



**HAL**  
open science

# Convergence in sup-norm of least-squares estimators in regression with random design and nonparametric heteroscedastic noise

Adrien Saumard

► **To cite this version:**

Adrien Saumard. Convergence in sup-norm of least-squares estimators in regression with random design and nonparametric heteroscedastic noise. 2010. hal-00528539v1

**HAL Id: hal-00528539**

**<https://hal.science/hal-00528539v1>**

Preprint submitted on 22 Oct 2010 (v1), last revised 21 Mar 2017 (v5)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Convergence in sup-norm of least-squares estimators in regression with random design and nonparametric heteroscedastic noise

A. Saumard  
University Rennes 1, IRMAR  
adrien.saumard@univ-rennes1.fr

October 19, 2010

## Abstract

Recent advances in the theoretical analysis of optimality in model selection via penalization procedures, and more precisely concerning the validity of the Slope Heuristics first formulated by Birgé and Massart [3] and then extended by Arlot and Massart [1], have led to investigate the consistency in sup-norm of M-estimators in order to derive controls of the excess risk and of the empirical excess risk of an M-estimator, that are optimal at the first order (see [13], [14] and [12]). Indeed, such controls are one of the keystones to justify the Slope Heuristics, as claimed in [1]. In [13] (and also in [14]), the author has been able to show the consistency of least-squares estimators in an heteroscedastic with random design regression setting, on suitable linear models of histograms and piecewise polynomials. We investigate in the present paper a systematical approach of convergence in sup-norm for least-squares regression on finite dimensional linear models. We give general constraints on the structure of these models that are sufficient to derive the consistency of the considered estimators, and these constraints appear to be slightly more restrictive than the classical assumption of localized basis. Nevertheless, our approach allows to consider for example some models of compactly supported wavelets, such as Haar expansions.

**Keywords:** Least-squares estimators, sup-norm, finite-dimensional models, localized basis.

## 1 Introduction

Let  $P$  be the unknown law of independent and identically distributed data  $(X_1, \dots, X_n)$ . The sup-norm, i.e. the norm that classically define the Banach space  $L_\infty(P)$  is, as far as one can say, one of the three fundamental norms in nonparametric statistics, the two others being the classical norms endowing  $L_2(P)$  and  $L_1(P)$ . Among the fundamental tools describing the behavior of the empirical process associated to  $(X_1, \dots, X_n)$  and indexed by a given class of functions, the celebrated Talagrand's type concentration inequalities indeed require a control in sup-norm of the indexes of the considered process (see e.g., Bousquet [5] and Klein-Rio [9], for optimal constants in the concentration of the empirical process at the right of its mean, and nearly optimal constants - and best available now, as far as we know - in the concentration at left of the empirical process, respectively). Also, when one wants to achieve an accurate control of the first moment of the empirical process, it appears that informations concerning sup-norm of indexes become unavoidable, as claimed by Talagrand concerning the generic chaining techniques applied to the empirical process and its symmetrised version, the Rademacher process, see Sections 2.6, 2.7 and 4 of [16]. Such sharp controls, by upper and by below, of the first moment of an empirical process indexed by functions, appeared to be an essential step in recent and innovating proofs related to the theoretical understanding of the so-called Slope Heuristics, discovered a few years ago by Birgé and Massart [3]. It is not surprising then that the convergence in sup-norm of the considered estimators appeared as essential in the proofs of the Slope Heuristics phenomena - see [13], [14], [12]. Before describing some existing results on sup-norm of convergence of statistical estimators and the problem to be addressed, let us recall to the reader a few fundamental aspects of researches related to the Slope Heuristics.

Concerned by the central practical issue, that consists in rightly calibrating a penalty in penalization procedures, Birgé and Massart [3] have shown the existence in a general Gaussian model selection framework,

of a minimal penalty such that a model selection via penalization procedure totally misbehaves under this minimal level of penalty. Moreover, the procedure behaves quite well, in the sense that it satisfies an oracle inequality, as soon as the penalty is uniformly higher than the minimal one. They also proved the existence of a slope in the selected dimensions around this level, which is used in practice to estimate Birgé and Massart's minimal penalty. A very beautiful fact that they have shown, and that explains the practical success of the method, is that a (nearly) optimal penalty is twice the minimal one in their setting. By optimal penalty, understand a penalty that achieves a nonasymptotic oracle inequality with leading constant almost one and tending to one when the number of data tends to infinity. Based on these heuristics, many successful simulations and confrontations to real data sets have been achieved, see [2] for a survey of practical issues about the Slope Heuristics. These heuristics have then been naturally extended by Arlot and Massart [1] to more general problems of selection of M-estimators and they conjectured that the mean of the empirical excess risk of the M-estimators on each model was a good and general candidate to be the minimal penalty. They proved their conjecture on an heteroscedastic with random design regression setting, when using linear histogram models. Then Lerasle [10] recovered Arlot and Massart propositions in least-squares density estimation. More recently, these results have been generalized in papers [13], [14] and [12] considering heteroscedastic regression on linear models and maximum likelihood estimation of density using histograms. In these works, the author highlights the fact that the convergence in sup-norm of the considered M-estimators is essential, and prove it on particular models, such as histograms and piecewise polynomials, at the rate  $\sqrt{D \ln n/n}$ , where  $D$  is the linear dimension of the considered linear model.

Our goal in this paper is to derive by a more systematical approach, the consistency in sup-norm of the least-squares estimators in a general regression setting, under structural constraints on the considered linear models. To our knowledge, no such generality in the question of the consistency in sup-norm of the least-squares estimators of a regression function has been addressed yet. Nonparametric minimax rates under various regularity assumptions for the estimation in sup-norm of a regression function are well-known, see [7] and Stone [15]. Korotselev [8] moreover found the exact asymptotic constant in the minimax problem, considering the estimation of a  $\beta$ -Lipschitz regression function. This discovery has been then extended by Donoho [6], using a beautiful method inspired by optimal recovery techniques, considering the estimation of a  $\beta$ -Lipschitz function in Gaussian white noise setting, closely related to nonparametric regression with fixed design and homoscedastic Gaussian noise. Donoho also claims that "the subject area is appealing because  $L_\infty$ -loss has special importance in connection with setting fixed-width simultaneous confidence bands for an unknown regression". This connexion seems in the idea, not so far from model selection interests, even if our interest in sup-norm consistency of least-squares estimators comes from the fact that it allows to derive upper and lower bounds in probability that are optimal - and equal - for the  $L_2$ -loss in certain finite dimensional models. We also highlight the work of Tsybakov [17], concerning again the convergence in sup-norm (and the pointwise convergence) in the Gaussian white noise model, but in Sobolev classes. It is shown in this latter article that a sharp adaptative estimator - in regard with the regularity parameter of the Sobolev classes - can be constructed using developments in Fourier basis and some ideas due to Lepski. The reader interested to references in sup-norm problems for other settings than the regression one, can consult references in [17], for example concerning density estimation related problems.

