Convergence in sup-norm of least-squares estimators in regression with random design and nonparametric heteroscedastic noise, and its application to optimal model selection

Adrien Saumard

Convergence in sup-norm of least-squares estimators in regression with random design and nonparametric heteroscedastic noise, and its application to optimal model selection

A. Saumard
Institut Télécom/Télécom ParisTech, CNRS UMR 5141

March 26, 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 ([3], [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 [14], [16], [13] and [15]). Indeed, such controls are one of the keystones to justify the Slope Heuristics, as claimed in [1]. In [14] (and also in [16]), 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. These constraints appear to be slightly more restrictive than the classical assumption of existence of an orthonormal localized basis of the model. Nevertheless, our approach allows to consider histograms and piecewise polynomials, but also, for example, some models of compactly supported wavelets, such as Haar expansions. Finally, our general result allows to stricly extend the previous theoretical justifications of the Slope Heuristics that have been achieved in the heteroscedastic regression framework.

Keywords: Least-squares estimators, sup-norm, finite-dimensional models, localized basis, model selection, slope heuristics.

1 Introduction

Let P be the unknown law of independent and identically distributed data $(X_1, ..., X_n)$. The sup-norm, i.e. the norm that classically defines 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 $(X_1, ..., X_n)$ 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 [6] and Klein-Rio [10], 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 symmetrized version, the Rademacher process, see Sections 2.6, 2.7 and 4 of [18]. Sharp controls, from upper and from below, of the first moment of an empirical process indexed by functions, appeared to be an essential step in recent 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 [14], [16], [13]. Before describing some

existing results on convergence in sup-norm of some estimators in regression settings 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 largely 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 [11] recovered Arlot and Massart propositions in least-squares density estimation. More recently, these results have been generalized in papers [14], [16] and [13] - see also [15] - 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 to achieve optimal upper and lower bounds of the excess risks, 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, concerning the regression framework, 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 Ibragimov and Khasminskii [8] and Stone [17]. Korotselev [9] moreover found the exact asymptotic constant in the minimax problem, considering the estimation of a β-Lipschitz regression function. This discovery has been then extended by Donoho [7], using a beautiful method inspired by optimal recovery techniques, considering the estimation of a β-Lipschitz function in Gaussian white noise setting. This latter framework is closely related to nonparametric regression with fixed design and homoscedastic Gaussian noise. Donoho also claims that "the subject area is appealing because L_∞-loss has special importance in connection with setting fixed-width simultaneous confidence bands for an unknown regression". This connection 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 equivalent - for the L_2-loss in certain finite dimensional models. We also highlight the work of Tsybakov [19], 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 adaptive estimator - in regard with the regularity parameter of the Sobolev classes - can be constructed using developments in Fourier basis combined with 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 [19], 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 argued before, this setting 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_∞-loss of the least-squares estimator, and also for more general M-estimators. We then state in Section 4 our main result, where we derive the consistency in sup-norm of the least-squares estimator at the rate $\sqrt{D \ln n/n}$, when the model

is fulfilled with an orthonormal basis in $L_2(P)$ satisfying a criterion which is a little more restrictive than the assumption of localized basis (see Section 7.4 of Massart [12]). We call these bases the "strongly localized bases". We give three explicit examples of such bases, namely histogram bases, piecewise polynomial bases and Haar bases, this latter example being the simplest case of compactly supported wavelet basis. We then conclude our paper in Section 5, by returning back to our "initial" problem of optimal model selection. By using our new general result on the convergence in sup-norm of least-squares estimator, we strictly extend the results of the previous studies of the Slope Heuristics in the heteroscedastic and random design regression framework. Indeed, we show that the Slope Heuristics are actually satisfied when considering the selection of linear models that contain strongly localized orthonormal bases. 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 $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, $i \in \{1, ..., n\}$ 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 $s_* \in L_2\left(P^X\right)$, ε_i are i.i.d. random variables with mean 0 and variance 1 conditionally to X_i and $\sigma : \mathcal{X} \longrightarrow \mathbb{R}$ is an heteroscedastic noise level. A generic random variable of law P, independent of $(\xi_1, ..., \xi_n)$, 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 $L^2\left(P^X\right)$ 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

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

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

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

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 $L_2\left(P^X\right)$. Recall that the excess risk $\ell(s_*, s)$ of a function $s \in M$ is then given by

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

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

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

Hence, the excess risk on M $\ell(s_M, s)$ 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

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

When considering the convergence in sup-norm in Sections 3 and 4, we will not discuss here 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 $P(Ks_n - Ks_M)$. However, this model selection task is tackled in Section 5 below, when we establish the validity of the Slope Heuristics in a general framework.

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

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

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

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

and

$$\mathcal{F}_{>C}^\infty := \{s \in M \ ; \|s - s_M\|_\infty > C\} = M \backslash \mathcal{F}_C^\infty \ . \tag{5}$$

Notice that since M is a linear vector space, $\mathcal{F}_C^\infty$ is the closed L_∞-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

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

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

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

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 state here our 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 (**Aslb**) 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 localized basis".

