned} \tag{33}$$

Using repeatedly the following elementary inequality that holds for all positive numbers α, y, z

$$2yz \leq \alpha y^2 + \alpha^{-1} z^2, \tag{34}$$

we get for any $m' \in \{m, \hat{m}\}$

$$2\|s-\Pi_{m'} s\|_n| < u_{m'}, \varepsilon >_n | \leq \alpha \|s-\Pi_{m'} s\|_n^2 + \alpha^{-1} < u_{m'}, \varepsilon >_n^2 .$$

On the other hand, by Pythagoras Theorem we have

$$\begin{aligned} \|s-\tilde{s}\|_n^2 &= \|s-\Pi_{\hat{m}} s\|_n^2 + \|\Pi_{\hat{m}} s - \tilde{s}\|_n^2 \\ &= \|s-\Pi_{\hat{m}} s\|_n^2 + G_{\hat{m}}^2. \end{aligned} \tag{35}$$

We derive from (33) that

$$\begin{aligned} (1-\alpha)\|s-\tilde{s}\|_n^2 &\leq (1+\alpha)\|s-\Pi_m s\|_n^2 + \mathrm{pen}(m) \ + \ (2-\alpha) G_{\hat{m}}^2 \\ &\quad + \alpha^{-1} < u_{\hat{m}}, \varepsilon >_n^2 \ - \mathrm{pen}(\hat{m}) \ + \ \alpha^{-1} < u_m, \varepsilon >_n^2 . \end{aligned} \tag{36}$$

We choose $\alpha = 2/(2+\eta) \in]0,1[$ but for legibility we keep using the notation α. Let $\bar{p}_1$ and $\bar{p}_2$ be two functions depending on η mapping $\mathcal{M}_n$ into $\mathbb{R}_+$. They will be specified later to satisfy

$$\mathrm{pen}(m') \geq (2-\alpha)\bar{p}_1(m') \ + \ \alpha^{-1}\bar{p}_2(m') \quad \text{for all } m' \in \mathcal{M}_n. \tag{37}$$

Since $\alpha^{-1}\bar{p}_2(m) \leq \mathrm{pen}(m)$ and $1+\alpha \leq 2$, we get from (36) and (37)

$$\begin{aligned} (1-\alpha)\|s-\tilde{s}\|_n^2 &\leq (1+\alpha)\|s-\Pi_m s\|_n^2 + \mathrm{pen}(m) \ + \alpha^{-1}\bar{p}_2(m) \\ &\quad + (2-\alpha)\left(G_{\hat{m}}^2 - \bar{p}_1(\hat{m})\right) \ + \ \alpha^{-1}\left(< u_{\hat{m}}, \varepsilon >_n^2 \ - \bar{p}_2(\hat{m})\right) \\ &\quad + \alpha^{-1}\left(< u_m, \varepsilon >_n^2 - \bar{p}_2(m)\right) \\ &\leq 2\left(\|s-\Pi_m s\|_n^2 + \mathrm{pen}(m)\right) \ + \ (2-\alpha)\left(G_{\hat{m}}^2 - \bar{p}_1(\hat{m})\right) \\ &\quad + \alpha^{-1}\left(< u_{\hat{m}}, \varepsilon >_n^2 \ - \bar{p}_2(\hat{m})\right) \ + \ \alpha^{-1}\left(< u_m, \varepsilon >_n^2 - \bar{p}_2(m)\right). \end{aligned}$$

As $2/(1-\alpha) = (2+4/\eta)$, we obtain that

$$\begin{aligned} (1-\alpha)\mathcal{H}(s) &\leq (2-\alpha)\left(G_{\hat{m}}^2 - \bar{p}_1(\hat{m})\right) \\ &\quad + \alpha^{-1}\left(< u_{\hat{m}}, \varepsilon >_n^2 - \bar{p}_2(\hat{m})\right) \ + \ \alpha^{-1}\left(< u_m, \varepsilon >_n^2 - \bar{p}_2(m)\right). \end{aligned}$$

For any $x > 0$,

$$\begin{aligned}
\mathbb{P}\left[(1-\alpha)\mathcal{H}(s) \geq \frac{x\sigma^2}{n}\right] &\leq \mathbb{P}\left[\exists m' \in \mathcal{M}_n,\ (2-\alpha)\left(G_{m'}^2 - \bar{p}_1(m')\right) \geq \frac{x\sigma^2}{3n}\right] \\
&\quad + \mathbb{P}\left[\exists m' \in \mathcal{M}_n,\ \alpha^{-1}\left(< u_{m'}, \varepsilon >_n^2 - \bar{p}_2(m')\right) \geq \frac{x\sigma^2}{3n}\right] \\
&\leq \sum_{m' \in \mathcal{M}_n} \mathbb{P}\left[(2-\alpha)\left(\|\Pi_{m'}\varepsilon\|_n^2 - \bar{p}_1(m')\right) \geq \frac{x\sigma^2}{3n}\right] \\
&\quad + \sum_{m' \in \mathcal{M}_n} \mathbb{P}\left[\alpha^{-1}\left(< u_{m'}, \varepsilon >_n^2 - \bar{p}_2(m')\right) \geq \frac{x\sigma^2}{3n}\right] \\
&= \sum_{m' \in \mathcal{M}_n} \mathbb{P}_{1,m'}(x) + \sum_{m' \in \mathcal{M}_n} \mathbb{P}_{2,m'}(x). \qquad (38)
\end{aligned}$$