The present study is made in a quite different spirit than the works cited above, since we take the more general regression framework concerning hypothesis on the noise, which is taken heteroscedastic and non-parametric, and we study the behavior of the least-squares estimator of the regression function on a linear parametric space. As we argue, this framework is directly motivated by modern issues in model selection theory, related to *effective* optimality of penalization procedures.

The paper is organized as follows. We describe in Section 2 the statistical framework of our study and we give in Section 3 a general, tractable formulation of probability upper bounds for the  $L_\infty$ -loss of the least-squares estimator. We then state in Section 4 our main result, where we derive the consistency in sup-norm of the least-squares at the rate  $\sqrt{D \ln n/n}$ , when the model is fulfilled with an orthonormal basis in  $L_2(P)$  satisfying a criterium which is a little more restrictive than the assumption of localized basis (see Section 7.4 of Massart [11]). We call these basis the "strongly localized basis". We give three explicit examples of such basis, namely histogram basis, piecewise polynomial basis and Haar basis, this latter example being the simplest case of compactly supported wavelet basis. The proofs are postponed to the end of the paper.

## 2 Regression Framework

Let  $(\mathcal{X}, \mathcal{T}_{\mathcal{X}})$  be a measurable space and set  $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ . We assume that  $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ ,  $i \in \{1, \dots, n\}$  are  $n$  i.i.d. observations with law  $P$ . The marginal law of  $X_i$  is denoted by  $P^X$ . We assume that the data satisfy the following relation

$$Y_i = s_*(X_i) + \sigma(X_i) \varepsilon_i,$$

where  $s_* \in L_2(P^X)$ ,  $\varepsilon_i$  are i.i.d. random variables with mean 0 and variance 1 conditionally to  $X_i$  and  $\sigma : \mathcal{X} \rightarrow \mathbb{R}$  is an heteroscedastic noise level. A generic random variable of law  $P$ , independent of  $(\xi_1, \dots, \xi_n)$ , is denoted by  $\xi = (X, Y)$ .

Hence,  $s_*$  is the regression function of  $Y$  with respect to  $X$ , that we want to estimate. Given a finite dimensional linear vector space  $M$ , we denote by  $s_M$  the linear projection of  $s_*$  onto  $M$  in  $L^2(P^X)$  and by  $D$  the linear dimension of the model  $M$ .

We consider on  $M$  a least-squares estimator  $s_n$  (possibly non unique), defined as follows

$$\begin{aligned} s_n &\in \arg \min_{s \in M} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - s(X_i))^2 \right\} \\ &= \arg \min_{s \in M} \{P_n(K(s))\}, \end{aligned} \tag{1}$$

where  $K : L_2(P^X) \rightarrow L_1(P)$  is the least-squares contrast, defined by

$$K(s) = (x, y) \in \mathcal{Z} \rightarrow (y - s(x))^2, \quad s \in L_2(P^X)$$

and

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, Y_i)}$$

is the empirical distribution of the data.

We denote by

$$\|s\|_2 = \left( \int_{\mathcal{X}} s^2 dP^X \right)^{1/2}$$

the quadratic norm on  $L_2(P^X)$ . Recall that we then have for all  $s \in M$ ,

$$P(Ks - Ks_*) = \|s - s_*\|_2^2 \geq 0$$

Moreover, we denote by  $s_M$  the linear projection of  $s_*$  onto  $M$  in  $L^2(P^X)$ , and by the Pythagorean theorem, we have

$$P(Ks - Ks_M) = \|s - s_M\|_2^2 \geq 0.$$

Hence, the excess risk on  $M$  is given by the quadratic norm. In addition, the least-squares contrast satisfies the following expansion, for all  $s \in M$  and for all  $z = (x, y) \in \mathcal{Z}$ ,

$$(Ks)(z) - (Ks_M)(z) = \psi_{1,M}(z) (s - s_M)(x) + ((s - s_M)(x))^2, \tag{2}$$

where

$$\psi_{1,M} : z = (x, y) \in \mathcal{Z} \mapsto \psi_{1,M}(z) = -2(y - s_M(x)).$$

We will not discuss how to choose a model  $M$  and to achieve an accurate trade-off between the bias of the model, given by  $P(Ks_M - Ks_*)$ , and the excess risk on  $M$  of the estimator  $s_n$ , given by  $P(Ks - Ks_M)$ . See for instance, [11] and [14].

### 3 The problem of convergence in sup-norm

We are interested by the quantity  $\|s_n - s_M\|_\infty$ , where  $\|\cdot\|_\infty$  is the sup-norm on  $\mathcal{Z}$ , that we want to bound from above with high probability. More precisely, for some  $\alpha > 0$ , the problem is to find  $C > 0$  such that

$$\mathbb{P}(\|s_n - s_M\|_\infty > C) \leq n^{-\alpha} . \quad (3)$$

We formulate below the problem stated in (3) in a more tractable but general formulation. Let us now define two slices of interest in  $M$ , that are localized in sup-norm. We set

$$\mathcal{F}_C^\infty := \{s \in M ; \|s - s_M\|_\infty \leq C\} \quad (4)$$

and

$$\mathcal{F}_{>C}^\infty := \{s \in M ; \|s - s_M\|_\infty > C\} = M \setminus \mathcal{F}_C^\infty . \quad (5)$$

Notice that since  $M$  is a linear vector space,  $\mathcal{F}_C^\infty$  is the closed  $L_\infty$ -ball in  $M$  centered at  $s_M$  and of radius  $C$ . By the definition of the least-squares estimator  $s_n$  given in (1), we then have

$$\begin{aligned} & \mathbb{P}(\|s_n - s_M\|_\infty \geq C) \\ & \leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_{>C}^\infty} P_n(Ks) \leq \inf_{s \in \mathcal{F}_C^\infty} P_n(Ks)\right) \\ & = \mathbb{P}\left(\sup_{s \in \mathcal{F}_{>C}^\infty} P_n(Ks_M - Ks) \geq \sup_{s \in \mathcal{F}_C^\infty} P_n(Ks_M - Ks)\right) . \end{aligned} \quad (6)$$

It is worth mentioning that inequality (6) is a general fact of M-estimation, as we only used the definition of the least-squares estimator as a M-estimator. Formulation (6) is now more tractable than the original inequality (3). Indeed, by using the fact that the least-squares contrast achieves an expansion, which is recalled in (2), we have

$$\begin{aligned} P_n(Ks_M - Ks) &= (P_n - P)(\psi_{1,M} \cdot (s_M - s)) - (P_n - P)\left((s - s_M)^2\right) - P(Ks - Ks_M) \\ &= (P_n - P)(\psi_{1,M} \cdot (s_M - s)) - (P_n - P)\left((s - s_M)^2\right) - \|s - s_M\|_2^2 . \end{aligned} \quad (7)$$

In the proof of Theorem 1, where we derive the rate of convergence in sup-norm of the least-squares estimator under general conditions on the model  $M$ , we use formula (7) to control and compare the two quantities of interest,

$$\sup_{s \in \mathcal{F}_C^\infty} P_n(Ks_M - Ks) \quad \text{and} \quad \sup_{s \in \mathcal{F}_{>C}^\infty} P_n(Ks_M - Ks) .$$

## 4 Results

We give here our mains results concerning the convergence in sup-norm of the least-squares estimator in a bounded heteroscedastic regression setting. We first give in Section 4.1 general constraints on the finite-dimensional model  $M$ , that allow us to derive in Section 4.2 a general theorem. We then give in Section 4.3 three classical examples that are covered by our theorem, namely histogram models, models of piecewise polynomials and Haar expansions on the unit interval.