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

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

and

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

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

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

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

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

It is directly seen that a strongly localized basis is a localized basis in the sense of Birgé and Massart [4] - and it is worth noting that the latter notion has been introduced by these authors in [4] 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

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

Now, (8) and (9) imply (11). Moreover, we require in (10) a control of the number of intersections between the supports of the elements 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 and are strongly localized in $L_2\left(P^X\right)$ under suitable 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 (**Aslb**) and use notations of (**Aslb**). Assume moreover that the following assumption holds:*

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

If there exists $A_+ > 0$ such that

$$\max\left\{D, p^2 A_p\right\} \leq A_+ \frac{n}{(\ln n)^2} \ , \tag{12}$$

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

$$\mathbb{P}\left(\|s_n - s_M\|_\infty \geq L_{A_{1,M}, r_M, \alpha} \sqrt{\frac{D \ln n}{n}} \right) \leq n^{-\alpha} \ . \tag{13}$$

Let us briefly comment on Theorem 1. We derive here a probability upper bound for the L_∞-loss of the least-square estimator of a regression function towards the orthonormal projection in a finite-dimensional linear model fulfilled with a strongly localized basis. As in [14], the polynomial probability bounds are obtained at the rate $\sqrt{D \ln n / n}$. Assumption (**Ab**) restricts our study to a bounded setting, a restriction also made in related works [1], [14] and [16]. Moreover, in (12), we ask in particular for a dimension D of the model not greater than the number of data, within the square of the logarithm of this number. Such a restriction in the dimension of the model is also classical for model selection purposes. We also need via (12),

$$p^2 A_p \leq A_+ \frac{n}{(\ln n)^2} \tag{14}$$

and we notice that quantity $p^2 A_p$ is closely related to D as by (**Aslb**), we have

$$A_p \leq \left(\sum_{i=1}^{p} \sqrt{A_i} \right)^2 \leq p^2 A_p \quad \text{and} \quad \left(\sum_{i=1}^{p} \sqrt{A_i} \right)^2 \leq r_M^2 D \ .$$

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 [14].

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 $c_{M,P}$,*

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} P^X(I)} \geq c_{M,P} > 0 \ . \tag{15}$$

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

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

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

$$\text{for all } \beta = (\beta_I)_{I \in \mathcal{P}} \in \mathbb{R}^D, \quad \left\| \sum_{I \in \mathcal{P}} \beta_I \varphi_I \right\|_\infty \leq c_{M,P}^{-1} \sqrt{D} \, |\beta|_\infty \ . \tag{16}$$

By Lemma 2, we deduce that if the partition $\mathcal{P}$ satisfies the assumption of lower regularity given in (15) then inequality (9) is satisfied for M, with $p = 1$ and $r_M = c_{M,P}^{-1} > 0$. Moreover, notice that for all $(i,j) \in \{1, ..., D\}^2$,

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

and in this case (**Aslb**) is straightforwardly satisfied.
Hence, from Theorem 1 and Lemma 2, we deduce the following result, describing the convergence in sup-norm in the case of histograms defined on a partition which is lower-regular with respect to the unknown probability P^X.

Corollary 3 *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 $c_{M,P}$,*

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

and that there exists a constant $A > 0$ such that $|Y| \leq A$ a.s.
If there exists $A_+ > 0$ such that

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

then we have, for all $n \geq n_0 (A_+, c_{M,P}, \alpha)$,

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

In Corollary 3, assumption (**Ab**) of Theorem 4.2 is replaced by a softer and also more tractable control :$|Y| \leq A$ *a.s.* The equivalence between the two statements in the case of histograms (with $A_{1,M} = 2A$), comes from the following relations,

$$\|Y\|_\infty \geq \|s_* = \mathbb{E}[Y | X = \cdot]\|_\infty \geq \|s_M\|_\infty \ ,$$

the first inequality being a general fact in L_2 and the second being valid in the histogram case but not in general. Moreover, by Corollary 3, we recover exactly Lemma 6 of [14], where the reader will find a direct proof, only valid in the case of histograms.

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

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

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 is given in [14] - where it is called Lemma 8 - and states the existence, under suitable assumptions, of a localized orthonormal basis in $(M, \|\cdot\|_2)$. Its proof, which is not totally trivial as it requires arguments from the theory of orthogonal polynomials, can be found in [14].

Lemma 4 *Let* 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* 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 $\left\{ \varphi_{I,j}, \ I \in \mathcal{P}, \ j \in \{0, ..., r\} \right\}$ of $(M, \|\cdot\|_2)$ such that,

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

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

$$\max_{j \in \{0,...,r\}} \left\| \varphi_{I,j} \right\|_\infty \leq L_{r,c_{\min}} \frac{1}{\sqrt{\text{Leb}(I)}} \ . \tag{17}$$

As a consequence, if it holds

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \text{Leb}(I)} \geq c_{M,\text{Leb}} > 0 \ , \tag{18}$$

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 $\beta = \left(\beta_{I,j}\right)_{I \in \mathcal{P}, j \in \{0,...,r\}} \in \mathbb{R}^D$,

$$\left\| \sum_{I,j} \beta_{I,j} \varphi_{I,j} \right\|_\infty \leq L_{r,c_{\min},c_{M,\text{Leb}}} \sqrt{D} \, |\beta|_\infty \ , \tag{19}$$

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

Lemma 4 states that if $\mathcal{X} = [0,1]$ is the unit interval and P^X has a density with respect to the Lebesgue measure Leb on $\mathcal{X}$ uniformly bounded away form 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 (17). Moreover, if we assume the lower regularity of the partition with respect to Leb then the orthonormal basis is localized, where the constant of localization in (19) 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, under assumptions of Lemma 4, the property of strongly localized basis is satisfied, by the same type of arguments than for histogram models ($p = 1$ and $A_c = r + 1$ are convenient).
From Theorem 1 and Lemma 4, we now state the following straightforward corollary concerning convergence in sup-norm of the least-squares estimator in the case of piecewise polynomials.

