ESAIM: Probability and Statistics July 2002, Vol. 6, 127–146
URL: http://www.emath.fr/ps/ DOI: 10.1051/ps:2002007

MODEL SELECTION FOR REGRESSION ON A RANDOM DESIGN*

Yannick Baraud[1]

Abstract. We consider the problem of estimating an unknown regression function when the design is random with values in $\mathbb{R}^k$. Our estimation procedure is based on model selection and does not rely on any prior information on the target function. We start with a collection of linear functional spaces and build, on a data selected space among this collection, the least-squares estimator. We study the performance of an estimator which is obtained by modifying this least-squares estimator on a set of small probability. For the so-defined estimator, we establish nonasymptotic risk bounds that can be related to oracle inequalities. As a consequence of these, we show that our estimator possesses adaptive properties in the minimax sense over large families of Besov balls $\mathcal{B}_{\alpha,l,\infty}(R)$ with $R > 0$, $l \geq 1$ and $\alpha > \alpha_l$ where α_l is a positive number satisfying $1/l - 1/2 \leq \alpha_l < 1/l$. We also study the particular case where the regression function is additive and then obtain an additive estimator which converges at the same rate as it does when $k = 1$.

Mathematics Subject Classification. 62G07, 62J02.

Received March 28, 2001. Revised May 5, 2002.

Introduction

Let $\mathcal{A}$ be some subset of $\mathbb{R}^k$. We consider the problem of estimating on $\mathcal{A}$ the unknown function s mapping $\mathbb{R}^k$ into $\mathbb{R}$ in the following regression framework:

$$Y_i = s(X_i) + \xi_i \quad i = 1, \cdots, n. \tag{1}$$

The X_i's are independent random variables with values in $\mathcal{A}$ and the ξ_i's are i.i.d. zero mean random variables admitting a finite variance denoted by σ^2. For simplicity, we assume all along that σ^2 is known. However, the results contained in this paper would only be slightly modified by replacing σ^2 by some suitable estimator as, for example, the one proposed in Baraud [1] (Sect. 6). Throughout this paper, we assume that the sequences of X_i's and ξ_i's are independent. For each $i \in \{1, \cdots, n\}$ we denote by μ_i the distribution of X_i and set $\mu = n^{-1} \sum_{i=1}^n \mu_i$. By assuming that the X_i's are not necessarily identically distributed we have in mind to handle the particular case of deterministic X_i's for which $\mu_i = \delta_{X_i}$. Throughout the paper we fix some reference measure ν supported on $\mathcal{A}$. Unlike μ, we suppose that ν is known and our assumptions concern ν. We equip

Keywords and phrases: Nonparametric regression, least-squares estimators, penalized criteria, minimax rates, Besov spaces, model selection, adaptive estimation.

* *The author wishes to thank the three anonymous referees and one of the Editors-in-chief of ESAIM for critically reading this paper and making useful remarks.*

[1] École Normale Supérieure, DMA, 45 rue d'Ulm, 75230 Paris Cedex 05, France; e-mail: yannick.baraud@ens.fr

the Hilbert space $\mathbb{L}^2(\mathcal{A},\nu)$ with its usual norm, $\|\ \|_\nu$, and assume that s belongs to $\mathbb{L}^2(\mathcal{A},\nu)$. Our aim is to establish risk bounds for our estimator with respect to $\|\ \|_\nu$.

We build our estimator of s as follows. We start with a collection of finite-dimensional linear spaces $\{S_m,\ m\in\mathcal{M}_n\}$, such that for all $m\in\mathcal{M}_n$, $S_m\neq\{0\}$. We call the S_m's *models*. The cardinality of the collection and the dimensions of the models are allowed to depend on n. The unknown function s may or may not belong to one of the models. For each $m\in\mathcal{M}_n$, we denote by D_m the dimension of S_m and by $\hat{s}_m$ the least-squares estimator of s in S_m. Let pen(.) be some function from $\mathcal{M}_n$ into $\mathbb{R}_+=[0,+\infty[$. We select $\hat{m}$ in $\mathcal{M}_n$ from the data as

$$\hat{m}=\arg\min_{m\in\mathcal{M}_n}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{s}_m(X_i))^2+\mathrm{pen}(m)\right], \tag{2}$$

and define $\tilde{s}$ as the least-squares estimator $\hat{s}_{\hat{m}}$. The estimator $\tilde{s}_*$ we propose is defined as follows: let $k_n=2\exp\left(\ln^2(n)\right)$,

$$\tilde{s}_*=\tilde{s}\ \text{ if }\ \|\tilde{s}\|_\nu\leq k_n\ \text{ and }\tilde{s}=0\text{ otherwise.} \tag{3}$$

Other choices of k_n are possible, the one proposed here will be convenient in the proofs.

The penalty function in (2) is chosen to be of the form $C_m\sigma^2D_m/n$ for all $m\in\mathcal{M}_n$. This way of penalizing is known as "complexity regularization" and was introduced by Barron and Cover [3] in density estimation. A suitable calibration of the quantity C_m in view of statistical inference is one of the main concerns of this paper. In order to explain the basic ideas underlying our approach, let us assume for a short time that the design is deterministic, that the collection $\{S_m,\ m\in\mathcal{M}_n\}$ is totally ordered for the inclusion and that the dimension of the linear subspace of $\mathbb{R}^n$, $\{(t(X_1),...,t(X_n))',\ t\in S_m\}$, is D_m for each $m\in\mathcal{M}_n$. In the sequel, we set

$$\|t\|_n^2=\frac{1}{n}\sum_{i=1}^{n}t^2(X_i),\ \forall t\in\mathbb{L}^2(\mathcal{A},\nu).$$

For deterministic X_i's, the quadratic risk of the least-squares estimator $\hat{s}_m$ with respect to $\|\ \|_n^2$ is given by

$$R_m=\mathbb{E}[\|s-\hat{s}_m\|_n^2]=\inf_{t\in S_m}\|s-t\|_n^2+\frac{D_m}{n}\sigma^2. \tag{4}$$

The quantities $\inf_{t\in S_m}\|s-t\|_n^2$ and σ^2D_m/n are respectively called the bias and the variance term and are monotone functions of D_m (the former decreases as the latter increases). Consequently, the estimator which achieves the smallest risk among the family of estimators $\{\hat{s}_m,m\in\mathcal{M}_n\}$ is the one which realizes the best trade-off between those two terms. Since the bias depends on the unknown function s, the computation of the index m which minimizes R_m is impossible. The aim of model selection procedures is to determine, solely from the data, some $\hat{m}$ among $\mathcal{M}_n$ in such a way that the risk of $\tilde{s}=\hat{s}_{\hat{m}}$ is as close as possible to the minimal one, *i.e.* $\inf_{m\in\mathcal{M}_n}R_m$. Results in this direction have been obtained in Baraud [1] (Cor. 3.1) under weak moment assumptions on the ξ_i's and Birgé and Massart [6] under Gaussian assumptions. In these two papers, it is shown that for a suitable choice of the penalty function pen(.) in (2), the estimator $\hat{s}_{\hat{m}}$ satisfies for some constant C

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right]\leq C\inf_{m\in\mathcal{M}_n}\left[\inf_{t\in S_m}\|s-t\|_n^2+\mathrm{pen}(m)\right]+\frac{\Sigma_n}{n}, \tag{5}$$

where Σ_n denotes a positive number. When pen(m) is of order $\sigma^2 D_m/n$ for all $m \in \mathcal{M}_n$ and Σ_n remains bounded for all values of n simultaneously, we derive from (4) and (5) that for some constant C',

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leq C' \inf_{m\in\mathcal{M}_n} \mathbb{E}\left[\|s-\hat{s}_m\|_n^2\right] \tag{6}$$

$$= C' \inf_{m\in\mathcal{M}_n} \left[\inf_{t\in S_m} \|s-t\|_n^2 + \frac{D_m}{n}\sigma^2\right]. \tag{7}$$

Inequality (6) shows that the selected estimator achieves, up to a constant, the infimum of the risks among the collection of estimators $\hat{s}_m$'s. This inequality is usually called an oracle inequality as introduced in Donoho and Johnstone [13]. An interesting feature of (7) is that the constant C' does not depend on s and n. This makes it possible to derive adaptation properties in the minimax sense for $\tilde{s}$ when the collection of models is suitably chosen. Such properties are obtained by balancing the bias and variance term in the right-hand side of (7) under the *a posteriori* assumption that s belongs to some class of smooth functions.

In the regression framework given by (1) with random design points, inequalities such as (5) (note that $\|\ \|_n$ is now random) can be deduced from Yang [20], Baraud [1] and Birgé and Massart [6] by deconditioning on the X_i's. In the present paper, our aim is to obtain an analogue of (5) with $\|\ \|_\nu$ in place of $\|\ \|_n$.

In the literature, the problem of building an estimator which achieves (up to a constant) the minimal risk among a family $\mathcal{F}$ of estimators given beforehand was addressed by several authors. In our regression setting with random design points, when the risk is measured with respect to an integral $\mathbb{L}^2$-norm, we mention the work of Barron *et al.* [2], Kohler [15], Wegkamp [19], Yang [21] and Catoni [9]. In the first three papers, the authors built their estimators by selecting one among the family of estimators $\mathcal{F}$. In Barron *et al.* [2] and Kohler [15] (see also Kohler [16]), the proposed selection rule was based on a penalized criterion which was related to ours. In the paper by Wegkamp, the data were split in two parts: one part was used to generate the collections of estimators, the other part to select one among the collection. In the two other papers, the estimator was based on a progressive mixture of those lying in $\mathcal{F}$. In all these papers, strong integrability conditions on the errors are required. Besides, the proposed estimators are depending on a known upper bound B on the sup-norm of the regression function $\|s\|_\infty$. Consequently, the risk bounds established for these estimators involve constants that are increasing functions of B and hence, these constants become large if one chooses B large enough to satisfy the condition $\|s\|_\infty \leq B$.