We first bound $\mathbb{P}_{2,m'}(x)$. Let t be some positive number,

$$\mathbb{P}\left[|< u_{m'}, \varepsilon >_n| \geq t\right] \leq t^{-p}\mathbb{E}\left[|< u_{m'}, \varepsilon >_n|^p\right]. \qquad (39)$$

By Rosenthal's inequality (recalled in the Appendix), we know that for some constant $C(p)$ that depends on p only

$$C^{-1}(p)n^p\mathbb{E}\left[|< u_{m'}, \varepsilon >_n|^p\right] \leq \mathbb{E}[|\varepsilon_1|^p]\sum_{i=1}^n |u_{m',i}|^p + \left(\sigma^2\sum_{i=1}^n u_{m',i}^2\right)^{p/2}.$$

Since $p \geq 2$, $\sigma^p \leq \mathbb{E}[|\varepsilon_i|^p]$ and $\sum_{i=1}^n |u_{m',i}|^p \leq (\sum_{i=1}^n u_{m',i}^2)^{p/2} = n^{p/2}\|u_{m'}\|_n^p \leq n^{p/2}$. Thus, we deduce from (39) that for some constant $C'(p)$ that only depends on p

$$\mathbb{P}\left[|< u_{m'}, \varepsilon >_n| \geq t\right] \leq C'(p)\mathbb{E}[|\varepsilon_1|^p]n^{-p/2}t^{-p}.$$

Let δ be some positive number depending on η only to be chosen later. We take t such that $nt^2 = \min(\delta, \alpha/3)(L_{m'}D_{m'} + x)\sigma^2$ and set $n\bar{p}_2(m') = \delta L_{m'}D_{m'}\sigma^2$. We get

$$\begin{aligned}
\mathbb{P}_{2,m'}(x) &= \mathbb{P}\left[n < u_{m'}, \varepsilon >_n^2 \geq \delta L_{m'}D_{m'}\sigma^2 + \frac{\alpha x\sigma^2}{3}\right] \\
&\leq \mathbb{P}\left[n < u_{m'}, \varepsilon >_n^2 \geq \min(\delta, \alpha/3)\left(L_{m'}D_{m'} + x\right)\sigma^2\right] \\
&\leq C''(p,\eta)\frac{\tau_p}{(L_{m'}D_{m'} + x)^{p/2}}, \qquad (40)
\end{aligned}$$

where we recall that $\tau_p = \mathbb{E}[|\varepsilon_1|^p]/\sigma^p$.

We now bound $\mathbb{P}_{1,m'}(x)$. If $D_{m'} = 0$, we take $\bar{p}_1(m') = 0$ and clearly $\mathbb{P}_{1,m'}(x) = 0$, thus it remains to bound $\mathbb{P}_{1,m'}(x)$ for those $m' \in \mathcal{M}_n$ such that $D_{m'} \geq 1$. By using

Corollary 5.1 with $A = \Pi_{m'} = \tilde{A}$ which satisfies $\mathrm{tr}(\tilde{A}) = D_{m'}$ and $\rho(\tilde{A}) = 1$, we obtain from (22) that for any positive $x_{m'}$

$$\mathbb{P}\left[nG_{m'}^2 \geq D_{m'}\sigma^2 + 2\sigma^2\sqrt{D_{m'}x_{m'}} + \sigma^2 x_{m'}\right] \leq C(p)\tau_p D_{m'} x_{m'}^{-p/2}.$$

Since for any $\beta > 0$,

$$2\sigma^2\sqrt{D_{m'}x_{m'}} \leq \beta D_{m'}\sigma^2 \; + \; \beta^{-1}\sigma^2 x_{m'}$$

we obtain that

$$\mathbb{P}\left[nG_{m'}^2 \geq (1+\beta)D_{m'}\sigma^2 + (1+\beta^{-1})x_{m'}\sigma^2\right] \leq C(p)\tau_p D_{m'} x_{m'}^{-p/2}. \tag{41}$$