Corollary 5 *Let $A_+, c_{\min}, c_{M,\mathrm{Leb}} > 0$ some positive constants. Let* 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* Leb*, satisfying*

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

Assume moreover that

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \mathrm{Leb}(I)} \geq c_{M,\mathrm{Leb}} > 0 \ ,$$

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

*and that (**Ab**) holds, that is*

$$\textit{there exists a constant } A_{1,M} > 0 \textit{ such that } \left|\psi_{1,M}(X,Y)\right| \leq A_{1,M} \textit{ a.s.}$$

Then it holds, for all $n \geq n_0(A_+, r, c_{M,\mathrm{Leb}}, c_{\min}, \alpha)$,

$$\mathbb{P}\left(\|s_n - s_M\|_\infty \geq L_{A_{1,M}, r, c_{M,\mathrm{Leb}}, c_{\min}, \alpha} \sqrt{\frac{D \ln n}{n}} \right) \leq n^{-\alpha} \ .$$

It is worth mentioning that Corollary 5 slightly improves Lemma 9 of [14], because the conclusions obtained in these two propositions are formally the same, but we avoid in Corollary 5 the use of a condition required in Lemma 9 of [14] and which states that the density f is uniformly upper bounded on $\mathcal{X}$. However, condition (**Ab**) used in Corollary 5 is slightly stronger than the corresponding one used in Lemma 9 of [14], where we only demand that there exists a constant $A > 0$ such that $|Y| \leq A$. Indeed, to deduce (**Ab**) from the latter assumption on the response variable, it suffices to show that the projection s_M is also uniformly upper bounded : $\|s_M\|_\infty \leq A$ for some $A > 0$. On contrary to the histogram case, this not guaranteed, but it is easily seen that such a property is satisfied as soon as we have

$$\max_{I \in \mathcal{P}} \left\{ \frac{P(I)}{\mathrm{Leb}(I)} \right\} \leq \tilde{A} \text{ for some } \tilde{A} > 0.$$

In particular, the latter control is achieved if the density f is uniformly upper bounded on $\mathcal{X}$ and in this case we recover exactly Lemma 9 of [14]. The reader interested can check the proof of Lemma 9 given [14], and see that our present approach, described in Section 3 and given precisely in Section 6, has two major advantages compared to it. First, our approach is here systematical and avoid the use of an explicit formula for the least-squares estimator s_n, whereas in [14] it heavily relies on an explicit expression of the considered estimator ; and secondly, despite the fact that we are more general in the present paper, we gain in simplicity and our proof of Theorem 1 is substantially condensed compared to the one proposed in [14] concerning piecewise polynomials.

4.3.3 Haar expansions

Let $\mathcal{X} = [0,1]$, $m \in \mathbb{N}$. We consider now some Haar expansions on $\mathcal{X}$, which are special cases of wavelet expansions with compact supports.

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

$$\Lambda(j) = \left\{(j,k) \ ; \ 1 \leq k \leq 2^j\right\} \ , \tag{20}$$

$$\Lambda(j,i,l) = \left\{(j,k) \ ; \ 2^{j-i}(l-1)+1 \leq k \leq 2^{j-i}l\right\} \ . \tag{21}$$

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 $i \leq j$, $\left\{\Lambda(j,i,l) \ ; \ 1 \leq l \leq 2^i\right\}$ is a partition of $\Lambda(j)$, which means that

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

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

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

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

$$M = \mathrm{Span}\left\{\rho_\lambda \ ; \ \lambda \in \Lambda_m\right\} \ . \tag{22}$$

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

Lemma 6 *Let $m \in \mathbb{N}$. Assume that $\mathcal{X} = [0,1]$ and let M be the model of dimension D given by (22). Then*

$$D = \mathrm{Card}(\Lambda_m) = 2^{m+1} \ . \tag{23}$$

Set for every integers $j \geq 0$, $1 \leq k \leq 2^j$,

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

$$\varphi_{j,k} : x \in [0,1] \mapsto \frac{1}{\sqrt{p_{j,k,+}^2 p_{j,k,-} + p_{j,k,-}^2 p_{j,k,+}}}\left(p_{j,k,+}\mathbf{1}_{[2^{-j}(k-1), 2^{-j}(k-1/2)]} - p_{j,k,-}\mathbf{1}_{\left(2^{-j}\left(k-\frac{1}{2}\right), 2^{-j}k\right]}\right) \ . \tag{24}$$

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

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

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

$$\left\|\varphi_{j,k}\right\|_\infty \leq \sqrt{\frac{2}{c_{\min}}}2^{j/2} \ . \tag{25}$$

Moreover, by setting $A_{-1} = 1$ and $A_j = 2^j$, $j \geq 0$, we have

$$\sum_{j=-1}^{m} \sqrt{A_j} \leq \left(\sqrt{2}+1\right)\sqrt{D} \tag{26}$$

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

$$\left\|\sum_{\lambda \in \Lambda_m} \beta_\lambda \varphi_\lambda\right\|_\infty \leq \sqrt{\frac{2}{c_{\min}}} \sum_{j=-1}^{m} \sqrt{A_j} \max_{k \in \Lambda_j} |\beta_k| \ . \tag{27}$$

Finally, if $\Lambda_{j,\mu} = \left\{\lambda \in \Lambda_j ; \mathit{Support}\left(\varphi_\mu\right) \bigcap \mathit{Support}\left(\varphi_\lambda\right) \neq \emptyset\right\}$ for $\mu \in \Lambda_m$ and $j \in \{-1, 0, 1, ..., m\}$,

$$\max_{\mu \in \Lambda_i} \mathrm{Card}\left(\Lambda_{j,\mu}\right) \leq A_j A_i^{-1} \vee 1 \ . \tag{28}$$

By Lemma 6, which proof is left to the reader, 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 (22) admits a strongly localized orthonormal basis for the $L_2\left(P^X\right)$-norm. More precisely, with notations of (**Aslb**), $r_M = \max\left\{\sqrt{2}+1, \sqrt{2c_{\min}^{-1}}\right\}$ and $A_c = 1$ are convenient.
We particularize now Theorem 1 to models of Haar expansions by using Lemma 6.