### 4.1 Strongly localized basis

We first give the general assumption on the linear model  $M$  needed to prove Theorem 1 below. An orthonormal basis  $(\varphi_k)_{k=1}^D$  of  $(M, \|\cdot\|_2)$  satisfying this assumption will be called a "strongly localised basis".

**(Aslb)** there exist  $r_M > 0$ ,  $p \in \mathbb{N}_*$ , a partition  $(\Pi_i)_{i=1}^p$  of  $\{1, \dots, D\}$ , positive constants  $(A_i)_{i=1}^p$  and an orthonormal basis  $(\varphi_k)_{k=1}^D$  of  $(M, \|\cdot\|_2)$  such that  $0 < A_1 \leq A_2 \leq \dots \leq A_p$ ,

$$\sum_{i=1}^p \sqrt{A_i} \leq r_M \sqrt{D} , \quad (8)$$

and

$$\text{for all } \beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D, \left\| \sum_{k=1}^D \beta_k \varphi_k \right\|_{\infty} \leq \sum_{i=1}^p \sqrt{A_i} \max_{k \in \Pi_i} |\beta_k|. \quad (9)$$

Moreover, for every  $(i, j) \in \{1, \dots, p\}$  and  $k \in \Pi_i$ , we set

$$\Pi_{j,k} = \left\{ l \in \Pi_j ; \text{Support}(\varphi_k) \cap \text{Support}(\varphi_l) \neq \emptyset \right\}$$

and we assume that there exists a positive constant  $A_c$  such that for all  $j \in \{1, \dots, p\}$ ,

$$\max_{k \in \Pi_i} \text{Card}(\Pi_{j,k}) \leq A_c (A_j A_i^{-1} \vee 1). \quad (10)$$

It is directly seen that a strongly localized basis is a localized basis in the sense of Birgé and Massart [4], the latter notion having been introduced by these authors to derive accurate exponential bounds of the excess risk of general bounded M-estimators on sieves. More precisely, an orthonormal basis  $(\varphi_k)_{k=1}^D$  of  $(M, \|\cdot\|_2)$  is a localized basis if there exists  $r_\varphi > 0$  such that

$$\text{for all } \beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D, \left\| \sum_{k=1}^D \beta_k \varphi_k \right\|_{\infty} \leq \sqrt{D} \max_{k \in \{1, \dots, D\}} |\beta_k|. \quad (11)$$

Now, (8) and (9) imply (11). Moreover, we require in (10) a control of the number of intersections of support of element of the considered orthonormal basis to be strongly localized. As we will see in Section 4.3 below, the main practical examples of localized basis seem to be also strongly localized, as for example Haar expansions, which are a simple example of compactly supported wavelets, are strongly localized in  $L_2(P^X)$  under rather mild assumptions on  $P^X$ .

## 4.2 A Structural Theorem

We give here our main result concerning the convergence in sup-norm of the least-squares estimator in a bounded heteroscedastic regression setting. Three corollaries corresponding to explicit models will be derived in Section 4.3.

**Theorem 1** *Let  $\alpha > 0$ . Assume that  $M$  is a linear vector space of finite dimension  $D$  satisfying (A slb). Assume moreover that the following assumption holds:*

**(Ab)** *There exists a constant  $A_1 > 0$  such that  $|\psi_{1,M}(X, Y)| \leq A_1$  a.s.*

*If there exists  $A_+ > 0$  such that*

$$\max \{p^2 A_p ; D\} \leq A_+ \frac{n}{(\ln n)^2},$$

*then we have, for all  $n \geq n_0(A_+, A_c, r_M)$ ,*

$$\mathbb{P} \left( \|s_n - s_M\|_{\infty} \geq L_{A, r_M, \alpha} \sqrt{\frac{D \ln n}{n}} \right) \leq n^{-\alpha}. \quad (12)$$

Let us briefly comment on Theorem 1. We derive here a probability upper bound for the  $L_{\infty}$ -loss of the least-square estimator of a regression function towards the orthonormal projection in a finite-dimensional linear model fulfilled with

### 4.3 Examples and corollaries

#### 4.3.1 Histogram models

Let  $\mathcal{P}$  be a finite partition of  $\mathcal{X}$ . Consider the model

$$M = \left\{ \sum_{I \in \mathcal{P}} \beta_I \mathbf{1}_I ; (\beta_I)_{I \in \mathcal{P}} \in \mathbb{R}^D \right\} ,$$

where  $D := |\mathcal{P}|$  is the linear dimension of  $M$  and corresponds to the number of elements in  $\mathcal{P}$ .

The following lemma states the existence of an orthonormal localized basis in  $(M, \|\cdot\|_2)$ , if the partition  $\mathcal{P}$  is lower-regular for the law  $P^X$ . This lemma is also stated and proved in [13].

**Lemma 2** *Let consider a linear model  $M$  of histograms defined on a finite partition  $\mathcal{P}$  on  $\mathcal{X}$ , and write  $|\mathcal{P}| = D$  the dimension of  $M$ . Moreover, assume that for a positive finite constant  $c_{M,P}$ ,*

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} P^X(I)} \geq c_{M,P} > 0 . \quad (13)$$

Set, for  $I \in \mathcal{P}$ ,

$$\varphi_I = (P^X(I))^{-1/2} \mathbf{1}_I .$$

Then the family  $(\varphi_I)_{I \in \Lambda_M}$  is an orthonormal basis in  $L_2(P^X)$  and we have,

$$\text{for all } \beta = (\beta_I)_{I \in \mathcal{P}} \in \mathbb{R}^D, \quad \left\| \sum_{I \in \mathcal{P}} \beta_I \varphi_I \right\|_{\infty} \leq c_{M,P}^{-1} \sqrt{D} |\beta|_{\infty} . \quad (14)$$

By Lemma 2, we deduce that if the partition  $\mathcal{P}$  satisfies the assumption of lower regularity given in (13) then inequality (9) is satisfied for  $M$ , with  $p = 1$  and  $r_M = c_{M,P}^{-1} > 0$ . Moreover, notice that for all  $(i, j) \in \{1, \dots, D\}^2$ ,

$$\text{Card}(\Pi_{i,j}) = \delta_{i,j}$$

and in this case (Aslb) is straightforwardly satisfied.