Now, for some number β depending on η only to be chosen later, we take

$$x_{m'} = (1+\beta^{-1})^{-1}\min(\delta, (2-\alpha)^{-1}/3)(L_{m'}D_{m'} + x)$$

and $n\bar{p}_1(m') = \delta L_{m'}D_{m'}\sigma^2 + (1+\beta)D_{m'}\sigma^2$. This gives

$$\begin{aligned} \mathbb{P}_{1,m'}(x) &= \mathbb{P}\left[nG_{m'}^2 \geq (1+\beta)D_{m'}\sigma^2 + \delta L_{m'}D_{m'}\sigma^2 + \frac{(2-\alpha)^{-1}}{3}x\sigma^2\right] \\ &\leq \mathbb{P}\left[nG_{m'}^2 \geq (1+\beta)D_{m'}\sigma^2 + (1+\beta^{-1})x_{m'}\sigma^2\right] \\ &\leq C'(p,\eta)\frac{D_{m'}\tau_p}{(L_{m'}D_{m'} + x)^{p/2}}. \end{aligned} \tag{42}$$

Gathering (38), (40) and (42), we get (28). It remains to choose β and δ for (37) to hold (we recall that $\alpha = 2/(2+\eta)$). This is the case if $(2-\alpha)(1+\beta) = (1+\eta)$ and $(2-\alpha+\alpha^{-1})\delta = 1$, therefore we take $\beta = \eta/2$ and $\delta = (1 + \eta/2 + 2(1+\eta)/(2+\eta))^{-1}$.

7.2. **Proof of Theorem 5.2.** The line of proof for Inequality (21) is similar to that given by Petrov (1995) (p.59) for the Rosenthal Inequality. We start by setting for each $x > 0$, $Z_x = \sup_{g\in\mathcal{G}}\sum_{i=1}^n g(U_i)\mathbb{1}_{|g(U_i)|\leq x}$. For any $t > 0$, we have

$$\begin{aligned} \mathbb{P}\left[|Z - \mathbb{E}[Z]| > t\right] &\leq \mathbb{P}\left[Z_x \neq Z\right] + \mathbb{P}\left[|Z - \mathbb{E}[Z]| > t, Z = Z_x\right] \\ &\leq \mathbb{P}\left[\max_{i=1,\cdots,n}\sup_{g\in\mathcal{G}}|g(U_i)| > x\right] + \mathbb{P}\left[|Z - \mathbb{E}[Z]| > t, Z = Z_x\right]. \end{aligned} \tag{43}$$

Let r be some nonnegative number that will be chosen later. As

$$\mathbb{E}\left[|Z - \mathbb{E}[Z]|^p\right] = \int_0^\infty pt^{p-1}\mathbb{P}\left[|Z - \mathbb{E}[Z]| > t\right]dt,$$

using (43), we get by taking $r = t/x$ that

$$\begin{aligned}
\mathbb{E}\left[|Z-\mathbb{E}[Z]|^p\right] &\leq \int_0^\infty pt^{p-1}\mathbb{P}\left[r \max_{i=1,\cdots,n}\sup_{g\in\mathcal{G}}|g(U_i)| > t\right] dt \\
&\quad + \int_0^\infty pt^{p-1}\mathbb{P}\left[|Z-\mathbb{E}[Z]| > t, Z = Z_x\right] dt \\
&= r^p\mathbb{E}\left[\max_{i=1,\cdots,n}\sup_{g\in\mathcal{G}}|g(U_i)|^p\right] \\
&\quad + \int_0^\infty pt^{p-1}\mathbb{P}\left[|Z-\mathbb{E}[Z]| > t, Z = Z_x\right] dt. \qquad (44)
\end{aligned}$$

But,

$$\mathbb{P}\left[|Z-\mathbb{E}[Z]| > t, Z = Z_x\right] \leq \mathbb{P}\left[|Z_x-\mathbb{E}[Z_x]| > t - |\mathbb{E}[Z-Z_x]|\right] \qquad (45)$$

and

$$\begin{aligned}
|\mathbb{E}[Z-Z_x]| &\leq \mathbb{E}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^n \left|g(U_i) - g(U_i)\mathbb{1}_{|g(U_i)|\leq x}\right|\right] \\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^n |g(U_i)|\mathbb{1}_{|g(U_i)|>x}\right] \leq \frac{1}{x}\mathbb{E}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^n g^2(U_i)\right] \\
&= \frac{V^2}{x}.
\end{aligned}$$

Hence, if $t \geq \sqrt{2r}V = t_0$ then $t - |\mathbb{E}[Z-Z_x]| \geq t/2$ and by Talagrand's Inequality (20) with $b = x$, we get

$$\begin{aligned}
&\int_0^\infty pt^{p-1}\mathbb{P}\left[|Z-\mathbb{E}[Z]| > t, Z = Z_x\right] dt \\
&\qquad \leq \int_0^{t_0} pt^{p-1}dt + \int_{t_0}^\infty pt^{p-1}\mathbb{P}\left[|Z_x-\mathbb{E}[Z_x]| > \frac{t}{2}\right] dt \\
&\qquad \leq t_0^p + C_0\int_0^\infty pt^{p-1}\exp\left(-\frac{t}{2C_0x}\ln\left(1+\frac{tx}{2V^2}\right)\right) dt \\
&\qquad \leq t_0^p + C_0\int_0^\infty pt^{p-1}\exp\left(-\frac{r}{2C_0}\ln\left(1+\frac{t^2}{2rV^2}\right)\right) dt.
\end{aligned}$$