Corollary 7 *Let $m \in \mathbb{N}$, $c_{\min}, A_+ > 0$. Assume that $\mathcal{X} = [0,1]$ and let M be the model of dimension D given by (22). Then*

$$D = \mathrm{Card}(\Lambda_m) = 2^{m+1} \ .$$

Assume that

$$m^2 \cdot 2^m \leq A_+ \frac{n}{(\ln n)^2} \tag{29}$$

and that P^X has a density f with respect to Leb *on $[0,1]$ such that for all $x \in [0,1]$,*

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

*Also assume that (**Ab**) holds, that is*

$$\textit{there exists a constant } A_{1,M} > 0 \textit{ such that } \left|\psi_{1,M}(X,Y)\right| \leq A_{1,M} \textit{ a.s.}$$

Then it holds, for all $n \geq n_0(A_+, c_{\min}, \alpha)$,

$$\mathbb{P}\left(\|s_n - s_M\|_\infty \geq L_{A_{1,M}, c_{\min}, \alpha}\sqrt{\frac{D \ln n}{n}}\right) \leq n^{-\alpha} \ .$$

Comparing to Corollaries 3 and 5, condition (29) is stronger than the condition $D \leq A_+ n (\ln n)^{-2}$. Indeed, we have to take into account the multiscale structure of Haar expansions, whereas in histograms or piecewise polynomials cases, the scale is unique (understand that $p = 1$ in (**Aslb**)). Corollary 7 also shows that the scope of Theorem 1 is strictly wider that results concerning convergence in sup-norm if least-squares estimators of a regression function proposed in [14].

5 Application to optimal model selection and the slope heuristics

As emphasized in the introduction of the present paper, the original motivation of our study of convergence in sup-norm of least-squares estimators comes from our recent analysis in ([14], [16], [13], [15]) of the so-called Slope Heuristics first discovered by Birgé and Massart ([3]) and then extended by Arlot and Massart ([1]). Indeed, the Slope Heuristics aim at performing an efficient model selection procedure, and in order to prove oracle inequalities with constants almost one and tending to one when the number of data tends to infinity, we need to derive in particular upper and lower bounds for the true excess risk and the empirical excess risk on a fixed model, that are optimal at the first order ([16]). To achieve this goal, we noticed in [14] and also in [13] and [15] that proving the convergence in sup-norm of the considered M-estimators was an essential step of the new methodology of proof that we have developed in order to generalize the precedent investigation of Arlot and Massart ([1]). Indeed, considering the case of heteroscedastic regression, we have been able in [14] and [16] to recover the results that Arlot and Massart obtained in [1] by considering linear histogram models, and we have extended it to models of piecewise polynomials. Moreover, we give in [14] and [16] more general statements that ensure the desired results as soon as the considered (general) linear models contain orthonormal localized bases and that the least-squares estimators are consistent in sup-norm towards the projections of the regression function onto the models.

Hence, we turn in the following to the application of the general result obtained in Section 4.2 to this context of model selection and also accurate bounds of excess risks on a fixed model. It is worth noting that condition (**Aslb**) ensures, *via* Theorem 1, that models of dimension not too heavy satisfy the condition of consistency in sup-norm of the least-squares estimators (assumption (**H5**) of ([14]) and assumption ($\mathbf{Ac}_\infty$)

of ([16])). Moreover, as noticed in Section 4.1, condition (**Aslb**) is stronger than the classical assumption of existence of an orthonormal localized basis - even if it seems that all the classical examples of localized basis do fulfill condition (**Aslb**) - and so it also ensures that corresponding assumptions (**H4**) of ([14]) and (**Alb**) of ([16]) are also satisfied. The application of Theorem 3 and 4 of ([14]), in which we derive sharp probability bounds for the excess risks on a fixed model, and Theorems 1 and 2 of ([16]) describing the validity of the Slope Heuristics in heteroscedastic regression, are then straightforward and detailed in the following.

5.1 Fixed model case

In this section, the model M is fixed and we establish sharp bounds for the true and empirical excess risks of the least-squares estimator on M, as addressed in ([14]). We thus apply Theorem 1 of the present paper to Theorem 3 of ([14]). This gives the following result.

Theorem 8 *Let $A_+, A_-, \alpha > 0$ and let M be a linear model of finite dimension D. Assume that there exists an orthonormal basis $\varphi = (\varphi_k)_{k=1}^D$ satisfying (**Aslb**) and assume that two positive constants $A, \sigma_{\min} > 0$ exist such that*

$$|Y_i| \leq A \text{ a.s. }, \tag{30}$$

$$\|s_M\|_\infty \leq A \ , \tag{31}$$

and

$$0 < \sigma_{\min} \leq \sigma(X) \quad a.s. \ . \tag{32}$$

If it moreover holds

$$A_- (\ln n)^2 \leq D \leq \max\left\{D, p^2 A_p\right\} \leq A_+ \frac{n}{(\ln n)^2} \ , \tag{33}$$

*then a positive finite constant A_0 exists, only depending on α, A_- and on the constants $A, \sigma_{\min}$ and the constant r_M defined in assumption (**Aslb**), such that by setting*

$$\varepsilon_n = A_0 \max\left\{\left(\frac{\ln n}{D}\right)^{1/4}, \ \left(\frac{D \ln n}{n}\right)^{1/4}\right\} \ , \tag{34}$$

we have for all $n \geq n_0(A_-, A_+, r_M, \sigma_{\min}, \alpha)$,

$$\mathbb{P}\left[P(Ks_n - Ks_M) \geq (1-\varepsilon_n)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \geq 1 - 5n^{-\alpha} \ , \tag{35}$$

$$\mathbb{P}\left[P(Ks_n - Ks_M) \leq (1+\varepsilon_n)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \geq 1 - 5n^{-\alpha} \ , \tag{36}$$

$$\mathbb{P}\left[P_n(Ks_M - Ks_n) \geq \left(1-\varepsilon_n^2\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \geq 1 - 2n^{-\alpha} \ , \tag{37}$$

$$\mathbb{P}\left[P_n(Ks_M - Ks_n) \leq \left(1+\varepsilon_n^2\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \geq 1 - 3n^{-\alpha} \ , \tag{38}$$

where $\mathcal{K}_{1,M}^2 = \frac{1}{D}\sum_{k=1}^D \mathrm{Var}\left(\psi_{1,M} \cdot \varphi_k\right)$.