#### 4.3.2 Piecewise polynomials

Assume that  $\mathcal{X} = [0, 1]$  is the unit interval,  $\mathcal{P}$  is a finite partition of  $\mathcal{X}$  made of intervals and let

$$M = \text{Span} \{ p_{I,j} : x \in \mathcal{X} \mapsto x^j \mathbf{1}_I ; (I, j) \in \mathcal{P} \times \{0, \dots, r\} \}$$

be the linear model of piecewise polynomials on  $\mathcal{X}$ , of degrees not larger than  $r$ . Notice that the linear dimension of  $M$  is  $(r+1)|\mathcal{P}|$ .

The following lemma states the existence, under suitable assumptions, of a localized orthonormal basis in  $(M, \|\cdot\|_2)$ .

**Lemma 3** *Let  $\text{Leb}$  denotes the Lebesgue measure on  $[0, 1]$ . Let assume that  $\mathcal{X} = [0, 1]$  and that  $P^X$  has a density  $f$  with respect to  $\text{Leb}$  satisfying, for a positive constant  $c_{\min}$ ,*

$$f(x) \geq c_{\min} > 0, \quad x \in [0, 1] .$$

Consider a linear model  $M$  of piecewise polynomials on  $[0, 1]$  with degree  $r$  or smaller, defined on a finite partition  $\mathcal{P}$  made of intervals. Then there exists an orthonormal basis  $\{\varphi_{I,j}, I \in \mathcal{P}, j \in \{0, \dots, r\}\}$  of  $(M, \|\cdot\|_2)$  such that,

$$\text{for all } j \in \{0, \dots, r\} \quad \varphi_{I,j} \text{ is supported by the element } I \text{ of } \mathcal{P},$$

and a constant  $L_{r,c_{\min}}$  depending only on  $r, c_{\min}$  exists, satisfying for all  $I \in \mathcal{P}$ ,

$$\max_{j \in \{0, \dots, r\}} \|\varphi_{I,j}\|_{\infty} \leq L_{r,c_{\min}} \frac{1}{\sqrt{\text{Leb}(I)}} . \quad (15)$$

As a consequence, if it holds

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \text{Leb}(I)} \geq c_{M, \text{Leb}} \quad (16)$$

a constant  $L_{r, c_{\min}, c_{M, \text{Leb}}}$  depending only on  $r, c_{\min}$  and  $c_{M, \text{Leb}}$  exists, such that for all  $\beta = (\beta_{I,j})_{I \in \mathcal{P}, j \in \{0, \dots, r\}} \in \mathbb{R}^{\mathcal{D}}$ ,

$$\left\| \sum_{I,j} \beta_{I,j} \varphi_{I,j} \right\|_{\infty} \leq L_{r, c_{\min}, c_{M, \text{Leb}}} \sqrt{D} |\beta|_{\infty} , \quad (17)$$

where  $D = (r+1) |\mathcal{P}|$  is the dimension of  $M$ .

Lemma 3 states that if  $\mathcal{X} = [0, 1]$  is the unit interval and  $P^X$  has a density with respect to the Lebesgue measure  $\text{Leb}$  on  $\mathcal{X}$  uniformly bounded away from zero, then there exists an orthonormal basis in  $(M, \|\cdot\|_2)$  of piecewise polynomials, where the sup-norm of its elements are suitably controlled by (15). Moreover, if we assume the lower regularity of the partition with respect to  $\text{Leb}$  then the orthonormal basis is localized, where the constant of localization in (17) depend on the maximal degree  $r$ . We notice that in the case of piecewise constant functions, we do not need to assume the existence of a density for  $P^X$  or to restrict ourselves to the unit interval. Moreover, Lemma 3 proves that under suitable assumptions, the property of strongly localised basis is recovered, exactly by the same argument than for histogram models.

### 4.3.3 Haar expansions

Let  $\mathcal{X} = [0, 1]$ . We consider now some Haar expansions on  $\mathcal{X}$ , which are special cases of wavelet expansions with compact supports. We believe that the following results, validating assumptions **(Alb)** and **(Asq)** for Haar expansions, can be generalized to some wavelet expansions with compact support, and this work is still in progress.

We set for every integers  $i, j, l \geq 0$ , satisfying  $i \leq j$  and  $1 \leq l \leq 2^i$ ,

$$\Lambda(j) = \{(j, k) ; 1 \leq k \leq 2^j\} , \quad (18)$$

$$\Lambda(j, i, l) = \{(j, k) ; 2^{j-i}(l-1) + 1 \leq k \leq 2^{j-i}l\} . \quad (19)$$

Moreover, we set

$$\Lambda(-1) = \{-1\} \quad \text{and} \quad \Lambda_m = \bigcup_{j=-1}^m \Lambda(j) .$$

Notice that for every integers  $i, j \geq 0$  such that  $i \leq j$ ,  $\{\Lambda(j, i, l) ; 1 \leq l \leq 2^i\}$  is a partition of  $\Lambda(j)$ , which means that

$$\Lambda(j) = \bigcup_{l=1}^{2^i} \Lambda(j, i, l) \quad \text{and for all } 1 \leq l, h \leq 2^i, \Lambda(j, i, l) \cap \Lambda(j, i, h) = \emptyset .$$

Let  $\phi = \mathbf{1}_{[0,1]}$ ,  $\rho = \mathbf{1}_{[0,1/2]} - \mathbf{1}_{(1/2,1]}$  and for every integers  $j \geq 0$ ,  $1 \leq k \leq 2^j$ ,

$$\rho_{j,k} : x \in [0, 1] \mapsto 2^{j/2} \rho(2^j x - k + 1) .$$

Set  $\rho_{-1} = \phi$  and let  $m \in \mathbb{N}$ . We consider the model

$$M = \text{Span} \{ \rho_{\lambda} ; \lambda \in \Lambda_m \} . \quad (20)$$

Notice that the linear dimension  $D$  of  $M$  satisfies  $D = 2^{m+1}$ . The following lemma gives an explicit localized orthonormal basis of  $(M, \|\cdot\|_2)$ .