We set $u = t/\sqrt{2r}V$, and as $t^{p-1} \leq (2rV^2)^{(p-1)/2}(1+u^2)^{(p-1)/2}$ we obtain (replacing t_0 by $\sqrt{2r}V$)

$$\begin{aligned}
&\int_0^\infty pt^{p-1}\mathbb{P}\left[|Z-\mathbb{E}[Z]| > t, Z = Z_x\right] dt \\
&\qquad \leq (2r)^{p/2}V^p + C_0p(2r)^{p/2}V^p\int_0^\infty (1+u^2)^{(p-1)/2-r/(2C_0)}du. \qquad (46)
\end{aligned}$$

Now, it remains to take $r = C_0(p+1)$ to guarantee the convergence of the last integral in the right-hand side of (46). Finally (21) follows by collecting (44) and (46).

7.3. **Proof of Theorem 6.1.** We start with the following claim

Claim : $\hat{\sigma}_n^2$ satisfies the following properties:

(i) $\mathbb{E}[\hat{\sigma}_n^2] \leq \sigma^2 + 2\mathrm{d}_n^2(s, V_n)$.

(ii) Let $0 < \delta < 1/2$ then

$$\mathbb{P}\left[\hat{\sigma}_n^2 \leq (1-2\delta)\sigma^2\right] \leq C(p,\delta)\frac{\tau_p}{n^\beta},$$

where $\beta = (p/2-1) \wedge p/4$.

Let us first assume that our claim is true. Given $\theta > 0$ one can find two positive numbers $\delta = \delta(\theta) \leq 1/2$ and $\eta = \eta(\theta)$ such that $(1+\theta)(1-2\delta) \geq (1+2\eta)$. For such a δ, let

$$\Omega_n = \left\{\hat{\sigma}_n^2 \geq (1-2\delta)\sigma^2\right\}.$$

We start by bounding $\mathbb{E}[\|s - \tilde{s}\|_n^{2q}\mathbf{1}_{\Omega_n}]$. On Ω_n we know that

$$\mathrm{pen}(m) \geq (1+2\eta)\frac{D_m}{n}\sigma^2 \text{ for all } m \in \mathcal{M}_n.$$

Arguing as in the proof of Theorem 3.1 (with $L_m = \eta$), we get under the assumptions of Theorem 6.1 that for some m chosen in $\mathcal{M}_n$ to minimize $\mathrm{d}_n^2(s, S_{m'}) + \sigma^2 D_{m'}/n$ among $m' \in \mathcal{M}_n$,

$$(\mathbb{E}[\mathcal{H}_m^q(s)\mathbf{1}_{\Omega_n}])^{1/q} \leq \Delta_p\frac{\sigma^2}{n} \tag{47}$$

where $\mathcal{H}_m$ is defined by

$$\mathcal{H}_m(s) = \left(\|s-\tilde{s}\|_n^2 - \kappa(\theta)\left(\mathrm{d}_n^2(s, S_m) + \frac{D_m}{n}\hat{\sigma}_n^2\right)\right)_+.$$

Since $q \leq 1$, by a convexity argument and Jensen's inequality we deduce from (47) that

$$\begin{aligned}\left(\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\mathbf{1}_{\Omega_n}\right]\right)^{1/q} &\leq C'(\theta,q)\left(\mathbb{E}\left[\left(\mathrm{d}_n^2(s,S_m) + \frac{D_m}{n}\hat{\sigma}_n^2\right)^q\right]\right)^{1/q} + \Delta_p\frac{\sigma^2}{n}\\ &\leq C'(\theta,q)\left[\mathrm{d}_n^2(s,S_m) + \frac{D_m}{n}\mathbb{E}[\hat{\sigma}_n^2]\right] + \Delta_p\frac{\sigma^2}{n}.\end{aligned}$$

As $D_m \leq n$, we obtain by using (i)

$$\left(\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\mathbf{1}_{\Omega_n}\right]\right)^{1/q} \leq C''(\theta,q)\left[\mathrm{d}_n^2(s,S_m)+\frac{D_m}{n}\sigma^2+2\mathrm{d}_n^2(s,V_n)+\Delta_p\frac{\sigma^2}{n}\right]. \tag{48}$$

We now bound $\mathbb{E}[\|s-\tilde{s}\|_n^{2q}\mathbf{1}_{^c\Omega_n}]$. From (35) we know that $\|s-\tilde{s}\|_n^2 \leq \|s\|_n^2+\|\varepsilon\|_n^2$, thus,