In Theorem 8 above, we achieve sharp upper and lower bounds for the true and empirical excess risks on M. Note that Theorem 1 of the present paper apply under assumptions of Theorem 8, since assumptions (30) and (31) allow to recover assumption (**Ab**) of Theorem 1 with $A_{1,M} = 4A$. The excess risks bound given in Theorem 8 are optimal at the first order since the leading constants are equal for upper and lower bounds. Moreover, Theorem 8 states the equivalence with high probability of the true and empirical excess risks for models of reasonable dimensions. This equivalence is actually the keystone of the Slope Heuristics, which is furthermore discussed in Section 5.2 below. Theorem 8 gives a strict generalization of Theorems 7 and 10 of [14], in which the author derived the same kind of results for histograms and piecewise polynomials respectively. Indeed, as discussed in Section 4.3 above, the conditions of Theorem 8 contains nonetheless the

cases of histograms and piecewise polynomials, but also the example of Haar expansions on the unit interval, for which no such controls of the excess risks were derived yet.
As in [14], we turn now to upper bounds in probability for the true and empirical excess risks on models with possibly small dimensions. In this context, we do not achieve sharp or explicit constants in the rates of convergence. The following result is needed in the proof of results stated in Section 5.2 below (see also [16]) and is a straightforward extension of Theorem 4 of [14], using Theorem 1 of the present paper.

Theorem 9 *Let $\alpha, A_+ > 0$ be fixed and let M be a linear model of finite dimension D. Assume that (**Aslb**) hold and that a positive constant $A > 0$ exists such that*

$$|Y_i| \leq A \text{ a.s. } , \tag{39}$$

$$\|s_M\|_\infty \leq A \ . \tag{40}$$

If it holds

$$1 \leq \max\left\{D, p^2 A_p\right\} \leq A_+ \frac{n}{(\ln n)^2} \ ,$$

then a positive constant A_u exists, only depending on A, r_M and α, such that for all $n \geq n_0(A, r_M)$,

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) \geq A_u \frac{D \vee \ln n}{n}\right] \leq 3n^{-\alpha} \tag{41}$$

and

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\right) \geq A_u \frac{D \vee \ln n}{n}\right] \leq 3n^{-\alpha} \ . \tag{42}$$

Notice that on contrary to the situation of Theorem 8, we do not assume that the noise is uniformly bounded away from zero over the space $\mathcal{X}$ (see (32) above). This assumption allows in Theorem 8 to derive lower bounds for the true and empirical excess risks, as well as to achieve sharp constants in the deviation bounds for models of reasonable dimensions. In Theorem 9, we just derive upper bounds and assumption (32) is not needed. The price to pay is that constants in the rates of convergence derived in (41) and (42) are possibly larger than the corresponding ones of Theorem 8, but our results still hold true for small models. Moreover, in the case of models with reasonable dimensions, that is dimensions satisfying assumption (33) of Theorem 8, the rate of decay is preserved compared to Theorem 8 and is proportional to D/n.

5.2 Slope Heuristics

We extend here the results obtained in ([16]), from the cases of histogram and piecewise polynomials to the case where assumption (**Aslb**) - that contains for instance the example of Haar expansions - is satisfied. The adaptations of Theorem 1 and 2 of ([16]) are again straightforward.

5.2.1 Model selection framework and the Slope Heuristics

Until the end of the present Section 5.2, we are given a finite collection of models $\mathcal{M}_n$, among which we aim to choose the best model in terms of prediction. This best model is called an oracle model M_* and is defined to be

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \left\{l\left(s_*, s_n\left(M\right)\right)\right\} \ . \tag{43}$$

The associated oracle estimator $s_n(M_*)$ thus achieves the best performance in terms of excess risk among the collection $\{s_n(M)\,; M \in \mathcal{M}_n\}$. Unfortunately, the oracle model is unknown as it depends on the unknown law P of the data, and we consider the case where we estimate it by a model selection procedure via penalization. Given some known penalty pen, that is a function from $\mathcal{M}_n$ to $\mathbb{R}_+$, we thus consider the following data-dependent model, also called selected model,

$$\widehat{M} \in \arg\min_{M \in \mathcal{M}_n} \left\{P_n\left(Ks_n\left(M\right)\right) + \text{pen}\left(M\right)\right\} \ . \tag{44}$$

Our goal is then to find a good penalty, such that the selected model $\widehat{M}$ satisfies an oracle inequality of the form

$$l\left(s_*, s_n\left(\widehat{M}\right)\right) \leq C \times \ell\left(s_*, s_n\left(M_*\right)\right) \ ,$$

with some positive constant C as close to one as possible and with high probability, typically more than $1 - Ln^{-2}$ for some positive constant L.

In order to introduce the slope phenomenon, let us rewrite the definition of the oracle model M_* given in (43). As for any $M \in \mathcal{M}_n$, the excess risk $l\left(s_*, s_n\left(M\right)\right) = P\left(Ks_n\left(M\right)\right) - P\left(Ks_*\right)$ is the difference between the risk of the estimator $s_n\left(M\right)$ and the risk of the target s_*, and as $P\left(Ks_*\right)$ is a constant of the problem, it holds

$$\begin{aligned} M_* &\in \arg\min_{M \in \mathcal{M}_n} \left\{P\left(Ks_n\left(M\right)\right)\right\} \\ &= \arg\min_{M \in \mathcal{M}_n} \left\{P_n\left(Ks_n\left(M\right)\right) + \mathrm{pen}_{\mathrm{id}}\left(M\right)\right\} \end{aligned}$$

where for all $M \in \mathcal{M}_n$,

$$\mathrm{pen}_{\mathrm{id}}\left(M\right) := P\left(Ks_n\left(M\right)\right) - P_n\left(Ks_n\left(M\right)\right) \ .$$