Under suitable assumptions on the collection of models and on the distribution of the design points, we do not assume that an upper bound on $\|s\|_\infty$ is known. Besides, we do not exclude the case where the errors have only few finite moments. The risk bounds we establish are similar to Inequality (5) except that the loss $\|\ \|_n^2$ is replaced by the loss $\|\ \|_\nu^2$. In the case where $\mathcal{A} = [0,1]$ and $\nu = \mathrm{d}x$, we derive from these bounds uniform risk bounds over families of Besov balls $\mathcal{B}_{\alpha,l,\infty}(R)$ with $R > 0$ and l, α in suitable intervals. These bounds allow to establish the rate optimality of our estimation procedure in the minimax sense. More precisely, by considering a collection of linear spaces based on piecewise polynomials of degree less than $r \geq 1$, we show that our estimator is adaptive in the minimax sense over the family of Besov balls $\mathcal{B}_{\alpha,l,\infty}(R)$ for which $l \geq 2$ and $\alpha \in]0, r[$. This result is obtained under weak moment assumptions on the errors. We also consider the case of inhomogeneous Besov balls ($l < 2$). Then, we deal with a collection of linear spaces generated by wavelets of regularity r and show that our estimator achieves the rate $n^{-2\alpha/(1+2\alpha)}$ over each Besov ball $\mathcal{B}_{\alpha,l,\infty}(R)$ with $R > 0$, $l \geq 1$ and $\alpha \in]\alpha_l, r[$ where α_l is some positive number satisfying

$$\max\left\{0, \frac{1}{l} - \frac{1}{2}\right\} \leq \alpha_l < \frac{1}{l}.$$

To our knowledge, in the regression framework with random design points, the minimax rate of estimation over the Besov ball $\mathcal{B}_{\alpha,l,\infty}(R)$ with $l \geq 1$ and $\alpha \in]\alpha_l, 1/l]$ had never been described. Only the lower bound $n^{-2\alpha/(1+2\alpha)}$

was available in the literature, see for example Korostelev and Tsybakov [17] or Yang and Barron [22]. By combining this lower bound with our risk bound, we deduce that the (quadratic) minimax rate of estimation over this Besov ball is of order $n^{-2\alpha/(1+2\alpha)}$. Moreover, in the regression context considered here, our procedure is the first one to provide an adaptive estimator over classes of Besov balls of the form $\mathcal{B}_{\alpha,l,\infty}(R)$ with $l \geq 1$, $R > 0$ and $\alpha \in]\underline{\alpha}, \bar{\alpha}[$ where $\underline{\alpha}, \bar{\alpha}$ are positive numbers satisfying $\alpha_l \leq \underline{\alpha} < \bar{\alpha} < +\infty$. As R is allowed to become large, these classes are not uniformly bounded in sup-norm for any choices of $\underline{\alpha}$ and $\bar{\alpha}$. Consequently, the procedures proposed by Barron *et al.* [2], Kohler [15], Wegkamp [19], Yang [21] and Catoni [9] fail to provide an adaptive estimator over such classes.

We also give some perspective on the case where s is an additive regression function (*i.e.* s of the form $s(x_1, \ldots, x_k) = s^{(1)}(x_1) + \ldots + s^{(k)}(x_k)$). This problem was also addressed in Yang [21] under a Gaussian assumption on the errors. Stone [18] showed that the optimal rate of convergence when n tends to infinity is then independent of k. For appropriate collections of models and under weak moment assumptions on the errors, we show that a model selection procedure can provide additive estimators that converge at a rate which is free of k.

The paper is organized as follows. The main result can be found in Section 1. The adaptation properties of the estimator in the one dimensional case are presented in Section 2. In Section 3, we study the case where the regression function is additive. Sections 4 and 5 are devoted to the proofs. Throughout the paper, C denotes a constant that may vary from line to line.

1. Presentation of the results

1.1. The main assumptions

In this section, we present our main assumptions. We classify them into three groups. The first group consists of two basic assumptions that we always assume to hold. The former concerns the distribution of the X_i's and the latter the collection of models.

($\mathrm{H_{Bas}}$)**:**

(i) The measure μ admits a density with respect to ν that is bounded from below by $h_0 > 0$ and from above by $h_1 < +\infty$.

(ii) The collection of models $\{S_m,\ m \in \mathcal{M}_n\}$ is finite and consists of linear subspaces of some larger linear space $\mathcal{S}_n \subset \mathbb{L}^2(\mathcal{A}, \nu) \cap \mathbb{L}^\infty(\mathcal{A}, \nu)$ satisfying

$$\dim(\mathcal{S}_n) = N_n < n.$$

In addition, $S_m \neq \{0\}$ for all $m \in \mathcal{M}_n$.

The space $\mathcal{S}_n$ may or may not belong to the collection. The quantities h_1 and h_0 need not be known when $\mu \neq \nu$.

The second group of assumptions is concerned with the distribution of the errors.

($\mathrm{H_{Mom}(a_0)}$)**:** Let $a_0 > 0$. There exists some $p > 2(2 + a_0)$ such that $\tau_p = \mathbb{E}[|\xi_1|^p] < +\infty$.

($\mathrm{H_{Gaus}}$)**:** The ξ_i's are i.i.d. Gaussian random variables.

By assuming ($\mathrm{H_{Gaus}}$), we extend the result established under ($\mathrm{H_{Mom}(a_0)}$) to more general collections of models.

The third group of assumptions deals with the structure of the $\mathbb{L}^2(\mathcal{A}, \nu)$-orthonormal bases of $\mathcal{S}_n$. Hereafter, if $\mathcal{X}$ is a finite set, $|\mathcal{X}|$ denotes its cardinality.

($\mathrm{H_{Con}}$)**:** There exists some constant $K \geq 1$ such that for all $t \in \mathcal{S}_n$

$$\sup_{x \in \mathcal{A}} |t(x)| = \|t\|_\infty \leq K\sqrt{N_n}\|t\|_\nu. \tag{8}$$

($\mathrm{H_{Loc}}$)**:** There exist a $\mathbb{L}^2(\mathcal{A},\nu)$-orthonormal basis of $\mathcal{S}_n$, $(\varphi_\lambda)_{\lambda\in\Lambda_n}$ (where Λ_n is an index set satisfying $|\Lambda_n| = N_n$) and a constant $K \geq 1$ such that
(i) for all $\lambda \in \Lambda_n$, $|\{\lambda' \in \Lambda_n, \varphi_\lambda\varphi_{\lambda'} \not\equiv 0\}| \leq K$;
(ii) $\sup_{\lambda\in\Lambda_n} \|\varphi_\lambda\|_\infty^2 \leq KN_n$.

Condition ($\mathrm{H_{Con}}$) is related to the structure of the orthonormal bases of $\mathcal{S}_n$. It is shown in Birgé and Massart [5] (Lem. 1) that $\mathcal{S}_n$ satisfies this condition if and only if there exists an orthogonal basis $(\varphi_\lambda)_{\lambda\in\Lambda_n}$ that satisfies

$$\left\| \sum_{\lambda\in\Lambda_n} \varphi_\lambda^2 \right\|_\infty \leq K^2 N_n. \tag{9}$$

Moreover if (8) is fulfilled then any orthonormal basis satisfies (9). If $\mathcal{S}_n$ satisfies ($\mathrm{H_{Loc}}$) then it satisfies ($\mathrm{H_{Con}}$) with the same constant K. Indeed, if ($\mathrm{H_{Loc}}$) is fulfilled then for all $x \in \mathcal{A}$

$$\sum_{\lambda\in\Lambda_n} \varphi_\lambda^2(x) \leq K \sup_{\lambda\in\Lambda_n} \|\varphi_\lambda\|_\infty^2 \tag{10}$$

$$\leq K^2 N_n, \tag{11}$$

which leads to (9) and ($\mathrm{H_{Con}}$). Therefore, Condition ($\mathrm{H_{Loc}}$) is more restrictive than ($\mathrm{H_{Con}}$). Condition ($\mathrm{H_{Loc}}$) is satisfied when $\mathcal{S}_n$ consists of piecewise polynomials or wavelets with compact supports for example. This condition allows us to weaken the constraint on the dimension of $\mathcal{S}_n$ in our main theorem. Finally, let us mention that the assumption that the basis $(\varphi_\lambda)_{\lambda\in\Lambda_n}$ is orthonormal in ($\mathrm{H_{Loc}}$) can be weakened by assuming that it is a Riesz basis which means that there exist two positive constants C_1 and C_2 such that

$$C_1 \sum_{\lambda\in\Lambda_n} \beta_\lambda^2 \leq \left\| \sum_{\lambda\in\Lambda_n} \beta_\lambda\varphi_\lambda \right\|_\nu^2 \leq C_2 \sum_{\lambda\in\Lambda_n} \beta_\lambda^2.$$

This allows to handle the case where $\mathcal{S}_n$ is generated by splines.

1.2. **The main theorem**

In this section, we establish risk bounds for the estimator $\tilde{s}_*$ defined by (3). We distinguish between two settings. For each of these we introduce specific penalty functions and notations.

In our first setting, we only assume that the errors satisfy some moment condition. When the collection is not too "rich", we shall see in the comment following the theorem that the existence of few finite moments is enough to obtain an oracle inequality similar to (7).

(K1): ($\mathrm{H_{Mom}(a_0)}$) holds. Given some $\theta > 0$, we define the penalty term as

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

We set

$$\Sigma_n = \sum_{m\in\mathcal{M}_n} D_m^{-a_0}. \tag{12}$$

In the second setting, we assume that the errors are Gaussian.