**Lemma 4** *Let  $m \in \mathbb{N}$ . Assume that  $\mathcal{X} = [0, 1]$  and let  $M$  be the model of dimension  $D$  given by (20). Set for every integers  $j \geq 0$ ,  $1 \leq k \leq 2^j$ ,*

$$p_{j,k,-} = P^X \left( \left[ 2^{-j}(k-1), 2^{-j}(k-1/2) \right] \right) , \quad p_{j,k,+} = P^X \left( \left( 2^{-j} \left( k - \frac{1}{2} \right), 2^{-j}k \right] \right)$$



$$\varphi_{j,k} : x \in [0, 1] \mapsto \frac{1}{\sqrt{p_{j,k,+}^2 + p_{j,k,-}^2}} \left( p_{j,k,+} \mathbf{1}_{[2^{-j}(k-1), 2^{-j}(k-1/2)]} - p_{j,k,-} \mathbf{1}_{(2^{-j}(k-\frac{1}{2}), 2^{-j}k]} \right). \quad (21)$$

Moreover we set  $\varphi_{-1} = \phi$ . Assume that  $P^X$  has a density  $f$  with respect to Leb on  $[0, 1]$  and that there exist  $c_{\max}, c_{\min} > 0$  such that for all  $x \in [0, 1]$ ,

$$f(x) \geq c_{\min} > 0.$$

Then  $\{\varphi_\lambda; \lambda \in \Lambda_m\}$  is a localized orthonormal basis of  $(M, \|\cdot\|_2)$  and it holds for every integers  $j \geq 0$ ,  $1 \leq k \leq 2^j$ ,

$$\|\varphi_{j,k}\|_\infty \leq \sqrt{\frac{2}{c_{\min}}} 2^{j/2}. \quad (22)$$

Moreover, for all  $\beta = (\beta_0, \beta_\lambda)_{\lambda \in \Lambda_m} \in \mathbb{R}^D$ , it holds

$$\left\| \sum_{\lambda \in \Lambda_m} \beta_\lambda \varphi_\lambda \right\|_\infty \leq L_{c_{\min}} \sqrt{D} |\beta|_\infty \quad (23)$$

and

$$\max_{\lambda \in \Lambda_m} \sum_{\eta \in \Lambda_m} \sqrt{\|\varphi_\lambda \cdot \varphi_\eta\|} \leq L_{c_{\min}} \sqrt{D}. \quad (24)$$

By Lemma 4, we see that if  $P^X$  has a density which is uniformly bounded away from zero on  $\mathcal{Z}$ , then the model  $M$  given by (20) admits a localized orthonormal basis for the  $L_2(P)$ -norm which satisfied **(Aslb)**.

## 5 Proofs

### 5.1 Proof of Lemma 4

**Proof of Lemma 4.** Inequality (22) and the fact that  $\{\varphi_\lambda; \lambda \in \Lambda_m\}$  is an orthonormal basis of  $(M, \|\cdot\|_2)$  are straightforward and we skip their proof. Moreover, for all  $\beta = (\beta_\lambda)_{\lambda \in \Lambda_m} \in \mathbb{R}^D$ , since  $D = 2^{m+1}$  we have,

$$\begin{aligned} \left\| \sum_{\lambda \in \Lambda_m} \beta_\lambda \varphi_\lambda \right\|_\infty &\leq |\beta_{-1}| + \sum_{j=0}^m \left\| \sum_{k=1}^{2^j} \beta_{j,k} \varphi_{j,k} \right\|_\infty \\ &\leq |\beta_{-1}| + \sqrt{\frac{2}{c_{\min}}} \sum_{j=0}^m 2^{j/2} \sup_{k \in \{1, \dots, 2^j\}} |\beta_{j,k}| \\ &\leq \left( 1 + \sqrt{\frac{2}{c_{\min}}} \frac{2^{(m+1)/2} - 1}{\sqrt{2} - 1} \right) |\beta|_\infty \\ &< \left( 1 + (1 + \sqrt{2}) \sqrt{\frac{2}{c_{\min}}} \right) \sqrt{D} |\beta|_\infty. \end{aligned}$$

Hence,  $L_{c_{\min}} = 1 + (1 + \sqrt{2}) \sqrt{\frac{2}{c_{\min}}}$  is convenient in (23). Let now  $(j, k) \in \bigcup_{j=0}^m \Lambda(j)$  be fixed. Notice that for any  $0 \leq i < j$  there exists a unique  $l_{(j,k)} \in \{1, \dots, 2^i\}$  such that

$$\text{Support}(\varphi_{i, l_{(j,k)}}) \cap \text{Support}(\varphi_{j,k}) \neq \emptyset.$$

Hence, for any  $(j, k) \in \bigcup_{j=0}^m \Lambda(j)$ , we have

$$\begin{aligned} &\sum_{\lambda \in \Lambda_m} \sqrt{\|\varphi_{j,k} \cdot \varphi_\lambda\|_\infty} \\ &= \sqrt{\|\varphi_{j,k} \cdot \varphi_{-1}\|_\infty} + \sum_{0 \leq i < j} \sqrt{\|\varphi_{j,k} \cdot \varphi_{i, l_{(j,k)}}\|_\infty} + \sum_{i=j}^m \sum_{l \in \Lambda(i, j, k)} \sqrt{\|\varphi_{j,k} \cdot \varphi_{i, l}\|_\infty}, \end{aligned} \quad (25)$$

where the  $\Lambda(i, j, k)$  are defined in (19). From (25) and (22), we thus deduce that

$$\begin{aligned}
& \sum_{\lambda \in \Lambda_m} \sqrt{\|\varphi_{j,k} \cdot \varphi_\lambda\|_\infty} \\
& \leq \sqrt{\|\varphi_{j,k}\|_\infty} \left( 1 + \sum_{0 \leq i < j} \sqrt{\|\varphi_{i,l(j,k)}\|_\infty} + \sum_{i=j}^m \sum_{l \in \Lambda(i,j,k)} \sqrt{\|\varphi_{i,l}\|_\infty} \right) \\
& \leq \left( \frac{2^{j+1}}{c_{\min}} \right)^{1/4} \left( 1 + \sum_{0 \leq i < j} \left( \frac{2^{i+1}}{c_{\min}} \right)^{1/4} + \sum_{i=j}^m 2^{i-j} \left( \frac{2^{i+1}}{c_{\min}} \right)^{1/4} \right) \\
& \leq \sqrt{\frac{2}{c_{\min}}} 2^{j/4} \left( \frac{2^{j/4}}{\sqrt{2}-1} + \right)
\end{aligned}$$

■

## 5.2 Proof of Theorem 1

Before stating the proof of Theorem 1, we need two preliminary lemmas.