The penalty function $\mathrm{pen}_{\mathrm{id}}$ is called the *ideal penalty*, as it allows to select the oracle, but it is unknown because it depends on the distribution of the data. As pointed out by Arlot and Massart [1], the leading idea of penalization in the efficiency problem is thus to give some sharp estimate of the ideal penalty, in order to perform an unbiased or asymptotically unbiased estimation of the risk over the collection of models, leading to a sharp oracle inequality for the selected model. A penalty term $\mathrm{pen}_{\mathrm{opt}}$ is said to be optimal if it achieves an oracle inequality with constant almost one, tending to one when the number n of data tends to infinity.
Concerning the estimation of the optimal penalty, Arlot and Massart [1] conjecture that the mean of the empirical excess risk $\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$ satisfies the following slope heuristics in a quite general framework:

(i) If a penalty $\mathrm{pen} : \mathcal{M}_n \longrightarrow \mathbb{R}_+$ is such that, for all model $M \in \mathcal{M}_n$,

$$\mathrm{pen}\left(M\right) \leq \left(1 - \delta\right)\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$$

with $\delta > 0$, then the dimension of the selected model $\widehat{M}$ is "very large" and the excess risk of the selected estimator $s_n\left(\widehat{M}\right)$ is "much larger" than the excess risk of the oracle.

(ii) If $\mathrm{pen} \approx \left(1 + \delta\right)\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$ with $\delta > 0$, then the corresponding model selection procedure satisfies an oracle inequality with a leading constant $C\left(\delta\right) < +\infty$ and the dimension of the selected model is "not too large". Moreover,

$$\mathrm{pen}_{\mathrm{opt}} \approx 2\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$$

is an optimal penalty.

The mean of the empirical excess risk on M, when M varies in $\mathcal{M}_n$, is thus conjectured to be the maximal value of penalty under which the model selection procedure totally misbehaves. It is called the *minimal penalty,* denoted by $\mathrm{pen}_{\mathrm{min}}$:

$$\text{for all } M \in \mathcal{M}_n, \qquad \mathrm{pen}_{\mathrm{min}}\left(M\right) = \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \ .$$

The optimal penalty is then close to two times the minimal one,

$$\mathrm{pen}_{\mathrm{opt}} \approx 2\,\mathrm{pen}_{\mathrm{min}} \ .$$

Let us now briefly explain why points (**i**) and (**ii**) below are natural. We give in Section 5.2.2 below precise results which validate the slope heuristics for models satisfying (**Aslb**) and which are, to our knowledge, the

more general results in the regression framework concerning the validation of the Slope Heuristics. If the penalty is the minimal one, then for all $M \in \mathcal{M}_n$,

$$\begin{aligned}
&P_n\left(Ks_n\left(M\right)\right) + \mathrm{pen}_{\min}\left(M\right) \\
&= P_n\left(Ks_n\left(M\right)\right) + \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \\
&= P\left(Ks_M\right) + \left(P_n - P\right)\left(Ks_M\right) + \left(\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] - P_n\left(Ks_M - Ks_n\left(M\right)\right)\right) \\
&\approx P\left(Ks_M\right) \ .
\end{aligned}$$

In the above lines, we neglect $\left(P_n - P\right)\left(Ks_M\right)$ as it is a centered quantity and if the empirical excess risk $P_n\left(Ks_n\left(M\right) - Ks_M\right)$ is close enough to its expectation, then the selected model almost minimizes its bias, and so its dimension is among the largest of the models and the excess risk of the selected estimator blows up. As shown by Boucheron and Massart [5], the empirical excess risk satisfies a concentration inequality in a general framework, which allows to neglect the difference with its mean, at least for models that are not too small.

Now, if the chosen penalty is less than the minimal one, $\mathrm{pen} \approx (1-\delta)\,\mathrm{pen}_{\min}$ with $\delta \in (0,1)$, the algorithm minimizes over $\mathcal{M}_n$,

$$\begin{aligned}
&P_n\left(Ks_n\left(M\right)\right) + \mathrm{pen}\left(M\right) \\
&\approx P\left(Ks_M\right) - \delta P_n\left(Ks_M - Ks_n\left(M\right)\right) + \left(P_n - P\right)\left(Ks_M\right) \\
&\quad + \left(1-\delta\right)\left(\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] - P_n\left(Ks_M - Ks_n\left(M\right)\right)\right) \\
&\approx P\left(Ks_M\right) - \delta P_n\left(Ks_M - Ks_n\left(M\right)\right) \ ,
\end{aligned}$$

where in the last identity we neglect the deviations of the empirical excess risk and the difference between the empirical and true risk of the projections s_M. As the empirical excess risk is increasing and the risk of the projection s_M is decreasing with respect to the complexity of the models, the penalized criterion is decreasing with respect to the complexity of the models, and the selected model is again among the largest of the collection.

If on the contrary, the chosen penalty is more than the minimal one, $\mathrm{pen} \approx (1+\delta)\,\mathrm{pen}_{\min}$ with $\delta > 0$, then the selected model minimizes the following criterion, for all $M \in \mathcal{M}_n$,

$$\begin{aligned}
&P_n\left(Ks_n\left(M\right)\right) + \mathrm{pen}\left(M\right) - P_n\left(Ks_*\right) \\
&\approx \ell\left(s_*, s_M\right) + \delta P_n\left(Ks_M - Ks_n\left(M\right)\right) + \left(P_n - P\right)\left(Ks_M - Ks_*\right) \\
&\quad + \left(1+\delta\right)\left(\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] - P_n\left(Ks_M - Ks_n\left(M\right)\right)\right) \\
&\approx \ell\left(s_*, s_M\right) + \delta P_n\left(Ks_M - Ks_n\left(M\right)\right) \ ,
\end{aligned} \tag{45}$$

So the selected model achieves a trade-off between the bias of the models which decreases with the complexity and the empirical excess risk which increases with the complexity of the models. The selected dimension will be then reasonable, and the trade-off between the bias and the complexity of the models is likely to give some oracle inequality.