(K2): ($\mathrm{H_{Gaus}}$) holds. Given some $\theta > 0$ and a sequence of nonnegative numbers $\{L_m,\ m \in \mathcal{M}_n\}$ we define the penalty term as

$$\mathrm{pen}(m) = (1+\theta)(1+\sqrt{2L_m})^2 \frac{D_m}{n}\sigma^2 \quad \text{for all } m \in \mathcal{M}_n.$$

We set

$$\Sigma_n = \sum_{m \in \mathcal{M}_n} \exp\left(-L_m D_m\right). \tag{13}$$

We have the following result:

Theorem 1.1. *Let s be some function in $\mathbb{L}^2(\mathcal{A},\nu)$. Assume that* ($\mathrm{H_{Bas}}$) *holds. Assume that either* ($\mathrm{H_{Loc}}$) *holds and that* $\dim(\mathcal{S}_n) = N_n \leq K^{-3} n/\ln^3(n)$ *or that only* ($\mathrm{H_{Con}}$) *holds and that* $N_n \leq K^{-1}\sqrt{n/\ln^3(n)}$. *Under either* (K1) *or* (K2)*, the estimator $\tilde{s}_*$ defined by (3) satisfies*

$$\mathbb{E}\left[\|s-\tilde{s}_*\|_\nu^2\right] \leq C\left[\inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|s-t\|_\nu^2 + \mathrm{pen}(m)\right) + \varepsilon_n(s)\right] \tag{14}$$

where

$$\varepsilon_n(s) = \frac{\Sigma_n}{n} + \left(\|s\|_\nu^2 + 1\right)\exp\left(-2\ln^2(n)\right).$$

The constant C depends on h_0, h_1, θ, K and also on a_0, p, τ_p under (K1).

The aim of Inequality (14) is to provide some perspective on the way the risk bound of $\tilde{s}_*$ depends on the penalty term and the collection of models. Note that the dependency of the risk bound with respect to the collection (*via* Σ_n) depends on the integrability properties of the errors. Inequality (14) also allows to derive an inequality which relates to the oracle one (7) established in the case of deterministic design points. To obtain it, some additional assumptions will be used:

($O1$) there exists some finite constant $\zeta > 0$ such that

$$\sup_{n\geq 1}\sup_{m\in\mathcal{M}_n}\frac{\mathrm{pen}(m)}{\sigma^2 D_m/n} \leq \zeta;$$

($O2$) there exists some finite constant $\Sigma > 0$ such that

$$\sup_{n\geq 1}\Sigma_n \leq \Sigma;$$

($O3$) the function s satisfies

$$\|s\|_\nu^2 \leq K' \exp\left(2\ln^2(n)\right)\frac{\sigma^2}{n},$$

for some positive constant K'.

Then, under ($O1$), ($O2$), ($O3$) we deduce from (14) that

$$\mathbb{E}\left[\|s-\tilde{s}_*\|_\nu^2\right] \leq C' \inf_{m\in\mathcal{M}_n}\left[\inf_{t\in S_m}\|s-t\|_\nu^2 + \frac{D_m}{n}\sigma^2\right], \tag{15}$$

for some constant C' independent from n and s. We call such an inequality an oracle-type inequality by analogy with Inequality (7).

In the setting given by (K1), ($O1$) is satisfied with $\zeta = 1+\theta$. Condition ($O2$) is fulfilled when the collection of models is not too "rich", that is, when it does not contain too many models with the same dimension. Typically,

when the collection contains one model per dimension D and when the errors admit a moment of order $p > 6$, $(O2)$ is satisfied by taking $\Sigma = \sum_{D\geq 1} D^{-a}$ for any $a \in]1, p/2 - 2[$.

In the setting given by (K2), $(O1)$ is satisfied if one takes all the L_m's equal to some positive constant L. Then, take $\zeta = (1+\theta)(1+\sqrt{2L})^2$. For this choice of L_m's, $(O2)$ is fulfilled if for some constant $L' < L$, the collection contains at most $\mathrm{e}^{L'D}$ models for each dimension D. Then, the choice

$$\Sigma = \sum_{D\geq 1} \mathrm{e}^{-(L-L')D} < +\infty$$

is suitable. Note that the Gaussian assumption of the errors allows to obtain an oracle-type inequality for more general collection of models. We shall take advantage of this property in our Section 2.4.

Under $(\mathrm{H_{Loc}})$, the constraint on N_n is mild. When $k = 1$, *i.e.* when s is defined on the real line, Condition $(\mathrm{H_{Loc}})$ is fulfilled by many spaces $\mathcal{S}_n$ of interest. This makes it possible to obtain adaptation properties for our estimator over large classes of functions (see our Sect. 2). In contrast, Condition $(\mathrm{H_{Loc}})$ is no longer satisfied in the additive setting described in Section 3. Nevertheless, for a suitable choice of the collection of models, Condition $(\mathrm{H_{Con}})$ still holds. Although the dimensions of the models of the collection are not allowed to be larger than $\sqrt{n}$, our estimator is proven to achieve the rate $n^{-2\alpha/(2\alpha+1)}$ over the set of additive functions s whose additive components belong to Besov balls $\mathcal{B}_{\alpha,2,\infty}(R)$ with $\alpha > 1/2$ (see Sect. 3.2). Let us recall that the constraint $\alpha > 1/2$ ensures that the functions belonging to $\mathcal{B}_{\alpha,2,\infty}$ are continuous.

The result of Theorem 1.1 holds for any choice of the positive number θ in the penalty term. From a nonasymptotic point of view, the computations do not allow to determine some "best" choice of θ (one for which the constant C in (14) is minimum for example). However, it can be shown by computations that a choice of θ close to 0 makes the constant C blow up and therefore should not be recommended. When the errors are Gaussian, some theoretical results in this direction were obtained in Birgé and Massart [7]. By replacing σ^2 in the penalty function by some suitable estimator (for example the one proposed in Baraud [1]), some good choice of θ for practical issues can be obtained by carrying out a simulation study in the same way as Birgé and Rozenholc [8] did in the density estimation framework.

2. Uniform risk bounds over Besov balls

Throughout this section, we take $\mathcal{A}$ as the interval $[0,1]$ and ν as the Lebesgue measure on $[0,1]$. The aim of this section is to establish uniform risk bounds over Besov balls for our estimator and to compare them with the minimax ones. In the sequel, we shall say that an estimator is minimax over a set if it reaches up to a constant the minimax rate of estimation over this set.

2.1. **Besov balls**

In this section, we recall what a Besov space and a Besov ball are. We restrict ourself to the case of functions on $[0,1]$. For a more general definition of those spaces, we refer to the book by DeVore and Lorentz [12].

Let us start with some definitions. Let d be a positive integer and h a positive number, the d-th order difference of a function f on $[0,1]$ is defined by

$$\Delta_h^d(f,x) = \sum_{k=1}^{d} \left(\begin{array}{c} d \\ k \end{array} \right) (-1)^{d-k} f(x+kh),$$

where x and $x+\mathrm{d}h$ belong to $[0,1]$. A real valued function f on $[0,1]$ belongs to the Besov space $\mathcal{B}_{\alpha,l,\infty}$ if f belongs to $\mathbb{L}^l([0,1],\mathrm{d}x)$ and if for $d=[\alpha]+1$ ($[\alpha]$ denotes the integer part of α)

$$|f|_{\alpha,l}=\sup_{y>0} y^{-\alpha} w_d(f,y)_l < +\infty$$

where

$$(w_d(f,y)_l)^l = \sup_{0<h\leq y} \int_0^{1-dh} \left|\Delta_h^d(f,x)\right|^l \mathrm{d}x.$$

For $f\in\mathcal{B}_{\alpha,l,\infty}$, the quantity

$$\|f\|_{\alpha,l}=\left(\int_0^1 |f(u)|^l \mathrm{d}u\right)^{1/l}+|f|_{\alpha,l}$$

is the Besov norm associated to this Besov space and for $R>0$ the Besov ball $\mathcal{B}_{\alpha,l,\infty}(R)$ is defined as

$$\mathcal{B}_{\alpha,l,\infty}(R)=\{f\in\mathcal{B}_{\alpha,l,\infty},\ \|f\|_{\alpha,l}\leq R\}\cdot$$

2.2. **The collections of models**

In this section, we present three particular collections of models. Let us start with a collection consisting of piecewise polynomials on regular grids. In the sequel, $\mathbb{1}\{\mathcal{X}\}$ denotes the indicator of the set $\mathcal{X}$.

(P) For some positive integer J_n, let $\mathcal{M}_n$ be the set of integers $\{0,\cdots,J_n\}$ and S_m the linear space consisting of the functions t on $[0,1]$ of the form

$$t(x)=\sum_{j=1}^{2^m} P_j(x)\mathbb{1}\left\{(j-1)2^{-m}\leq x<j2^{-m}\right\},$$

where the P_j's are polynomials of degree less than r. We set $\mathcal{S}_n=S_{J_n}$.

For each $m\in\mathcal{M}_n$, the dimension of S_m is $D_m=r2^m$. In particular, N_n and J_n are related by the equality $N_n=r2^{J_n}$. One can verify that the linear space $\mathcal{S}_n$ satisfies ($\mathrm{H_{Loc}}$).

The two other collections are based on wavelets. Hereafter, for all integers $j\geq 0$, we denote by $\Lambda(j)$ the set $\{(j,k),\ k=1,\cdots,2^j\}$ and consider an $\mathbb{L}^2([0,1],\mathrm{d}x)$-orthonormal system of compactly supported wavelets

$$\{\phi_{J_0,k},\ (J_0,k)\in\Lambda(J_0)\}\cup\left\{\psi_{j,k},\ (j,k)\in\bigcup_{J=J_0}^{+\infty}\Lambda(J)\right\}$$

of regularity r built by Cohen *et al.* [10] (J_0 denotes some integer). Those wavelets are derived from Daubechies' wavelets [11] at the interior of $[0,1]$ and are boundary corrected at the endpoints. For both collections and for some integer $J_n\geq 1$, we take $\mathcal{S}_n$ as the linear span generated by the $\phi_{J_0,k}$'s for $(J_0,k)\in\Lambda(J_0)$ together with $\psi_{j,k}$'s for $(j,k)\in\bigcup_{J=J_0}^{J_n-1}\Lambda(J)$. We have that $N_n=\dim(\mathcal{S}_n)=2^{J_n}$. Besides, it follows from the construction of these wavelets that the linear space $\mathcal{S}_n$ satisfies ($\mathrm{H_{Loc}}$). Indeed, according to Cohen *et al.* [10] (Sect. 4), there exists a family of orthonormal wavelets $\phi_{J_n,k}$'s for $(J_n,k)\in\Lambda(J_n)$ which generates $\mathcal{S}_n$ and satisfies ($\mathrm{H_{Loc}}$). This family is derived from that of the $\phi_{J_0,k}$'s by rescaling.