**Lemma 5** *Let  $\alpha > 0$ . Consider a finite-dimensional linear model  $M$  of linear dimension  $D$  and assume that  $(\varphi_k)_{k=1}^D$  is a localized orthonormal basis of  $(M, \|\cdot\|_2)$  with index of localization  $r_M > 0$ . More explicitly, we thus assume that for all  $\beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D$ ,*

$$\left\| \sum_{k=1}^D \beta_k \varphi_k \right\|_\infty \leq r_M \sqrt{D} |\beta|_\infty .$$

If **(Ab)** holds for some positive constant  $A_+$ ,

$$D \leq A_+ \frac{n}{(\ln n)^2} ,$$

then there exists a positive constant  $L_\alpha^{(2)}$  such that for all  $n \geq n_0(A_+, r_M)$ , we have

$$\mathbb{P} \left( \max_{k \in \{1, \dots, D\}^2} |(P_n - P)(\varphi_k \cdot \varphi_l)| \geq L_\alpha^{(2)} \min \{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \} \sqrt{\frac{\ln n}{n}} \right) \leq n^{-\alpha} . \quad (26)$$

**Proof of Lemma 5.** For any  $(k, l) \in \{1, \dots, D\}^2$ , we have

$$\mathbb{E} \left[ (\varphi_k \cdot \varphi_l)^2 \right] \leq \min \{ \|\varphi_k\|_\infty^2 ; \|\varphi_l\|_\infty^2 \}$$

and

$$\begin{aligned}
\|\varphi_k \cdot \varphi_l\|_\infty & \leq \min \{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \} \times \max \{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \} \\
& \leq \min \{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \} \times r_M \sqrt{D}
\end{aligned}$$

Hence, we apply Bernstein's inequality (see Proposition 2.9 in [11]) and we get,

$$\mathbb{P} \left( |(P_n - P)(\varphi_k \cdot \varphi_l)| \geq \min \{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \} \sqrt{\frac{2\gamma \ln n}{n}} + \frac{r_M \sqrt{D} \gamma \ln n}{3n} \right) \leq 2n^{-\gamma} . \quad (27)$$

Since, for all  $n \geq n_0(A_+, r_M)$ ,

$$\frac{r_M \sqrt{D} \ln n}{n} \leq r_M \sqrt{\frac{A_+ \ln n}{n}} (\ln n)^{-1/2} \leq \sqrt{\frac{2 \ln n}{\gamma n}} ,$$

we get from (27) that for all  $n \geq n_0(A_+, r_M)$ ,

$$\begin{aligned}
& \mathbb{P} \left( \max_{(k,l) \in \{1, \dots, D\}^2} |(P_n - P)(\varphi_k \cdot \varphi_l)| \geq 2 \min \{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \} \sqrt{\frac{\gamma \ln n}{n}} \right) \\
& \leq \sum_{(k,l) \in \{1, \dots, D\}^2} \mathbb{P} \left( |(P_n - P)(\varphi_k \cdot \varphi_l)| \geq 2 \min \{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \} \sqrt{\frac{\gamma \ln n}{n}} \right) \\
& \leq \sum_{(k,l) \in \{1, \dots, D\}^2} \mathbb{P} \left( |(P_n - P)(\varphi_k \cdot \varphi_l)| \geq \min \{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \} \sqrt{\frac{2\gamma \ln n}{n}} + \frac{r_M \sqrt{D} \gamma \ln n}{3n} \right) \\
& \leq 2D^2 n^{-\gamma} \leq n^{-\gamma+2} .
\end{aligned} \tag{28}$$

We deduce from (28) that (26) holds with  $L_\alpha^{(2)} = 2\sqrt{\gamma} > 0$ . ■

**Lemma 6** *Let  $\alpha > 0$ . Consider a finite-dimensional linear model  $M$  of linear dimension  $D$  and assume that  $(\varphi_k)_{k=1}^D$  is a localized orthonormal basis of  $(M, \|\cdot\|_2)$  with index of localization  $r_M > 0$ . More explicitly, we thus assume that for all  $\beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D$ ,*

$$\left\| \sum_{k=1}^D \beta_k \varphi_k \right\|_\infty \leq r_M \sqrt{D} |\beta|_\infty . \tag{29}$$

If **(Ab)** holds and for some positive constant  $A_+$ ,

$$D \leq A_+ \frac{n}{(\ln n)^2} ,$$

then there exists a positive constant  $L_{A_+, r_M, \alpha}^{(1)}$  such that for all  $n \geq n_0(A_+)$ , we have

$$\mathbb{P} \left( \max_{k \in \{1, \dots, D\}} |(P_n - P)(\psi_{1,M} \cdot \varphi_k)| \geq L_{A_+, r_M, \alpha}^{(1)} \sqrt{\frac{\ln n}{n}} \right) \leq n^{-\alpha} . \tag{30}$$

**Proof of Lemma 6.** Let  $\beta > 0$ . By Bernstein's inequality, we get by straightforward computations that there exists  $L_{A_+, r_M, \beta} > 0$  such that, for all  $k \in \{1, \dots, D\}$ ,

$$\mathbb{P} \left( |(P_n - P)(\psi_{1,M} \cdot \varphi_k)| \geq L_{A_+, r_M, \beta} \sqrt{\frac{\ln n}{n}} \right) \leq n^{-\beta} .$$

Now the result follows from a simple union bound with  $\beta = \alpha + 1$ . ■

**Proof of Theorem 1.** Let  $C > 0$ . Recall that in (4) and (5) we set,

$$\mathcal{F}_C^\infty := \{s \in M ; \|s - s_M\|_\infty \leq C\}$$

and

$$\mathcal{F}_{>C}^\infty := \{s \in M ; \|s - s_M\|_\infty > C\} = M \setminus \mathcal{F}_C^\infty .$$

Take a localized orthonormal basis  $(\varphi_k)_{k=1}^D$  of  $(M, \|\cdot\|_2)$  satisfying **(Aslb)**. By Lemma 6 we get that there exists  $L_{A_+, r_M, \alpha}^{(1)} > 0$  such that, by setting

$$\Omega_1 = \left\{ \max_{k \in \{1, \dots, D\}} |(P_n - P)(\psi_{1,M} \cdot \varphi_k)| \leq L_{A_+, r_M, \alpha}^{(1)} \sqrt{\frac{\ln n}{n}} \right\} ,$$

we have for all  $n \geq n_0(A_+)$ ,  $\mathbb{P}(\Omega_1) \geq 1 - n^{-\alpha}$ . Moreover, we set

$$\Omega_2 = \bigcap_{i,j=1}^p \left\{ \max_{k \in \Pi_i} \left\{ \sup_{(\beta_l)_{l \in \Pi_{j,k}}; \sum_{l \in \Pi_{j,k}} \beta_l^2 \leq 1} \left| (P_n - P) \left( \varphi_k \cdot \sum_{l \in \Pi_{j,k}} \beta_l \varphi_l \right) \right| \right\} \leq L_\alpha^{(2)} \sqrt{\frac{\min\{A_i; A_j\} \ln n}{n}} \right\},$$