Finally, if we take $\delta = 1$ in the above case, that is $\mathrm{pen} \approx 2 \times \mathrm{pen}_{\min}$ and if we assume that the empirical excess risk is equivalent to the excess risk,

$$P_n\left(Ks_M - Ks_n\left(M\right)\right) \sim P\left(Ks_n\left(M\right) - Ks_M\right) \ , \tag{46}$$

then according to (45) the selected model almost minimizes

$$P\left(Ks_M - Ks_*\right) + P_n\left(Ks_M - Ks_n\left(M\right)\right) \approx \ell\left(s_*, s_M\right) + P\left(Ks_n\left(M\right) - Ks_M\right) \approx \ell\left(s_*, s_n\left(M\right)\right) \ .$$

Hence,

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \approx \ell\left(s_*, s_n\left(M_*\right)\right)$$

and the procedure is nearly optimal. Note that Theorem 8 of Section 5.1 actually shows that (46) is satisfied, under the right assumptions, for models satisfying (**Aslb**).

5.2.2 Results

Let us first state the following set of assumptions that will be called in Theorems 10 and 11 below.

Set of assumptions : (SA)

(P1) Polynomial complexity of $\mathcal{M}_n$: $\mathrm{Card}\left(\mathcal{M}_n\right) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(Auslb) Existence of a "strongly localized basis" in the sense of the definition given in Section 4.1 above, for each model $M \in \mathcal{M}_n$ and with uniform constants: there exist $r_{\mathcal{M}}, A_c > 0$ such that for every $M \in \mathcal{M}_n$, there exist $p_M \in \mathbb{N}_*$, a partition $(\Pi_i)_{i=1}^{p_M}$ of $\{1, ..., D_M\}$, positive constants $(A_i)_{i=1}^{p_M}$ and an orthonormal basis $(\varphi_k)_{k=1}^{D}$ of $(M, \|\cdot\|_2)$ such that $0 < A_1 \leq A_2 \leq ... \leq A_{p_M} < +\infty$,

$$\sum_{i=1}^{p_M} \sqrt{A_i} \leq r_{\mathcal{M}} \sqrt{D_M} \ ,$$

and

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

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

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

and we assume that for all $j \in \{1, ..., p\}$,

$$\max_{k \in \Pi_i} \mathrm{Card}\left(\Pi_{j,k}\right) \leq A_c \left(A_j A_i^{-1} \vee 1 \right) \ .$$

(P2) Upper bound on dimensions of models in $\mathcal{M}_n$: there exists a positive constant $A_{\mathcal{M},+}$ such that for every $M \in \mathcal{M}_n$, $1 \leq D_M \leq \max\left\{ D_M, p_M^2 A_{p_M} \right\} \leq A_{\mathcal{M},+} n \left(\ln n\right)^{-2} \leq n$.

(P3) Richness of $\mathcal{M}_n$: there exist $M_0, M_1 \in \mathcal{M}_n$ such that $D_{M_0} \in [\sqrt{n}, c_{rich}\sqrt{n}]$ and $D_{M_1} \geq A_{rich} n \left(\ln n\right)^{-2}$.

(Ab') A positive constant A exists, that bounds the data and the projections s_M of the target s_* over the models M of the collection $\mathcal{M}_n$: $|Y_i| \leq A < \infty$, $\|s_M\|_\infty \leq A < \infty$ for all $M \in \mathcal{M}_n$.

(An) Uniform lower-bound on the noise level: $\sigma\left(X_i\right) \geq \sigma_{\min} > 0$ *a.s.*

(Ap$_u$) The bias decreases as a power of D_M: there exist $\beta_+ > 0$ and $C_+ > 0$ such that

$$\ell\left(s_*, s_M\right) \leq C_+ D_M^{-\beta_+} \ .$$

Theorem 10 *Under the set of assumptions (**SA**), for $A_{\mathrm{pen}} \in [0, 1)$ and $A_d > 0$, we assume that with probability at least $1 - A_d n^{-2}$ we have*

$$0 \leq \mathrm{pen}\left(M_1\right) \leq A_{\mathrm{pen}} \mathbb{E}\left[P_n\left(K s_M - K s_n\left(M_1\right) \right) \right] \ , \tag{47}$$

*where the model M_1 is defined in assumption (**P3**) of (**SA**). Then there exist two positive constants A_1, A_2 independent of n such that, with probability at least $1 - A_1 n^{-2}$, we have, for all $n \geq n_0\left(\left(\boldsymbol{SA}\right), A_{\mathrm{pen}}\right)$,*

$$D_{\widehat{M}} \geq A_2 n \ln\left(n\right)^{-2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \ln\left(n\right) \inf_{M \in \mathcal{M}_n} \left\{ \ell\left(s_*, s_n\left(M\right)\right) \right\} \ . \tag{48}$$

Thus, Theorem 10 justifies the first part (**i**) of the slope heuristics exposed in Section 5.2.1. As a matter of fact, it shows that there exists a level such that if the penalty is smaller than this level for one of the largest models, then the dimension of the output is among the largest dimensions of the collection and the excess risk of the selected estimator is much bigger than the excess risk of the oracle. Moreover, this level is given by the mean of the empirical excess risk of the least-squares estimator on each model.
The following theorem validates the second part of the slope heuristics.