Let us now introduce the second collection of models.

(W) Let $\mathcal{M}_n=\{J_0,...,J_n-1\}$ and for each $m\in\mathcal{M}_n$, let S_m be the linear space generated by the $\phi_{J_0,k}$'s for $(J_0,k)\in\Lambda(J_0)$ together with $\psi_{j,k}$'s for $(j,k)\in\bigcup_{J=J_0}^{m}\Lambda(J)$.

Note that for each $m \in \mathcal{M}_n$, $\dim(S_m) = 2^{J_0} + \sum_{j=J_0}^{m} 2^j = 2^{m+1}$.

In order to define our third collection, let us introduce some additional notations. For $a > 2$, $x \in]0,1[$ and integers j, J such that $j \geq J$, we set

$$K_{j,J} = [\mathcal{L}(2^{J-j})2^J] \text{ and } \mathcal{L}(x) = \left(1 - \frac{\ln x}{\ln 2}\right)^{-a}, \tag{16}$$

where $[x]$ denotes the integer part of x. We define our collection of models as follows.

(W') For all $J \in \{J_0, ..., J_n\}$, we set

$$\mathcal{M}_n^J = \left\{ \bigcup_{j=J_0}^{J-1} \Lambda(j) \bigcup_{j=J}^{J_n-1} m_j,\ m_j \subset \Lambda(j),\ |m_j| = K_{j,J} \right\},$$

and $\mathcal{M}_n = \bigcup_{J=J_0}^{J_n-1} \mathcal{M}_n^J$. For each $m \in \mathcal{M}_n$, we define S_m as the linear span generated by the $\phi_{J_0,k}$'s for $(J_0,k) \in \Lambda(J_0)$ together with the $\psi_{j,k}$'s for $(j,k) \in m$.

In words, choosing a model among this collection amounts to choosing some integer J among $\{J_0, \cdots, J_n\}$ and a sequence of subsets $m_j \subset \Lambda(j)$ with $j \in \{J, \cdots, J_n - 1\}$ such that for each j, the cardinality of m_j is $K_{j,J}$. As the sequence of integers $K_{j,J}$ is nonincreasing with j, the cardinality of those sets decreases as j increases. For each $J \in \{J_0, \cdots, J_n\}$ and $m \in \mathcal{M}_n^J$, the dimension of S_m satisfies

$$2^J = 2^{J_0} + \sum_{j=J_0}^{J-1} 2^j \leq D_m \leq 2^J + \sum_{j=J}^{J_n-1} K_j \leq 2^J \left(1 + \sum_{j \geq 1} j^{-a}\right).$$

Therefore, for those $m \in \mathcal{M}_n^J$, the dimension of S_m is of order 2^J.

2.3. **Convergence rates over Besov balls $\mathcal{B}_{\alpha,2,\infty}(R)$**

Obtaining a minimax estimator over one of these balls is easy. For each $\alpha \in]0,1[$, by modifying, as we did for $\tilde{s}$, the least-squares estimator over some suitable model S_m of the collection **(P)** (or **(W)**) one gets a minimax estimator. By suitable we mean that the dimension of the space must be chosen of order $n^{1/(2\alpha+1)}$ and therefore this choice unfortunately depends on the unknown parameter α. The advantage of model selection procedures is to provide a data driven choice of the index m for which one does almost as well as the optimal one. Consequently, by using those collections, the so-defined estimator is simultaneously minimax (up to a constant in the rate) over the whole range of α's in $]0, r[$. This property is usually called adaptation in the minimax sense.

Theorem 2.1. *Assume that* $(\mathrm{H_{Bas}})(i)$ *holds and that* $\mathbb{E}[|\xi_1|^{4+\delta}] < \infty$ *for some* $\delta > 0$. *Consider the collection* **(P)** *or* **(W)** *with*

$$J_n = \left[\log\left(n/(r\log^4(n))\right)/\log(2)\right], \tag{17}$$

and let $\tilde{s}_*$ *be the estimator defined by (3) with* $\mathrm{pen}(m) = 2\sigma^2 D_m/n$. *Then for all* $R > 0$ *and* $\alpha \in]0, r[$, $\tilde{s}_*$ *satisfies*

$$\sup_{s \in \mathcal{B}_{\alpha,2,\infty}(R)} \mathbb{E}\left[\|s - \tilde{s}_*\|_\nu^2\right] \leq C(\alpha, R) n^{-2\alpha/(1+2\alpha)}.$$

Proof. The arguments of the proof are similar to those given in Baraud [1] (Sect. 4) and details can be found there. We restrict to the case of the collection **(P)**, the arguments being similar for the collection **(W)**. The choice of J_n ensures that $N_n = \dim(\mathcal{S}_n)$ is or order $n/\ln^4(n)$. We apply Theorem 1.1 with the collection **(P)** and

take $a_0 = \delta/4$, $\theta = 1$. One can check that (K1) holds true, that the quantity Σ_n remains bounded independently of n and that the constraint on N_n is fulfilled as $(\mathrm{H_{Loc}})$ is satisfied. Moreover, for all $s \in \mathcal{B}_{\alpha,2,\infty}(R)$ we have $\|s\|_\nu \leq R$ and thus it remains to check that for some constant C that does not depend on n,

$$\inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|s-t\|_\nu^2+\frac{D_m}{n}\sigma^2\right)\leq Cn^{-2\alpha/(2\alpha+1)}.$$

Thanks to the approximation properties of the linear spaces S_m's defined in **(P)** (see Birgé and Massart [4], Th. 1) we have that for all $s \in \mathcal{B}_{\alpha,2,\infty}(R)$, $\inf_{t\in S_m}\|s-t\|_\nu^2 \leq C(\alpha,R)D_m^{-2\alpha}$ for all $m \in \mathcal{M}_n$. By choosing D_m of order $n^{1/(2\alpha+1)}$ we get the result. □

2.4. Convergence rates over Besov balls $\mathcal{B}_{\alpha,l,\infty}(R)$ with $l \geq 1$

In order to obtain a minimax estimator over those balls, we consider the collections of models **(W')**. We have the following result:

Theorem 2.2. *Assume that* $(\mathrm{H_{Bas}})(i)$ *and* $(\mathrm{H_{Gaus}})$ *hold. Consider the collection* **(W')** *with* J_n *defined by (17) and let us set*

$$L(a)=1+\sum_{j=0}^{+\infty}\frac{1+(a+\ln(2))j}{(1+j)^a}.$$

For all $l \geq 1$*, let us define*

$$\alpha_l=\begin{cases}\dfrac{1}{2}\left(\dfrac{1}{l}-\dfrac{1}{2}\right)\left[1+\sqrt{\dfrac{2+3l}{2-l}}\right] & \text{if } \ l\in[1,2[\\ 0 & \text{if } \ l\geq 2.\end{cases}$$

The estimator defined by (3) with

$$\mathrm{pen}(m)=2(1+\sqrt{2L(a)})^2\sigma^2D_m/n,\ \forall m\in\mathcal{M}_n$$

satisfies

$$\sup_{s\in\mathcal{B}_{\alpha,l,\infty}(R)}\mathbb{E}\left[\|s-\tilde{s}_*\|_\nu^2\right]\leq C(\alpha,R,a,r)n^{-2\alpha/(1+2\alpha)},$$

for all $l \geq 1$*,* $R > 0$ *and* $\alpha \in]\alpha_l, r[$.

It can be checked that for all $l \geq 1$,

$$\max\left\{\frac{1}{l}-\frac{1}{2},0\right\}\leq\alpha_l<\frac{1}{l}.$$

Besides the adaptation properties of $\tilde{s}_*$, we deduce from this result that the minimax rate of estimation over each Besov ball $\mathcal{B}_{\alpha,l,\infty}(R)$ with $l \geq 1$ and $\alpha \in]\alpha_l, 1/l[$ is of order $n^{-2\alpha/(2\alpha+1)}$. Indeed, this result provides an upper bound for the minimax rate, the corresponding lower bound can be easily obtained from that on the Hölder class $\mathcal{B}_{\alpha,\infty,\infty}(R)$ given in Korostelev and Tsybakov [17] (Th. 2.8.4, p. 85) (see also Yang and Barron [22], Sect. 6).

Proof. By using the inequality

$$C_N^k=\frac{N!}{k!(N-k)!}\leq\exp\left[k\left(1+\ln(N/k)\right)\right],$$

we have that for all $J \in \{J_0, \cdots, J_n\}$,

$$\begin{aligned}\ln\left(|\mathcal{M}_n^J|\right) &\leq \sum_{j\geq J} \ln\left(C_{2^j}^{K_j}\right) \leq \sum_{j\geq J} \frac{2^J}{(1+j-J)^a}[1+(j-J)\ln(2)+a\ln(1+j-J)] \\ &\leq \sum_{j\geq J} \frac{2^J}{(1+j-J)^a}[1+(a+\ln(2))(j-J)] = 2^J(L(a)-1),\end{aligned}$$

and as for all $m \in \mathcal{M}_n^J$, $D_m \geq 2^J$, we derive

$$\Sigma_n = \sum_{m\in\mathcal{M}_n} \mathrm{e}^{-L(a)D_m} \leq \sum_{J=0}^{+\infty} \sum_{m\in\mathcal{M}_n^J} \mathrm{e}^{-L(a)D_m} \leq \sum_{J\geq 0} \mathrm{e}^{2^J(L(a)-1)-L(a)2^J} = \sum_{J\geq 0} \mathrm{e}^{-2^J} = \Sigma < +\infty.$$