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\mathbf{1}_{^c\Omega_n}\right] \leq \|s\|_n^{2q}\mathbb{P}\left[{}^c\Omega_n\right]+\mathbb{E}\left[\|\varepsilon\|_n^{2q}\mathbf{1}_{^c\Omega_n}\right].$$

Hölder's inequality with $k=p/(2q)>1$ gives

$$\mathbb{E}\left[\|\varepsilon\|_n^{2q}\mathbf{1}_{^c\Omega_n}\right] \leq \left(\mathbb{E}[\|\varepsilon\|_n^p]\right)^{1/k}\mathbb{P}^{1-1/k}\left[{}^c\Omega_n\right],$$

since $\mathbb{E}[\|\varepsilon\|_n^p] \leq \mathbb{E}[|\varepsilon_1|^p]$ for $p \geq 2$ we obtain by using (ii) that

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\mathbf{1}_{^c\Omega_n}\right] \leq \left(\|s\|_n^{2q}+\left(\mathbb{E}[|\varepsilon_1|^p]\right)^{2q/p}\right)n^{-\beta(1-2q/p)}. \tag{49}$$

Assume that

$$\beta(1-2q/p) \geq q. \tag{50}$$

We deduce from (49) and (50) that

$$\begin{aligned}\left(\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\mathbf{1}_{^c\Omega_n}\right]\right)^{1/q} &\leq \left(\|s\|_n^{2q}+\left(\mathbb{E}[|\varepsilon_1|^p]\right)^{2q/p}\right)^{1/q}\frac{1}{n}\\ &\leq 2^{1/q-1}\left(\tau_p+\frac{\|s\|_n^2}{\sigma^2}\right)\frac{\sigma^2}{n}\end{aligned} \tag{51}$$

and (26) follows by collecting (48) and (51). Therefore it remains to check (50). When $2 \leq p \leq 4$ then $\beta = p/4$ and thus (50) holds if and only if $p \geq 6q$. Since $q \leq 1$, this is true because $p \geq 2(1+2q) \geq 6q$. When $p \geq 4$ and $p \geq 2(1+2q)$ then

$$\beta(1-2q/p) \geq (p/2-1)(1-q/2) \geq 2q((1-q/2) \geq q$$

and therefore (50) is satisfied.

Let us now prove our claim. First we endow $\mathbb{R}^n$ with the inner product $<,>_n$ defined at the beginning of Section 7.1. In the sequel we shall denote the same way s and the vector ${}^t(s(x_1),...,s(x_n))$. Let Π_n be the orthogonal projector onto $V_n \subset \mathbb{R}^n$. We have the following decomposition

$$\begin{aligned}\hat{\sigma}_n^2 &= \frac{n}{n-[n/2]}\|Y-\Pi_nY\|_n^2\\ &= \frac{n}{n-[n/2]}\left(\|s-\Pi_ns\|_n^2+\|\varepsilon-\Pi_n\varepsilon\|_n^2+2<s-\Pi_ns,\varepsilon>_n\right).\end{aligned} \tag{52}$$

Using that $n/(n-[n/2]) \leq 2$, we obtain (i) by taking the expectation on both side of (52). Let $a_n \in V_n^{\perp}$ such that $\|a\|_n^2 = 1$, we set

$$u_n = \begin{cases} (s-\Pi_n s)/\|s-\Pi_n s\|_n & \text{if } \Pi_n s \neq s, \\ a_n & \text{otherwise.} \end{cases}$$

We have

$$\begin{aligned} 2\left| < s-\Pi_n s, \varepsilon >_n \right| &= 2\|s-\Pi_n s\|_n \left| < u_n, \varepsilon >_n \right| \\ &\leq \|s-\Pi_n s\|_n^2 + < u_n, \varepsilon >_n^2, \end{aligned}$$

thus we derive from (52)

$$\begin{aligned} \hat{\sigma}_n^2 &\geq \|\varepsilon - \Pi_n \varepsilon\|_n^2 - < u_n, \varepsilon >_n^2 \\ &= \|\varepsilon\|_n^2 - \left(\|\Pi_n \varepsilon\|_n^2 + < u_n, \varepsilon >_n^2\right) \\ &= \|\varepsilon\|_n^2 - \|\tilde{\Pi}_n s\|_n^2, \end{aligned} \tag{53}$$

where $\tilde{\Pi}_n$ denotes the orthogonal projector onto $V_n \oplus \mathbb{R} u_n$. As a consequence of (53),

$$\begin{aligned} \mathbb{P}\left[\hat{\sigma}_n^2 \leq (1-2\delta)\sigma^2\right] &\leq \mathbb{P}\left[\|\varepsilon\|_n^2 - \sigma^2 \leq -\delta\left(1-\frac{[n/2]}{n}\right)\sigma^2\right] \\ &\quad + \mathbb{P}\left[\|\tilde{\Pi}_n s\|_n^2 - \frac{[n/2]}{n}\sigma^2 \geq \delta\left(1-\frac{[n/2]}{n}\right)\sigma^2\right] \\ &= \mathbb{P}_1 + \mathbb{P}_2. \end{aligned} \tag{54}$$