where  $L_\alpha^{(2)}$  is defined in Lemma 5. By Lemma 5, we have that for all  $n \geq n_0(A_+, r_M)$ ,  $\mathbb{P}(\Omega_2) \geq 1 - n^{-\alpha}$  and so, for all  $n \geq n_0(A_+, r_M)$ ,

$$\mathbb{P}(\Omega_1 \cap \Omega_2) \geq 1 - 2n^{-\alpha}. \quad (31)$$

We thus have for all  $n \geq n_0(A_+, r_M)$ ,

$$\begin{aligned} & \mathbb{P}(\|s_n - s_M\|_\infty > C) \\ & \leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_{>C}^\infty} P_n(Ks - Ks_M) \leq \inf_{s \in \mathcal{F}_C^\infty} P_n(Ks - Ks_M)\right) \\ & = \mathbb{P}\left(\sup_{s \in \mathcal{F}_{>C}^\infty} P_n(Ks_M - Ks) \geq \sup_{s \in \mathcal{F}_C^\infty} P_n(Ks_M - Ks)\right) \\ & \leq \mathbb{P}\left(\left\{\sup_{s \in \mathcal{F}_{>C}^\infty} P_n(Ks_M - Ks) \geq \sup_{s \in \mathcal{F}_{C/2}^\infty} P_n(Ks_M - Ks)\right\} \cap \Omega_1 \cap \Omega_2\right) + 2n^{-\alpha}. \end{aligned} \quad (32)$$

Now, for any  $s \in M$  such that

$$s - s_M = \sum_{k=1}^D \beta_k \varphi_k, \quad \beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D,$$

we have

$$\begin{aligned} & P_n(Ks_M - Ks) \\ & = (P_n - P)(\psi_{1,M} \cdot (s_M - s)) - (P_n - P)\left((s - s_M)^2\right) - P(Ks - Ks_M) \\ & = \sum_{k=1}^D \beta_k (P_n - P)(\psi_{1,M} \cdot \varphi_k) - \sum_{k,l=1}^D \beta_k \beta_l (P_n - P)(\varphi_k \cdot \varphi_l) - \sum_{k=1}^D \beta_k^2. \end{aligned}$$

We set for any  $(k, l) \in \{1, \dots, D\}^2$ ,

$$R_{n,k}^{(1)} = (P_n - P)(\psi_{1,M} \cdot \varphi_k) \quad \text{and} \quad R_{n,k,l}^{(2)} = (P_n - P)(\varphi_k \cdot \varphi_l).$$

Moreover, we set a function  $h_n$ , defined as follows,

$$h_n : \beta = (\beta_k)_{k=1}^D \mapsto \sum_{k=1}^D \beta_k R_k^{(1)} - \sum_{k,l=1}^D \beta_k \beta_l R_{k,l}^{(2)} - \sum_{k=1}^D \beta_k^2.$$

We thus have for any  $s \in M$  such that  $s - s_M = \sum_{k=1}^D \beta_k \varphi_k$ ,  $\beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D$ ,

$$P_n(Ks_M - Ks) = h_n(\beta). \quad (33)$$

In addition we set for any  $\beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D$ ,

$$|\beta|_{M,\infty} = \sum_{i=1}^p \sqrt{A_i} \max_{k \in \Pi_i} |\beta_k|. \quad (34)$$

It is straightforward to see that  $|\cdot|_{M,\infty}$  is a norm on  $\mathbb{R}^D$ . We also set for a real  $D \times D$  matrix  $B$ , its operator norm  $\|A\|_M$  associated to the norm  $|\cdot|_{M,\infty}$  on the  $D$ -dimensional vectors. More explicitly, we set for any  $A \in \mathbb{R}^{D \times D}$ ,

$$\|B\|_M := \sup_{\beta \in \mathbb{R}^D, \beta \neq 0} \frac{|B\beta|_{M,\infty}}{|\beta|_{M,\infty}}.$$

We have, for any  $A = (A_{k,l})_{k,l=1..D} \in \mathbb{R}^{D \times D}$ ,

$$\begin{aligned}
\|B\|_M &= \sup_{\beta \in \mathbb{R}^D, |\beta|_{M,\infty}=1} \left\{ \sum_{i=1}^p \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \sum_{l=1}^D B_{k,l} \beta_l \right\} \right\} \\
&= \sup_{\beta \in \mathbb{R}^D, |\beta|_{M,\infty}=1} \left\{ \sum_{i=1}^p \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \sum_{j=1}^p \sum_{l \in \Pi_j} B_{k,l} \beta_l \right\} \right\} \\
&= \sup_{\beta \in \mathbb{R}^D, |\beta|_{M,\infty}=1} \left\{ \sum_{i=1}^p \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \sum_{j=1}^p \sqrt{A_j} \max_{l \in \Pi_j} |\beta_l| \left( \sqrt{A_j^{-1}} \sum_{l \in \Pi_j} |B_{k,l}| \right) \right\} \right\} \\
&= \sum_{i=1}^p \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \max_{j \in \{1, \dots, p\}} \left\{ \sqrt{A_j^{-1}} \sum_{l \in \Pi_j} |B_{k,l}| \right\} \right\} . \tag{35}
\end{aligned}$$

Notice that by inequality (9) of **(Aslb)**, it holds

$$\mathcal{F}_{>C}^\infty \subset \left\{ s \in M ; s - s_M = \sum_{k=1}^D \beta_k \varphi_k \ \& \ |\beta|_{M,\infty} \geq C \right\} \tag{36}$$

and

$$\mathcal{F}_{C/2}^\infty \supset \left\{ s \in M ; s - s_M = \sum_{k=1}^D \beta_k \varphi_k \ \& \ |\beta|_{M,\infty} \leq C/2 \right\} . \tag{37}$$

Hence, from (32), (33) (36) and (37) we deduce that if we find on  $\Omega_1 \cap \Omega_2$  a value of  $C$  such that

$$\sup_{\beta \in \mathbb{R}^D, |\beta|_{M,\infty} \geq C} h_n(\beta) < \sup_{\beta \in \mathbb{R}^D, |\beta|_{M,\infty} \leq C/2} h_n(\beta) , \tag{38}$$

then inequality (12) follows and Theorem 1 is proved. Taking the partial derivatives of  $h_n$  with respect to the coordinates of its arguments, it then holds for any  $(k, l) \in \{1, \dots, D\}^2$  and  $\beta = (\beta_i)_{i=1}^D \in \mathbb{R}^D$ ,

$$\frac{\partial h_n}{\partial \beta_k}(\beta) = R_{n,k}^{(1)} - 2 \sum_{i=1}^D \beta_i R_{n,k,i}^{(2)} - 2\beta_k \tag{39}$$

We look now at the set of solutions  $\beta$  of the following system,

$$\frac{\partial h_n}{\partial \beta_k}(\beta) = 0 , \forall k \in \{1, \dots, D\} . \tag{40}$$