We now argue as in the proof of Theorem 2.1. Condition ($\mathrm{H_{Loc}}$) and the corresponding constraint on N_n are fulfilled. Besides, (K2) is satisfied with $L_m = L(a)$ for all $m \in \mathcal{M}_n$ and $\theta = 1$. The Besov ball $\mathcal{B}_{\alpha,l,\infty}(R)$ being compact in $\left(\mathbb{L}^2([0,1],\mathrm{d}x), \|\ \|_\nu\right)$ for $\alpha > 1/l - 1/2$ ($\alpha_l > 1/l - 1/2$) we get that the quantities

$$\{\|s\|_\nu,\ s \in \mathcal{B}_{\alpha,l,\infty}(R)\}$$

are uniformly bounded. In addition, we know from Birgé and Massart [4] that there exists some model S_m among the collection which satisfies both that D_m is of order $n^{1/(2\alpha+1)}$ and

$$\|s - \pi_m s\|_\nu^2 \leq C(\alpha,R)[D_m^{-2\alpha} + N_n^{-2(\alpha+(1/l-1/2)_+)}],$$

where $(x)_+$ denotes the positive part of x. If $l \geq 2$ then $N_n^{-2\alpha}$ is smaller than $n^{-2\alpha/(2\alpha+1)}$ at least for n large enough whatever $\alpha > 0$. If $l \in [1,2[$ then for n large enough $N_n^{-2(\alpha+1/2-1/l)}$ is smaller than $n^{-2\alpha/(2\alpha+1)}$ for $\alpha > \alpha_l$. This concludes the proof. □

3. The additive regression framework

In this section we study the case where the regression function s is additive, that is of the form

$$s(x) = s^{(1)}(x_1) + \cdots + s^{(k)}(x_k)\ \forall x \in \mathbb{R}^k \tag{18}$$

for some real valued functions $s^{(1)}, \cdots, s^{(k)}$. In order to ensure that such a decomposition is unique we add the constraint that for $i = 2, \cdots, k$, $\int s^{(i)}\ \mathrm{d}\nu = 0$. We assume that $\mathcal{A}$ takes the form $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_k$ for measurable sets $\mathcal{A}_i \subset \mathbb{R}$ and that the measure ν is a product measure of the form $\nu = \nu_1 \otimes \cdots \otimes \nu_k$ where for each $i = 1, \cdots, k$, ν_i is supported on $\mathcal{A}_i$. We set for $i = 1, \cdots, k$

$$\mathbb{L}^{(i)}(\mathcal{A}) = \left\{s \in \mathbb{L}^2(\mathcal{A},\nu),\ \exists s^{(i)} \in \mathbb{L}^2(\mathcal{A}_i,\nu_i)\ \forall x \in \mathcal{A},\ s(x) = s^{(i)}(x_i)\right\},$$

and

$$\mathbb{L}_0^{(i)}(\mathcal{A}) = \left\{s \in \mathbb{L}^{(i)}(\mathcal{A}),\ \int_{\mathcal{A}_i} s^{(i)}(x)\mathrm{d}\nu^{(i)}(x) = 0\right\}.$$

It is easy to see that the spaces $\mathbb{L}^{(1)}$ and $\mathbb{L}_0^{(i)}$ for $i = 2, \cdots, k$ are pairwise orthogonal in $\mathbb{L}^2(\mathcal{A},\nu)$.

3.1. **The models**

The collections of models are obtained by proceeding as follows. We consider k collections $\{S_{m_i}^{(i)},\ m_i \in \mathcal{M}_n^{(i)}\}$ $(i = 1, \cdots, k)$ that satisfy the following conditions:

- for all $i = 1, \cdots, k$, the collection $\{S_{m_i}^{(i)},\ m_i \in \mathcal{M}_n^{(i)}\}$ satisfies condition $(\mathrm{H_{Bas}})(ii)$ for some functional space $\mathcal{S}_n^{(i)} \subset \mathbb{L}^2(\mathcal{A}, \nu)$ with $\dim(\mathcal{S}_n^{(i)}) < n/k$;
- the space $\mathcal{S}_n^{(1)}$ is a subset of $\mathbb{L}^{(1)}(\mathcal{A})$ and for all $i = 2, \cdots, k$, $\mathcal{S}_n^{(i)} \subset \mathbb{L}_0^{(i)}(\mathcal{A})$.

Then, we take $\mathcal{M}_n = \mathcal{M}_n^{(1)} \times \cdots \times \mathcal{M}_n^{(k)}$ (or a subset of this set), define the collections of models $\{S_m,\ m \in \mathcal{M}_n\}$ for $m = (m_1, \cdots, m_k) \in \mathcal{M}_n$ by

$$S_m = S_{m_1}^{(1)} + \cdots + S_{m_k}^{(k)} \tag{19}$$

and set

$$\mathcal{S}_n = \mathcal{S}_n^{(1)} + \cdots + \mathcal{S}_n^{(k)}. \tag{20}$$

The sums in (19) and (20) involve linear spaces that are pairwise orthogonal in $\mathbb{L}^2(\mathcal{A}, \nu)$. As a function mapping $\mathbb{R}$ into $\mathbb{R}$ can be extended in an obvious way to a function mapping $\mathbb{R} \times \mathbb{R}^{k-1}$ into $\mathbb{R}$, the collections of models described in Section 2.2 can be used as possible examples of collections $\{S_{m_i}^{(i)},\ m_i \in \mathcal{M}_n^{(i)}\}$'s. Finally, we note that if the $\{S_{m_i}^{(i)},\ m_i \in \mathcal{M}_n^{(i)}\}$'s satisfy $(\mathrm{H_{Con}})$ then the same does for $\{S_m,\ m \in \mathcal{M}_n\}$:

Lemma 3.1. *Let $S^{(1)}, \cdots, S^{(k)}$ be a sequence of k pairwise orthogonal subspaces of $\mathbb{L}^2(\mathcal{A}, \nu)$ such that*

$$\forall i \in \{1, \cdots, k\},\ \forall t_i \in S^{(i)},\quad \|t_i\|_\infty \leq K\sqrt{D_i}\|t_i\|_\nu$$

with $D_i = \dim(S^{(i)})$. Then for all $t \in S = S^{(1)} \oplus \cdots \oplus S^{(k)}$, $\|t\|_\infty \leq K\sqrt{D}\|t\|_\nu$ with $D = \dim(S)$.

Proof. For all $(t_1, \cdots, t_k) \in S^{(1)} \times \cdots \times S^{(k)}$,

$$\|\sum_{i=1}^k t_i\|_\infty \leq \sum_{i=1}^k \|t_i\|_\infty \leq K\sum_{i=1}^k \sqrt{D_i}\|t_i\|_\nu \leq K\left(\sum_{i=1}^k D_i\right)^{1/2}\left(\sum_{i=1}^k \|t_i\|_\nu^2\right)^{1/2} = K\sqrt{D}\|\sum_{i=1}^k t_i\|_\nu.$$

□

3.2. **Uniform risk bounds**

Let us take $\mathcal{A} = [0,1]^k$. For $\alpha, R > 0$, let us denote by $\mathcal{B}_{\alpha,2,\infty}^{(k)}(R)$ the set of additive functions s given by (18) such that for all $i = 1, \cdots, k$, $s^{(i)}$ belongs to the Besov ball $\mathcal{B}_{\alpha_i,2,\infty}(R_i)$ with $\alpha_i \geq \alpha$ and $R_i \leq R$. By assuming that s is additive it is known from Stone [18] that one can avoid the "curse of dimensionality" in the estimation rate. More precisely, it is possible to achieve a rate which does not depend on k and is optimal when $k = 1$. The dependency on k appears in the constant factor only. This result is presented below.

Theorem 3.1. *Assume that $(\mathrm{H_{Bas}})(i)$ holds and that for some $\delta > 0$, $\mathbb{E}[|\xi_1|^{4+\delta}] < \infty$. Consider k collections of models $\{S_{m_i}^{(i)},\ m_i \in \mathcal{M}_n^{(i)}\}$ $(i = 1, \cdots, k)$ such that for each $i = 1, \cdots, k$, $\{S_{m_i}^{(i)},\ m_i \in \mathcal{M}_n^{(i)}\}$ is given either by* **(P)** *or* **(W)** *with*

$$J_n = \left[\log\left(\sqrt{n}/(rk\log^2(n))\right)/\log(2)\right].$$

Let $\tilde{s}_*$ be the estimator defined by (3) with $\mathrm{pen}(m) = 2\sigma^2 D_m/n$. Then for all $R > 0$, $\alpha \in]1/2, r[$ the estimator $\hat{s}$ satisfies

$$\sup_{s \in \mathcal{B}^{(k)}{}_{\alpha,2,\infty}(R)} \mathbb{E}\left[\|s - \tilde{s}_*\|_\nu^2\right] \leq C(\alpha, R, k)n^{-2\alpha/(1+2\alpha)}.$$

When α is an integer, similar rates have been obtained in Yang [20] on Sobolev classes. His results hold without any restriction on α but require some information on R.

Proof. With our choice of J_n, for all $i = 1, \cdots, k$ $\dim(\mathcal{S}_n{}^{(i)})$ is of order $\sqrt{n}/(k\log^2(n))$. Thus, we know from Lemma 3.1 that $\mathcal{S}_n$ satisfies $(\mathrm{H_{Con}})$ since the $\mathcal{S}_n{}^{(i)}$'s do. Let us take $a_0 = \delta/2$. Whatever the choice of $S_{m_i}^{(i)}$ among the collections described by **(P)** or **(W)** we have that $D_{m_i} \geq 2^{m_i}$, we derive that

$$\begin{aligned}\Sigma_n &= \sum_{(m_1,\cdots,m_k)\in \mathcal{M}_n{}^{(1)}\times\cdots\times\mathcal{M}_n{}^{(k)}} (D_{m_1}+\cdots+D_{m_k})^{-a_0} \leq \sum_{m_1=0}^{+\infty}\cdots\sum_{m_k=0}^{+\infty}(2^{m_1}+\cdots+2^{m_k})^{-a_0} \\ &\leq k^{-a_0}\sum_{m_1=0}^{+\infty}\cdots\sum_{m_k=0}^{+\infty}2^{-a_0(m_1+\cdots+m_k)/k}\end{aligned}$$

the last line being true by convexity of the function $x \to 2^x$. Consequently, Σ_n can by bounded by some constant which is free from n. We conclude the proof by using similar arguments as in the proof of Theorem 2.1. As for all $m = (m_1, \cdots, m_k) \in \mathcal{M}_n$

$$\|s - \pi_m s\|_\nu^2 = \sum_{i=1}^k \|s^{(i)} - \pi_{m_i} s^{(i)}\|_{\nu^{(i)}}^2,$$

by choosing D_{m_i} of order $n^{1/(2\alpha+1)}$ for all $i \in \{1, \cdots, k\}$ we get to the result. Note that these choices of D_{m_i}'s are possible indeed as for $\alpha > 1/2$, $n^{1/(2\alpha+1)} \leq N_n$ at least for n large enough. □

4. Proof of Theorem 1.1

The proof relies on the following propositions the proofs of which are deferred to the next section. In the sequel, $\mathbb{1}\{\mathcal{X}\}$ denotes the indicator of the set $\mathcal{X}$ and

$$\tilde{\rho}_n = \sup_{t\in\mathcal{S}_n\setminus\{0\}} \frac{\|t\|_\nu^2}{\|t\|_n^2}\cdot \tag{21}$$

Proposition 4.1. *Assume that* (K1) *or* (K2) *holds. Then $\tilde{s} = \hat{s}_{\hat{m}}$ for $\hat{m}$ defined by (2) satisfies for all $\rho_0 > 0$,*

$$\mathbb{E}\left[\|s-\tilde{s}\|_\nu^2 \mathbb{1}\{\tilde{\rho}_n \leq \rho_0\}\right] \leq C \inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|s-t\|_\nu^2 + \mathrm{pen}(m)\right) + \frac{\Sigma_n}{n} \tag{22}$$

where the constant C depends on ρ_0 but neither on s nor n.