Let us bound $\mathbb{P}_1$. By Markov's Inequality, we get

$$\mathbb{P}_1 \leq \mathbb{P}\left[\left|\sum_{i=1}^n \varepsilon_i^2 - n\sigma^2\right| \geq n\delta\sigma^2/2\right] \leq C'(p)\delta^{-p/2}\sigma^{-p}n^{-p/2}\mathbb{E}\left[\left|\sum_{i=1}^n \varepsilon_i^2 - n\sigma^2\right|^{p/2}\right].$$

If $p \geq 4$ then we can use the Rosenthal Inequality (59) to obtain

$$\mathbb{E}\left[\left|\sum_{i=1}^n \varepsilon_i^2 - n\sigma^2\right|^{p/2}\right] \leq C'(p)\mathbb{E}[|\varepsilon_1|^p]n^{p/4}.$$

If $2 \leq p \leq 4$, we use Inequality (60) and we get

$$\mathbb{E}\left[\left|\sum_{i=1}^n \varepsilon_i^2 - n\sigma^2\right|^{p/2}\right] \leq C'(p)\mathbb{E}[|\varepsilon_1|^p]n.$$

Therefore

$$\mathbb{P}_1 \leq C''(p)\delta^{-p/2}\tau_p n^{-\beta}. \tag{55}$$

We now bound $\mathbb{P}_2$. We use Inequality (22) of Corollary 5.1 with $A = \tilde{A} = \tilde{\Pi}_n$, $\mathrm{tr}(\tilde{\Pi}_n) = [n/2] \leq n/2$, $\rho(\tilde{\Pi}_n) = 1$ and $x = \delta^2 n/18$. Keeping in mind that $\delta < 1$, we check that $2\sqrt{[n/2]x} + x \leq \delta(1 - [n/2]/n)n$. Thus,

$$\mathbb{P}_2 \leq C'(p)\delta^{-p/2}\tau_p n^{1-p/2} \leq C'(p)\delta^{-p/2}\tau_p n^{-\beta}. \tag{56}$$

Putting (54), (55) and (56) gives (ii). This concludes the proof of the claim.

7.4. **Proof of Proposition 4.1.** We first claim that:

For each collection of models (a) or (b) the following holds:

(i) there exist $q > 0$ and a constant $\Sigma = \Sigma(p, q)$ that does not depend on n such that

$$\sum_{\substack{m \in \mathcal{M}_n \\ D_m \geq 1}} D_m^{p/2-1-q} \leq \Sigma.$$

(ii) There exists a universal constant R, such that for any n and any $t \in \mathcal{S}_n$, $\|t\|_n \leq R\|t\|_\nu$.

(iii) For each $l \geq 2$, $\alpha > 1/l$ and $s \in \mathcal{B}_{\alpha,l,\infty}$

$$\mathrm{d}_\nu(s, S_m) \leq C(\alpha)|s|_{\alpha,2} D_m^{-\alpha} \quad \text{and} \quad \mathrm{d}_\infty(s, S_m) \leq C'(\alpha)|s|_{\alpha,l} D_m^{-\alpha+1/l}.$$

Let us assume the claim is true. Since $\dim(\mathcal{S}_n)$ is of order n, for all $s \in \mathcal{B}_{\alpha,l,\infty}$ such that $|s|_{\alpha,l} \leq L$, we know from (iii) that $\mathrm{d}_\infty(s, \mathcal{S}_n) = \mathcal{O}(n^{-\alpha+1/l})$ and as $|s|_{\alpha,2} \leq |s|_{\alpha,l} \leq L$, for all $m \in \mathcal{M}_n$, $\mathrm{d}_\nu(s, S_m) \leq C(\alpha)LD_m^{-\alpha}$. Since (ii) holds let us apply Corollary 3.2. By choosing D_m of order $n^{1/(2\alpha+1)}$ we get that $\inf_{m \in \mathcal{M}_n}(\mathrm{d}_\nu^2(s, S_m) + \sigma^2 D_m/n)^{1/2} = \mathcal{O}(n^{-\alpha/(2\alpha+1)})$. Thanks to (i) (13) leads to

$$\left(\sup_{|s|_{\alpha,l} \leq L} \mathbb{E}\left[\|s - \tilde{s}\|_\nu^{2q}\right]\right)^{1/2q} = \mathcal{O}(n^{-\alpha/(2\alpha+1)}) + \mathcal{O}(n^{-\alpha+1/l}) = \mathcal{O}(n^{-\alpha/(2\alpha+1)}),$$

if $\alpha \geq (1 + \sqrt{1+2l})/2l$ i.e $l \geq (2\alpha+1)/(2\alpha^2)$. This proves (19).