We define the  $D \times D$  matrix  $R_n^{(2)}$  to be

$$R_n^{(2)} := \left( R_{n,k,l}^{(2)} \right)_{k,l=1..D}$$

and by (39), the system given in (40) can be written

$$2 \left( I_D + R_n^{(2)} \right) \beta = R_n^{(1)} , \tag{S}$$

where  $R_n^{(1)}$  is a  $D$ -dimensional vector defined by

$$R_n^{(1)} = \left( R_{n,k}^{(1)} \right)_{k=1..D} .$$

Let us give an upper bound of the norm  $\|R_n^{(2)}\|_M$ , in order to show that the matrix  $I_D + R_n^{(2)}$  is nonsingular. On  $\Omega_2$  we have

$$\begin{aligned}
\|R_n^{(2)}\|_M &= \sum_{i=1}^p \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \max_{j \in \{1, \dots, p\}} \left\{ \sqrt{A_j^{-1}} \sum_{l \in \Pi_j} |R_{n,k,l}^{(2)}| \right\} \right\} \\
&= \sum_{i=1}^p \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \max_{j \in \{1, \dots, p\}} \left\{ \sqrt{A_j^{-1}} \sum_{l \in \Pi_{j,k}} |R_{n,k,l}^{(2)}| \right\} \right\} \\
&\leq \sum_{i=1}^p \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \max_{j \in \{1, \dots, p\}} \left\{ \sqrt{A_j^{-1}} |\Pi_{j,k}| \max_{l \in \Pi_j} |(P_n - P)(\varphi_k \cdot \varphi_l)| \right\} \right\} \\
&\leq A_c L_\alpha^{(2)} \sqrt{\frac{\ln n}{n}} \sum_{i=1}^p \max_{k \in \Pi_i} \left\{ \max_{j \in \{1, \dots, p\}} \left\{ \sqrt{\frac{A_i}{A_j}} \left( \frac{A_j}{A_i} \vee 1 \right) \sqrt{\min\{A_i; A_j\}} \right\} \right\} \quad (41)
\end{aligned}$$

We deduce from (8) and (41) that on  $\Omega_2$ ,

$$\|R_n^{(2)}\|_M \leq L_{A_c, \alpha} \cdot p \sqrt{\frac{A_p \ln n}{n}}. \quad (42)$$

Hence, from (42) and the fact that  $p^2 A_p \leq A_+ \frac{n}{(\ln n)^2}$ , we get that for all  $n \geq n_0(A_+, A_c, r_M, \alpha)$ , it holds on  $\Omega_2$ ,

$$\|R_n^{(2)}\|_M \leq \frac{1}{2},$$

the matrix  $(I_d + R_n^{(2)})$  is nonsingular, of inverse  $(I_d + R_n^{(2)})^{-1} = \sum_{u=0}^{+\infty} (-R_n^{(2)})^u$ , and so the system (S) admits a unique solution  $\beta^{(n)}$ , given by

$$\beta^{(n)} = \frac{1}{2} (I_d + R_n^{(2)})^{-1} R_n^{(1)}.$$

Now, on  $\Omega_1$  we have by (8),

$$\left| R_n^{(1)} \right|_{M, \infty} \leq \left( \sum_{i=1}^p \sqrt{A_i} \right) \max_{k \in \{1, \dots, D\}} |(P_n - P)(\psi_{1,M} \cdot \varphi_k)| \leq r_M L_{A_1, r_M, \alpha}^{(1)} \sqrt{\frac{D \ln n}{n}} \quad (43)$$

and we deduce that for all  $n \geq n_0(A_+, A_c, r_M, \alpha)$ , it holds on  $\Omega_2 \cap \Omega_1$ ,

$$\left| \beta^{(n)} \right|_{M, \infty} \leq \frac{1}{2} \left\| (I_d + R_n^{(2)})^{-1} \right\|_M \left| R_n^{(1)} \right|_{M, \infty} \leq r_M L_{A_1, r_M, \alpha}^{(1)} \sqrt{\frac{D \ln n}{n}}. \quad (44)$$

Moreover, by the formula (33) we have

$$h_n(\beta) = P_n(K s_M) - P_n \left( Y - \sum_{k=1}^D \beta_k \varphi_k \right)^2$$

and we thus see that  $h_n$  is concave. Hence, for all  $n \geq n_0(A_+, A_c, r_M, \alpha)$ , we get that on  $\Omega_2$ ,  $\beta^{(n)}$  is the unique maximum of  $h_n$  and on  $\Omega_2 \cap \Omega_1$ , by (44), concavity of  $h_n$  and unicity of  $\beta^{(n)}$ , we get

$$h_n(\beta^{(n)}) = \sup_{\beta \in \mathbb{R}^D, |\beta|_{M, \infty} \leq C/2} h_n(\beta) > \sup_{\beta \in \mathbb{R}^D, |\beta|_{M, \infty} \geq C} h_n(\beta),$$

which concludes the proof.  $\blacksquare$

## References

- [1] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.
- [2] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope Heuristics : Overview and Implementation. Technical Report 7223, INRIA, 2010.
- [3] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.
- [4] L. Birgé and P. Massart. Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.
- [5] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.
- [6] David L. Donoho. Asymptotic minimax risk for sup-norm loss: Solution via optimal recovery. *Probability Theory and Related Fields*, 99:145–170, 1994.
- [7] I.A. Ibragimov and R.Z. Khasminskii. Bounds for the risks of non-parametric regression estimates. *Theory Probab. Appl.*, 27:84–89, 1982.
- [8] I.A. Ibragimov and R.Z. Khasminskii. Exact asymptotically minimax estimator for nonparametric regression in uniform norm. *Theory Probab. Appl.*, 38:775–782, 1993.
- [9] R. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Annals of Probability*, 1:63–87 (electronic), 2005.
- [10] Matthieu Lerasle. Optimal model selection in density estimation, 2009. arXiv:0910.1654.
- [11] P. Massart. *Concentration Inequalities and Model Selection*. Springer-Verlag, 2007.
- [12] Adrien Saumard. Nonasymptotic quasi-optimality of aic and the slope heuristics in maximum likelihood estimation of density using histogram models, October 2010. hal-00512310, v1.
- [13] Adrien Saumard. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression, August 2010. hal-00512304, v1.
- [14] Adrien Saumard. The slope heuristics in heteroscedastic regression, August 2010. hal-00512306, v1.
- [15] C.J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.
- [16] Michel Talagrand. *The Generic Chaining*. Springer Monographs in Mathematics. Springer-Verlag, New York, 2005. Upper and lower bounds of stochastic processes.
- [17] A.B. Tsybakov. Pointwise and sup-norm sharp adaptive estimation of functions on the sobolev classes. *The Annals of Statistics*, 26(6):2420–2469, 1998.