Proposition 4.2. *Assume that* $(\mathrm{H_{Bas}})$ *holds and that either* $(\mathrm{H_{Con}})$ *or* $(\mathrm{H_{Loc}})$ *is satisfied. If* $(\mathrm{H_{Loc}})$ *holds, let $c_n = n/(K^3 N_n \ln(n))$ and let $c_n = n/(N_n^2 K^2 \ln(n))$ otherwise. Then for all $\rho_0 > h_0^{-1}$ we have*

$$\mathbb{P}\left(\tilde{\rho}_n > \rho_0\right) \leq N_n^2 \exp\left(-\frac{(h_0 - \rho_0^{-1})^2}{4h_1} c_n \ln(n)\right). \tag{23}$$

Let us now turn to the proof of the theorem. Let $\rho_0 = 2h_0^{-1}$ and set

$$\begin{aligned}\mathbb{E}_1 &= \mathbb{E}\left[\|s-\tilde{s}_*\|_\nu^2 \mathbb{1}\left\{\tilde{\rho}_n \leq \rho_0, \|\tilde{s}\|_\nu^2 \leq k_n\right\}\right], \\ \mathbb{E}_2 &= \mathbb{E}\left[\|s-\tilde{s}_*\|_\nu^2 \mathbb{1}\left\{\tilde{\rho}_n \leq \rho_0, \|\tilde{s}\|_\nu^2 > k_n\right\}\right]\end{aligned}$$

and

$$\mathbb{E}_3 = \mathbb{E}\left[\|s-\tilde{s}_*\|_\nu^2 \mathbb{1}\{\tilde{\rho}_n > \rho_0\}\right].$$

We have

$$\mathbb{E}\left[\|s-\tilde{s}_*\|_\nu^2\right]=\mathbb{E}_1+\mathbb{E}_2+\mathbb{E}_3,$$

and it remains to bound from above the three quantities $\mathbb{E}_1, \mathbb{E}_2$ and $\mathbb{E}_3$.

Upper bound for $\mathbb{E}_1$: We have

$$\mathbb{E}_1=\mathbb{E}\left[\|s-\tilde{s}\|_\nu^2 \mathbb{1}\left\{\tilde{\rho}_n\le\rho_0, \|\tilde{s}\|_\nu^2\le k_n\right\}\right]\le \mathbb{E}\left[\|s-\tilde{s}\|_\nu^2\mathbb{1}\left\{\tilde{\rho}_n\le\rho_0\right\}\right],$$

and by Proposition 4.1

$$\mathbb{E}_1\le C\inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|s-t\|_\nu^2+\mathrm{pen}(m)\right)+\frac{\Sigma_n}{n}$$

for some constant C not depending on s nor n.

Upper bound for $\mathbb{E}_2$: We have

$$\mathbb{E}_2=\|s\|_\nu^2\mathbb{P}\left(\|\tilde{s}\|_\nu>k_n,\ \tilde{\rho}_n\le\rho_0\right).$$

On the set $\{\tilde{\rho}_n\le\rho_0\}$ we know that

$$\|\tilde{s}\|_\nu\le\rho_0^{1/2}\|\tilde{s}\|_n\le\rho_0^{1/2}\left[\|s\|_n+\left(n^{-1}\sum_{i=1}^n\xi_i^2\right)^{1/2}\right]$$

and therefore by Markov's inequality

$$\begin{aligned}\mathbb{P}\left(\|\tilde{s}\|_\nu>k_n,\ \tilde{\rho}_n\le\rho_0\right) &\le \mathbb{P}\left(\rho_0\|s\|_n^2>\exp(2\ln^2(n))\right)+\mathbb{P}\left(\rho_0\sum_{i=1}^n\xi_i^2>n\exp(2\ln^2(n))\right)\\ &\le \rho_0\left(\|s\|_\mu^2+\sigma^2\right)\exp(-2\ln^2(n))\le 2h_0^{-1}\left(h_1\|s\|_\nu^2+\sigma^2\right)\exp(-2\ln^2(n))\\ &\le C\left(\|s\|_\nu^2+1\right)\exp(-2\ln^2(n))\end{aligned}$$

for some constant C depending on h_0, h_1 and σ.

Upper bound for $\mathbb{E}_3$: We have

$$\mathbb{E}_3\le 2\left(\|s\|_\nu^2+k_n^2\right)\mathbb{P}\left(\tilde{\rho}_n>\rho_0\right).$$

Under the assumptions of Theorem 1.1, we can apply Proposition 4.2 for which we know that $c_n\ge\ln^2(n)$. As $N_n<n$ we get

$$\mathbb{P}\left(\tilde{\rho}_n>\rho_0\right)\le n^2\exp\left(-\frac{h_0^2}{16h_1}\ln^3(n)\right)$$

and therefore

$$\begin{aligned}\mathbb{E}_3 &\le 2\left[\|s\|_\nu^2+4\exp\left(2\ln^2(n)\right)\right]\exp\left(-\frac{h_0^2}{16h_1}\ln^3(n)+2\ln(n)\right)\\ &\le C\left(\|s\|_\nu^2+1\right)\exp(-2\ln^2(n))\end{aligned}$$

for some constant C depending on h_0 and h_1.

The result follows by gathering these bounds.

5. Proofs of the Propositions 4.1 and 4.2

5.1. **Proof of Proposition 4.1**

Let $\mathbb{E}_\xi$ denotes the conditional expectation on the X_i's. Let us first establish some preliminary result.

Lemma 5.1. *Let $\tilde{s}$ be some estimator of s belonging to $\mathcal{S}_n$ and satisfying*

$$\mathbb{E}_\xi[\|s-\tilde{s}\|_n^2] \leq R_n^2. \tag{24}$$

Then under $(\mathrm{H_{Bas}})(i)$*, for any* $\rho_0 > 0$

$$\mathbb{E}\left[\|s-\tilde{s}\|_\nu^2 \mathbb{1}\{\tilde{\rho}_n \leq \rho_0\}\right] \quad \leq \quad (1+2h_1\rho_0)\inf_{t\in\mathcal{S}_n}\|s-t\|_\nu^2 + \ 2\rho_0\mathbb{E}[R_n^2]. \tag{25}$$

Proof. Let us set $\bar{s}_n$ the $\mathbb{L}^2(\mathcal{A},\nu)$-projection of s onto $\mathcal{S}_n$. By Pythagoras' theorem we have $\|s-\tilde{s}\|_\nu^2 = \|s-\bar{s}_n\|_\nu^2 + \|\bar{s}_n-\tilde{s}\|_\nu^2$. On the set where

$$\tilde{\rho}_n = \sup_{t\in\mathcal{S}_n\setminus\{0\}}\frac{\|t\|_\nu^2}{\|t\|_n^2} \leq \rho_0,$$

we have that $\|\bar{s}_n-\tilde{s}\|_\nu^2 \leq \rho_0\|\bar{s}_n-\tilde{s}\|_n^2$ and thus

$$\|\bar{s}_n-\tilde{s}\|_\nu^2 \leq 2\rho_0\left(\|s-\bar{s}_n\|_n^2 + \|s-\tilde{s}\|_n^2\right).$$

We derive that

$$\|s-\tilde{s}\|_\nu^2\mathbb{1}\{\tilde{\rho}_n \leq \rho_0\} \leq \|s-\bar{s}_n\|_\nu^2 + 2\rho_0\left(\|s-\bar{s}_n\|_n^2 + \|s-\tilde{s}\|_n^2\right),$$

and by taking the expectation with respect to the ξ_i's and using (24) we deduce that

$$\mathbb{E}_\xi\left[\|s-\tilde{s}\|_\nu^2\mathbb{1}\{\tilde{\rho}_n \leq \rho_0\}\right] \leq \|s-\bar{s}_n\|_\nu^2 + 2\rho_0\left(\|s-\bar{s}_n\|_n^2 + R_n^2\right).$$

By averaging now over the X_i's (note that $\mathbb{E}[\|t\|_n^2] = \|t\|_\mu^2$) we obtain that

$$\mathbb{E}\left[\|s-\tilde{s}\|_\nu^2\mathbb{1}\{\tilde{\rho}_n \leq \rho_0\}\right] \leq \|s-\bar{s}_n\|_\nu^2 + 2\rho_0\|s-\bar{s}_n\|_\mu^2 + \ 2\rho_0\mathbb{E}[R_n^2]$$

and it remains to use $(\mathrm{H_{Bas}})(i)$, namely that $\|s-\bar{s}_n\|_\mu^2 \leq h_1\|s-\bar{s}_n\|_\nu^2$, to get the result. □

5.1.1. *Proof of (22) under* (K1)