Let us now prove the claim. For (iii) we refer to DeVore and Lorentz (1993) (p.354 for the collection (a), p.205 for (b) and p.181 to obtain the approximation in $\| \ \|_\infty$ via the comparison of moduli of smoothness).

For (i), we only treat the case of the collection (a), the arguments being similar for the collection (b). In this case, $\dim(S_m)= ([\alpha]+1)2^m$, thus

$$
\begin{aligned}
\sum_{\substack{m\in\mathcal{M}_n\\ D_m\geq 1}} D_m^{p/2-1-q} &\leq C([\alpha],p,q)\sum_{m=0}^{m_n} 2^{-m(p/2-1-q)}\\
&\leq C([\alpha],p,q)\sum_{m=0}^{+\infty} 2^{-m(p/2-1-q)} = \Sigma < +\infty,
\end{aligned}
$$

for any $q < p/2-1$.

It remains to show (ii). Let us first consider the case of the collection (b). Then $\mathbb{1}$, $\sqrt{2}\cos(2\pi kx)$, $\sqrt{2}\sin(2\pi kx)$ $k=1,...,2^{m_n-2}$ is an orthonormal basis of $\mathcal{S}_n$ with respect to the inner product of $\mathbb{L}^2([0,1],dx)$. It follows from easy computations that it is also orthonormal with respect to the inner product defined by $< t,t' >_n = n^{-1}\sum_{i=1}^n t(i/n)t'(i/n)$. By virtue of Lemma 3.1 R_n can be chosen equal to 1 in (12). Note that r_n in (14) can also be chosen equal to 1 and by applying (15), (19) holds with $\|s-\tilde{s}\|_n$ replaced by $\|s-\tilde{s}\|_\nu$. We now consider the case of the collection (a). Let $(\varphi^{(l)})_{l=0,...,[\alpha]}$ be an $\mathbb{L}^2$-orthonormal basis of S_0 (with no loss of generality we assume that the $\varphi^{(l)}$s vanish out side the interval $[0,1]$). Clearly the family

$$
\varphi_{k,l}(x) = 2^{m_n/2}\varphi^{(l)}(2^{m_n}x-k) \quad k=0,...,2^{m_n}-1, \quad l=0,...,[\alpha]
$$

is a $\mathbb{L}^2$-orthonormal basis of $\mathcal{S}_n$. Note that if $k\neq k'$ then for all l,l', $\varphi_{k,l}$ and $\varphi_{k',l'}$ have disjoint supports. For each $t\in\mathcal{S}_n$, let us decompose t onto this basis, we can write that $t=\sum_{k=0}^{2^{m_n}-1}\sum_{l=0}^{[\alpha]} a_{k,l}\varphi_{k,l}$, we get

$$
\begin{aligned}
\|t\|_n^2 &= \frac{1}{n}\sum_{i=1}^n\left(\sum_{k=0}^{2^{m_n}-1}\sum_{l=0}^{[\alpha]} a_{k,l}\varphi_{k,l}(\frac{i}{n})\right)^2\\
&= \frac{1}{n}\sum_{i=1}^n\sum_{k=0}^{2^{m_n}-1}\left(\sum_{l=0}^{[\alpha]} a_{k,l}\varphi_{k,l}(\frac{i}{n})\right)^2\\
&\leq ([\alpha]+1)\frac{1}{n}\sum_{i=1}^n\sum_{k=0}^{2^{m_n}-1}\sum_{l=0}^{[\alpha]} a_{k,l}^2\varphi_{k,l}^2(\frac{i}{n})\\
&\leq ([\alpha]+1)\|t\|_\nu^2\max_{k,l}\|\varphi_{k,l}\|_n^2. \qquad (57)
\end{aligned}
$$

But,

$$\begin{aligned}
\|\varphi_{k,l}\|_n^2 &= \frac{2^{m_n}}{n}\sum_{i=1}^{n}\left(\varphi^{(l)}(2^{m_n}\frac{i}{n}-k)\right)^2 \\
&\leq \frac{1}{2}\sum_{i\geq kn2^{-m_n}}^{[(k+1)n2^{-m_n}]}\left(\varphi^{(l)}(2^{m_n}\frac{i}{n}-k)\right)^2 \\
&\leq \frac{3}{2}\max_{l=0,\ldots,[\alpha]}\|\varphi^{(l)}\|_\infty^2, \qquad (58)
\end{aligned}$$

since $n2^{-m_n}\leq 2$. Putting together (57) and (58) we obtain (ii) with $R^2=1.5(1+[\alpha])\max_{l=0,\ldots,[\alpha]}\|\varphi^{(l)}\|_\infty^2$.

7.5. **Proof of Lemma 3.1.** We denote by $|.|_2$ the Euclidean norm of $\mathbb{R}^{|\Lambda_n|}$. Let $(\mu_\lambda)_{\lambda\in\Lambda_n}$ be the sequence of the eigenvalues of Φ_n, since Φ_n is positive all of them are nonnegative. As Φ_n is symmetric, there exists an orthogonal matrix U_n such that ${}^t\!U_n\Phi_nU_n=\Psi_n$ where Ψ_n is diagonal. We have that

$$\begin{aligned}
\sup_{t\in\mathcal{S}_n\backslash\{0\}}\frac{\|t\|_n^2}{\|t\|_\nu^2} &= \sup_{|a|_2=1}\|\sum_{\lambda\in\Lambda_n}a_\lambda\varphi_\lambda\|_n^2=\sup_{|a|_2=1}{}^t a\Phi_n a=\sup_{|U_na|_2=1}{}^t(Ua)\Phi_n(Ua) \\
&= \sup_{|a|_2=1}{}^t a\Psi_n a==\sup_{|a|_2=1}\sum_{\lambda\in\Lambda_n}\mu_\lambda a_\lambda^2=\max\{\mu_\lambda/\ \lambda\in\Lambda_n\}.
\end{aligned}$$

Similarly,

$$\inf_{t\in\mathcal{S}_n\backslash\{0\}}\frac{\|t\|_n^2}{\|t\|_\nu^2}=\inf\{\mu_\lambda/\ \lambda\in\Lambda_n\}$$

and thus

$$\sup_{t\in\mathcal{S}_n\backslash\{0\}}\frac{\|t\|_\nu^2}{\|t\|_n^2}=\left(\inf_{t\in\mathcal{S}_n\backslash\{0\}}\frac{\|t\|_n^2}{\|t\|_\nu^2}\right)^{-1}=\sup\{\mu_\lambda^{-1}/\ \lambda\in\Lambda_n\}=\rho\left(\Phi_n^{-1}\right).$$

8. Appendix

In this section we recall two moment inequalities on sum of independent centered random variables which are repeatedly used throughout this paper.

Theorem 8.1 (Rosenthal's Inequality)**.** *Let $U_1,\ldots,U_n$ be independent centered random variables with values in $\mathbb{R}$. For any $p\geq 2$ we have,*

$$\mathbb{E}\left[\left|\sum_{i=1}^{n}U_i\right|^p\right]\leq C(p)\left(\sum_{i=1}^{n}\mathbb{E}[|U_i|^p]\ +\ \left(\sum_{i=1}^{n}\mathbb{E}[U_i^2]\right)^{p/2}\right). \qquad (59)$$

For the proof of this inequality, we refer to Petrov (1995).

The next result explores the case where $p \in [1, 2]$. To our knowledge the result is due to von Bahr and Esseen (1965).

Theorem 8.2. *Let $U_1, ..., U_n$ be independent centered random variables with values in $\mathbb{R}$. For any $1 \leq p \leq 2$ we have,*

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} U_i\right|^p\right] \leq 8 \sum_{i=1}^{n} \mathbb{E}[|U_i|^p]. \tag{60}$$

References

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

Birgé, L., and Massart, P. (1997). From model selection to adaptative estimation. *In Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics (D.Pollard, E. Torgensen and G. Yang, eds.),* 55 – 87. *Springer-Verlag, New York.*

DeVore, R.A. and Lorentz, G.G. (1993). *Constructive approximation.* Springer-Verlag, Berlin.

Efroimovich, S., and Pinsker, M. (1984). Learning algorithm for nonparametric filtering. *Auto. Remote Control* **11**, 1434 – 1440.

Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22**, 835 – 866.

Laurent, B., and Massart, P. (1998). Adaptive estimation of a quadratic functional by model selection. *Technical Report. 98.81, Université de Paris-Sud.*

Li, K.C. (1987). Asymptotic optimality for C_p,C_l, cross-validation and generalized cross validation: discrete index set. *Ann. Statist.* **15**, 958 – 975.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661 – 675.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, 758 – 765.

Petrov, V.V. (1995). *Limit theorems of probability theory. Sequences of independent random variables.* Oxford Studies in Probability 4. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York.

Polyak, B.T. and Tsybakov, A.B. (1990). Asymptotic optimality of the C_p - test for the orthogonal series estimation of regression. *Theory Probab. Appl.* **35**, 293 – 306.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45 – 54.

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

van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **18**, 907 – 924.

von Bahr, B. and Esseen, C.G. (1965). Inequalities for the rth absolute moment of a sum of random variables $1 \leq r \leq 2$. *Annals Math. Statist.* **36**, 299 − 303.

Acknowledgments : The author is deeply grateful to an anonymous referee for a number of constructive suggestions.

Y. Baraud, Ecole Normale Supérieure, DMI

E-mail address: baraud@dmi.ens.fr