We condition on the X_i's. Under $(\mathrm{H_{Mom}(a_0)})$, we know from Corollary 3.1 in Baraud [1]

$$\mathbb{E}_\xi\left[\|s-\tilde{s}\|_n^2\right] \leq R_n^2 = C\left[\inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|s-t\|_n^2 + \mathrm{pen}(m)\right) + \frac{\Sigma_n}{n}\right],$$

with a constant C depending on $p, a_0, \theta, q, \sigma$. By applying Lemma 5.1 with R_n^2 the result follows as under $(\mathrm{H_{Bas}})$

$$\begin{aligned}\mathbb{E}[R_n^2] \quad &\leq \quad C\left[\inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\mathbb{E}[\|s-t\|_n^2] + \mathrm{pen}(m)\right) + \frac{\Sigma_n}{n}\right] \\ &\leq \quad C\left[\inf_{m\in\mathcal{M}_n}\left(h_1\inf_{t\in S_m}\|s-t\|_\nu^2 + \mathrm{pen}(m)\right) + \frac{\Sigma_n}{n}\right].\end{aligned}$$

5.1.2. *Proof of (22) under* (K2)

We condition on the X_i's and apply Theorem 2 in Birgé and Massart [6] with $\mathbb{H} = \mathbb{R}^n$, $\langle t, u\rangle = n^{-1}\sum_{i=1}^n t_i u_i$ (we identify the function t with the $\mathbb{R}^n$ vector $(t(X_1), ..., t(X_n))'$), $K = 1 + \theta$ and $\varepsilon^2 = \sigma^2$. Under ($\mathrm{H_{Gaus}}$) we obtain that

$$\mathbb{E}_\xi\left[\|s - \tilde{s}\|_n^2\right] \leq C\left[\inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|s-t\|_n^2 + \mathrm{pen}(m)\right) + \frac{\Sigma_n}{n}\right],$$

for some constant C depending on θ, σ. By arguing as in the Section 5.1.1 we derive the desired result from Lemma 5.1.

5.2. **Proof of Proposition 4.2**

The proof of (23) relies on the lemma below. In the sequel, S denotes some finite dimensional linear subspace of $\mathbb{L}^2(\mathcal{A}, \nu) \cap \mathbb{L}^\infty(\mathcal{A}, \nu)$. We shall denote by $(\varphi_\lambda)_{\lambda\in\Lambda}$ one of its orthonormal basis and set

$$V = \left(\sqrt{\int \varphi_\lambda^2\varphi_{\lambda'}^2\, d\nu}\right)_{(\lambda,\lambda')\in\Lambda\times\Lambda} \quad \text{and} \quad B = \left(\|\varphi_\lambda\varphi_{\lambda'}\|_\infty\right)_{(\lambda,\lambda')\in\Lambda\times\Lambda}.$$

For a symmetric matrix A, $\bar{\rho}(A)$ denotes the quantity

$$\sup\sum_{\lambda,\lambda'}|a_\lambda||a_{\lambda'}||A_{\lambda\lambda'}|$$

where the supremum is taken over the sequences $(a_\lambda)_{\lambda\in\Lambda}$ satisfying $\sum_\lambda a_\lambda^2 = 1$. We now define the quantity $L(\varphi)$ by

$$L(\varphi) = \max\{\bar{\rho}^2(V), \bar{\rho}(B)\}. \tag{26}$$

Under the previous notations the following result holds.

Lemma 5.2. *Under Condition* ($\mathrm{H_{Bas}}$)(i)*, for all* $\rho_0 > h_0^{-1}$*,*

$$\mathbb{P}\left(\sup_{t\in S\setminus\{0\}}\frac{\|t\|_\nu^2}{\|t\|_n^2} > \rho_0\right) \leq |\Lambda|^2\exp\left(-n\frac{(h_0 - \rho_0^{-1})^2}{4h_1L(\varphi)}\right). \tag{27}$$

Proof. Let $(\psi_\lambda)_{\lambda\in\Lambda}$ an orthonormal basis of S with respect to the inner product of $\mathbb{L}^2(\mathcal{A}, \mu)$. Let us introduce the Gram matrices

$$\Phi(X) = \left(\frac{1}{n}\sum_{i=1}^n\varphi_\lambda(X_i)\varphi_{\lambda'}(X_i)\right)_{(\lambda,\lambda')\in\Lambda\times\Lambda}$$

and

$$\Psi(X) = \left(\frac{1}{n}\sum_{i=1}^n\psi_\lambda(X_i)\psi_{\lambda'}(X_i)\right)_{(\lambda,\lambda')\in\Lambda\times\Lambda}.$$

In the sequel, $\rho(A)$ denotes the spectral radius of a symmetric matrix A and $\mathbb{B}(\mu)$ and $\mathbb{B}(\nu)$ respectively denote the unit ball of S with respect to the measure μ respectively ν. The proof of the lemma is divided into consecutive claims.

Claim 1: The following identities hold:

$$\rho(\Psi(X)-I)=\sup_{t\in\mathbb{B}(\mu)}|\nu_n(t^2)| \text{ and } \rho\left(\Psi^{-1}(X)\right)=\sup_{t\in S\setminus\{0\}}\frac{\|t\|_\mu^2}{\|t\|_n^2}\cdot \tag{28}$$

This result is derived from classical algebra, for the details of the proof we refer to Baraud [1] (proof of Lem. 3.1, Sect. 7.5).

Claim 2: Under $(\mathrm{H}_{\mathrm{Bas}})(i)$,

$$\rho(\Phi^{-1}(X))=\sup_{t\in S\setminus\{0\}}\frac{\|t\|_\nu^2}{\|t\|_n^2}\le h_0^{-1}\rho(\Psi^{-1}(X)) \tag{29}$$

and

$$\rho\left(\Psi(X)-I\right)=\sup_{t\in\mathbb{B}(\mu)}|\nu_n(t^2)|\le h_0^{-1}\sup_{t\in\mathbb{B}(\nu)}|\nu_n(t^2)|. \tag{30}$$

Since under $(\mathrm{H}_{\mathrm{Bas}})(i)$, for all $t\in S$, $h_0\|t\|_\nu^2\le\|t\|_\mu^2$, Claim 2 is a straightforward consequence of Claim 1.

Claim 3: For all $\rho_0>h_0^{-1}$, on the set $\{\sup_{t\in\mathbb{B}(\nu)}|\nu_n(t^2)|\le h_0-\rho_0^{-1}\}$, we have that $\{\rho(\Phi^{-1}(X))\le\rho_0\}$.

On the set $\{\sup_{t\in\mathbb{B}(\nu)}|\nu_n(t^2)|\le h_0-\rho_0^{-1}\}$, we derive from (30) that

$$\rho\left(\Psi(X)-I\right)\le 1-(h_0\rho_0)^{-1}.$$

Since $h_0\rho_0>1$, on this set we can ensure that $\Psi_n^{-1}(X)$ exists and satisfies

$$\rho\left(\Psi^{-1}(X)\right)=\sup_{t\in S\setminus\{0\}}\frac{\|t\|_\mu^2}{\|t\|_n^2}\le h_0\rho_0. \tag{31}$$

The result then follows from (29).

Claim 4: For all $x>0$

$$\mathbb{P}\left(\exists(\lambda,\lambda')\in\Lambda^2/|\nu_n(\varphi_\lambda\varphi_{\lambda'})|>V_{\lambda,\lambda'}\sqrt{2h_1x}+B_{\lambda,\lambda'}x\right)\le|\Lambda|^2\exp\left(-nx\right). \tag{32}$$

From Bernstein's inequality (see Birgé and Massart [5] for this particular form of the inequality) we know that for all $x>0$

$$\mathbb{P}\left(|\nu_n(\varphi_\lambda\varphi_{\lambda'})|\ge\mathbb{E}_\mu^{1/2}\left[\varphi_\lambda^2\varphi_{\lambda'}^2\right]\sqrt{2x}+\|\varphi_\lambda\varphi_{\lambda'}\|_\infty x\right)\le\exp\left(-nx\right).$$

Under condition $(\mathrm{H}_{\mathrm{Bas}})(i)$, $\mathbb{E}_\mu\left[\varphi_\lambda^2\varphi_{\lambda'}^2\right]\le h_1V_{\lambda,\lambda'}^2$ and thus

$$\begin{aligned}\mathbb{P}\left(\exists(\lambda,\lambda')\in\Lambda^2/|\nu_n(\varphi_\lambda\varphi_{\lambda'})|>V_{\lambda,\lambda'}\sqrt{2h_1x}+B_{\lambda,\lambda'}x\right)&\le\sum_{(\lambda,\lambda')\in\Lambda^2}\mathbb{P}\left(|\nu_n(\varphi_\lambda\varphi_{\lambda'})|\ge V_{\lambda,\lambda'}\sqrt{2h_1x}+B_{\lambda,\lambda'}x\right)\\&\le|\Lambda|^2\exp\left(-nx\right)\end{aligned}$$

which ends the proof of the claim.

By using the claims we are now able to prove the lemma. We derive from Claim 3 that for all $\rho_0 > h_0^{-1}$

$$\mathbb{P}\left(\rho(\Phi^{-1}(X)) > \rho_0\right) = \mathbb{P}\left(\sup_{t\in S\setminus\{0\}} \frac{\|t\|_\nu^2}{\|t\|_n^2} > \rho_0\right) \leq \mathbb{P}\left(\sup_{t\in\mathbb{B}(\nu)} |\nu_n(t^2)| > h_0 - \rho_0^{-1}\right).$$

Now, for all $t \in \mathbb{B}(\nu)$ we write $t = \sum_{\lambda\in\Lambda} a_\lambda\varphi_\lambda$ with $\sum_{\lambda\in\Lambda} {a_\lambda}^2 \leq 1$ and obtain that

$$\sup_{t\in\mathbb{B}(\nu)} \left|\nu_n(t^2)\right| \leq \sup_{|a|_2\leq 1} \left|\sum_{\lambda\in\Lambda}\sum_{\lambda'\in\Lambda} a_\lambda a_{\lambda'}\nu_n(\varphi_\lambda\varphi_{\lambda'})\right| \leq \sup_{|a|_2\leq 1} \sum_{\lambda\in\Lambda}\sum_{\lambda'\in\Lambda} |a_\lambda||a_{\lambda'}||\nu_n(\varphi_\lambda\varphi_{\lambda'})|.$$

For $x = (h_0 - \rho_0^{-1})^2/(4h_1L(\varphi))$ on the set $\{\forall(\lambda,\lambda') \in \Lambda^2, |\nu_n(\varphi_\lambda\varphi_{\lambda'})| \leq V_{\lambda,\lambda'}\sqrt{2h_1x} + B_{\lambda,\lambda'}x\}$ we have

$$\begin{aligned}\sup_{|a|_2\leq 1} \sum_{\lambda\in\Lambda}\sum_{\lambda'\in\Lambda} |a_\lambda||a_{\lambda'}||\nu_n(\varphi_\lambda\varphi_{\lambda'})| &\leq \sqrt{2h_1x}\bar\rho\left(V\right) + x\bar\rho\left(B\right)\\ = (h_0 - \rho_0^{-1})\left(\frac{1}{\sqrt{2}}\left(\frac{\bar\rho^2\left(V\right)}{L(\varphi)}\right)^{1/2} + \frac{h_0-\rho_0^{-1}}{4h_1}\frac{\bar\rho\left(B\right)}{L(\varphi)}\right) &\leq (h_0-\rho_0^{-1})\left(\frac{1}{\sqrt{2}}+\frac{1}{4}\right) \leq h_0 - \rho_0^{-1}\end{aligned}$$

and therefore for such x,

$$\mathbb{P}\left(\sup_{t\in S\setminus\{0\}} \frac{\|t\|_\nu^2}{\|t\|_n^2} > \rho_0\right) \leq \mathbb{P}\left(\exists(\lambda,\lambda') \in \Lambda^2, |\nu_n(\varphi_\lambda\varphi_{\lambda'})| > V_{\lambda,\lambda'}\sqrt{2h_1x} + B_{\lambda,\lambda'}x\right)$$

and then the result follows from Claim 4. □

5.2.1. *Proof of (23): Part I*

In the sequel, we prove (23) under ($\mathrm{H_{Con}}$) with $c_n = n/(N_n^2K^2\ln(n))$. We use the notations introduced in Section 5.1.2 and take $S = \mathcal{S}_n$. From Lemma 5.2, to obtain the result it is enough to show that for some orthonormal basis of $\mathcal{S}_n$,

$$L(\varphi) = \max\{\bar\rho^2\left(V\right), \bar\rho\left(B\right)\} \leq n/(c_n\ln(n)).$$

Under ($\mathrm{H_{Con}}$), for any $\mathbb{L}^2(\mathcal{A},\nu)$-orthonormal basis of $\mathcal{S}_n$, $(\varphi_\lambda)_{\lambda\in\Lambda_n}$, we know from (9) and (8) (applying it with $t = \varphi_\lambda$) that $\|\sum_{\lambda\in\Lambda_n}\varphi_\lambda^2\|_\infty \leq K^2N_n$ and that $\max_{\lambda\in\Lambda_n}\|\varphi_\lambda\|_\infty \leq K\sqrt{N_n}$. Thus, we derive from Cauchy–Schwarz's inequality that

$$\bar\rho^2\left(V\right) \leq \sum_{\lambda,\lambda'\in\Lambda_n}\int \varphi_\lambda^2\varphi_{\lambda'}^2\,\mathrm{d}\nu \leq |\Lambda_n|\,\|\sum_{\lambda\in\Lambda_n}\varphi_\lambda^2\|_\infty \leq K^2N_n^2$$

and

$$\bar\rho\left(B\right) \leq \left(\sum_{\lambda,\lambda'\in\Lambda_n}\|\varphi_\lambda\|_\infty^2\|\varphi_{\lambda'}\|_\infty^2\right)^{1/2} \leq K^2N_n^2.$$

Finally, under the assumption that $N_n \leq \sqrt{n/(K^2c_n\ln(n))}$ we get that $L(\varphi) \leq K^2N_n^2 \leq n/(c_n\ln(n))$ which leads to the result.

5.2.2. *Proof of (23): Part II*

In this section, we prove (23) under ($\mathrm{H_{Loc}}$) with $c_n = n/(K^3 N_n \ln(n))$. We argue as in Section 5.2.1 and use the notations introduced in Section 5.1.2. In the sequel, we take $S = \mathcal{S}_n$. Let $(\varphi_\lambda)_{\lambda\in\Lambda_n}$ be an orthonormal basis of $\mathcal{S}_n$ satisfying ($\mathrm{H_{Loc}}$) and let us set for each $\lambda \in \Lambda_n$

$$\Delta(\lambda) = \{\lambda' \in \Lambda_n,\ \varphi_\lambda \varphi_{\lambda'} \not\equiv 0\}\cdot$$

Under ($\mathrm{H_{Loc}}$)(i), for all $\lambda \in \Lambda_n$, $\Delta(\lambda) \le K$. On the one hand, under ($\mathrm{H_{Loc}}$)(ii) for all $\lambda \in \Lambda_n$ and $\lambda' \in \Delta(\lambda)$ we have $\int \varphi_\lambda^2 \varphi_{\lambda'}^2 \mathrm{d}\nu \le K N_n$ and thus,

$$\bar\rho\,(V) = \sup_{|a|_2\le 1} \sum_{\lambda,\lambda'\in\Lambda_n} |a_\lambda||a_{\lambda'}| \left(\int \varphi_\lambda^2 \varphi_{\lambda'}^2\right)^{1/2} \le \sqrt{KN_n} \sup_{|a|_2\le 1} \sum_{\lambda\in\Lambda_n} |a_\lambda| \sum_{\lambda'\in\Delta(\lambda)} |a_{\lambda'}| = \sqrt{KN_n} W_n, \tag{33}$$

where

$$W_n = \sup_{|a|_2\le 1} \sum_{\lambda\in\Lambda_n} |a_\lambda| \sum_{\lambda'\in\Delta(\lambda)} |a_{\lambda'}|.$$

On the other hand, for $\lambda \in \Lambda_n$ and $\lambda' \in \Delta(\lambda)$ $\|\varphi_\lambda \varphi_{\lambda'}\|_\infty \le K N_n$ and thus,

$$\bar\rho\,(B) = \sup_{|a|_2\le 1} \sum_{\lambda,\lambda'\in\Lambda_n} |a_\lambda||a_{\lambda'}| \|\varphi_\lambda \varphi_{\lambda'}\|_\infty \le K N_n W_n. \tag{34}$$

Let us now show that $W_n \le K$. Indeed we have

$$W_n^2 \le \sup_{|a|_2\le 1} \sum_{\lambda\in\Lambda_n} \left(\sum_{\lambda'\in\Delta(\lambda)} |a_{\lambda'}|\right)^2 \le K \sup_{|a|_2\le 1} \sum_{\lambda\in\Lambda_n} \sum_{\lambda'\in\Delta(\lambda)} a_{\lambda'}^2 = K \sup_{|a|_2\le 1} \sum_{\lambda'\in\Lambda_n} a_{\lambda'}^2 |\Delta(\lambda')| \le K^2. \tag{35}$$

By gathering (33, 34) and (35) we derive that $L(\varphi) \le K^3 N_n$ (since $K \ge 1$). Under the assumption that $N_n \le n/(K^3 c_n \ln(n))$ we have that $L(\varphi) \le n/(c_n \ln(n))$ and the result follows from Lemma 5.2.

References

[1] Y. Baraud, Model selection for regression on a fixed design. *Probab. Theory Related Fields* **117** (2000) 467-493.
[2] A. Barron, L. Birgé and P. Massart, Risk bounds for model selection *via* penalization. *Probab. Theory Related Fields* **113** (1999) 301-413.
[3] A.R. Barron and T.M. Cover, Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37** (1991) 1738.
[4] L. Birgé and P. Massart, An adaptive compression algorithm in Besov spaces. *Constr. Approx.* **16** (2000) 1-36.
[5] L. Birgé and P. Massart, Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* **4** (1998) 329-375.
[6] L. Birgé and P. Massart, Gaussian model selection. *JEMS* **3** (2001) 203-268.
[7] L. Birgé and Massart, *A generalized C_p criterion for Gaussian model selection*, Technical Report. University Paris 6, PMA-647 (2001).
[8] L. Birgé and Y. Rozenholc, *How many bins should be put in a regular histogram*, Technical Report. University Paris 6, PMA-721 (2002).
[9] O. Catoni, Statistical learning theory and stochastic optimization, in *École d'été de probabilités de Saint-Flour*. Springer (2001).
[10] A. Cohen, I. Daubechies and P. Vial, Wavelet and fast wavelet transform on an interval. *Appl. Comp. Harmon. Anal.* **1** (1993) 54-81.
[11] I. Daubechies, *Ten lectures on wavelets*. SIAM: Philadelphia (1992).
[12] R.A. DeVore and G.G. Lorentz, *Constructive approximation*. Springer-Verlag, Berlin (1993).
[13] D.L. Donoho and I.M. Johnstone, Ideal spatial adaptation *via* wavelet shrinkage. *Biometrika* **81** (1994) 425-455.
[14] D.L. Donoho and I.M. Johnstone, Minimax estimation *via* wavelet shrinkage. *Ann. Statist.* **26** (1998) 879-921.

[15] M. Kohler, Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *J. Statist. Plann. Inference* **89** (2000) 1-23.
[16] M. Kohler, Nonparametric regression function estimation using interaction least square splines and complexity regularization. *Metrika* **47** (1998) 147-163.
[17] A.P. Korostelev and A.B. Tsybakov, *Minimax theory of image reconstruction*. Springer-Verlag, New York NY, *Lecture Notes in Statis.* (1993).
[18] C.J. Stone, Additive regression and other nonparametric models. *Ann. Statist.* **13** (1985) 689-705.
[19] M. Wegkamp, *Model selection in non-parametric regression*, Preprint. Yale University (2000).
[20] Y. Yang, Model selection for nonparametric regression. *Statist. Sinica* **9** (1999) 475-499.
[21] Y. Yang, Combining different procedures for adaptive regression. *J. Multivariate Anal.* **74** (2000) 135-161.
[22] Y. Yang and A. Barron, Information-Theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** (1999) 1564-1599.