Theorem 11 *Assume that the set of assumptions (**SA**) holds.*
Moreover, for some $\delta \in [0,1)$ and $A_d, A_r > 0$, assume that an event of probability at least $1 - A_d n^{-2}$ exists on which, for every model $M \in \mathcal{M}_n$ such that $D_M \geq A_{\mathcal{M},+} (\ln n)^3$, it holds

$$(2-\delta)\,\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \leq \mathrm{pen}\left(M\right) \leq (2+\delta)\,\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \tag{49}$$

together with

$$\mathrm{pen}\left(M\right) \leq A_r \frac{\left(\ln n\right)^3}{n} \tag{50}$$

*for every model $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$. Then, for $\frac{1}{2} > \eta > \left(1-\beta_+\right)_+/2$, there exist a positive constant A_3 only depending on $c_{\mathcal{M}}$ given in (**SA**) and on A_p, a positive constant A_4 only depending on constants in the set of assumptions (**SA**), a positive constant A_5 only depending on constants in the set of assumptions (**SA**) and on A_r such that with probability at least $1 - A_3 n^{-2}$, it holds for all $n \geq n_0\left(\left(\boldsymbol{SA}\right), \eta, \delta\right)$,*

$$D_{\widehat{M}} \leq n^{\eta+1/2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{A_4\left(\ln n\right)^{-1/4}}{\left(1-\delta\right)^2}\right)\ell\left(s_*, s_n\left(M_*\right)\right) + A_5\frac{\left(\ln n\right)^3}{n}\ . \tag{51}$$

Assume that in addition, the following assumption holds,

(Ap) *The bias decreases like a power of D_M : there exist $\beta_- \geq \beta_+ > 0$ and $C_+, C_- > 0$ such that*

$$C_- D_M^{-\beta_-} \leq \ell\left(s_*, s_M\right) \leq C_+ D_M^{-\beta_+}\ .$$

Then it holds with probability at least $1 - A_3 n^{-2}$, for all $n \geq n_0\left(\left(\boldsymbol{SA}\right), C_-, \beta_-, \eta, \delta\right)$,

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{\widehat{M}} \leq n^{\eta+1/2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{A_4\left(\ln n\right)^{-1/4}}{\left(1-\delta\right)^2}\right)\ell\left(s_*, s_n\left(M_*\right)\right)\ . \tag{52}$$

From Theorems 10 and 11, we identify the minimal penalty with the mean of the empirical excess risk on each model,

$$\mathrm{pen}_{\min}\left(M\right) = \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]\ .$$

Moreover, Theorem 11 states that if the penalty is close to two times the minimal procedure, then the selected estimator satisfies a pathwise oracle inequality with constant almost one, and so the model selection procedure is approximately optimal.

6 Proof of Theorem 1

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

Lemma 12 *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 and if for some positive constant A_+,*

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

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

$$\mathbb{P}\left(\max_{k \in \{1,\dots,D\}^2} |(P_n - P)(\varphi_k \cdot \varphi_l)| \geq L^{(2)}_{\alpha, r_M} \min\left\{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \right\} \sqrt{\frac{\ln n}{n}} \right) \leq n^{-\alpha} \ . \tag{53}$$

Proof of Lemma 12. For any $(k,l) \in \{1,...,D\}^2$, we have

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

and

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

Hence, we apply Bernstein's inequality (see Proposition 2.9 in [12]) and we get, for all $\gamma > 0$,

$$\mathbb{P}\left(|(P_n - P)(\varphi_k \cdot \varphi_l)| \geq \min\left\{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \right\} \left(\sqrt{\frac{2\gamma \ln n}{n}} + \frac{r_M \sqrt{D} \gamma \ln n}{3n} \right) \right) \leq 2n^{-\gamma} \ . \tag{54}$$

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

$$\frac{r_M \sqrt{D} \ln n}{n} \leq \frac{r_M \sqrt{A_+}}{\sqrt{\ln n}} \cdot \sqrt{\frac{\ln n}{n}} \leq r_M \sqrt{\frac{\ln n}{n}} \ ,$$

we get from (54) that for all $n \geq n_0(A_+)$,

$$\begin{aligned} & \mathbb{P}\left(\max_{(k,l) \in \{1,\dots,D\}^2} |(P_n - P)(\varphi_k \cdot \varphi_l)| \geq \left(\sqrt{2\gamma} + \frac{\gamma r_M}{3} \right) \min\left\{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \right\} \sqrt{\frac{\ln n}{n}} \right) \\ \leq & \sum_{(k,l) \in \{1,\dots,D\}^2} \mathbb{P}\left(|(P_n - P)(\varphi_k \cdot \varphi_l)| \geq \left(\sqrt{2\gamma} + \frac{\gamma r_M}{3} \right) \min\left\{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \right\} \sqrt{\frac{\ln n}{n}} \right) \\ \leq & \sum_{(k,l) \in \{1,\dots,D\}^2} \mathbb{P}\left(|(P_n - P)(\varphi_k \cdot \varphi_l)| \geq \min\left\{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \right\} \sqrt{\frac{2\gamma \ln n}{n}} + \frac{r_M \sqrt{D} \gamma \ln n}{3n} \right) \\ \leq & \ 2D^2 n^{-\gamma} \leq n^{-\gamma+2} \ . \end{aligned} \tag{55}$$

We deduce from (55) that (53) holds with $L^{(2)}_\alpha = \sqrt{2\alpha + 4} + (\alpha + 2)\, r_M/3 > 0$. ■

Lemma 13 *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 $\beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D$,*

$$\left\| \sum_{k=1}^{D} \beta_k \varphi_k \right\|_\infty \leq r_M \sqrt{D}\, |\beta|_\infty \ . \tag{56}$$

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

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

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

$$\mathbb{P}\left(\max_{k\in\{1,\ldots,D\}} \left| (P_n - P)\left(\psi_{1,M}\cdot\varphi_k\right)\right| \geq L^{(1)}_{A_{1,M},r_M,\alpha}\sqrt{\frac{\ln n}{n}} \right) \leq n^{-\alpha} \ . \tag{57}$$

Proof of Lemma 13. Let $\beta > 0$. By Bernstein's inequality, we get by straightforward computations (of the spirit of the proof of Lemma 12) that there exists $L_{A_{1,M},r_M,\beta} > 0$ such that, for all $k \in \{1,\ldots,D\}$,

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

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

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

and

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

Take an orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|_2)$ satisfying (**Aslb**). By Lemma 13, we get that there exists $L^{(1)}_{A_{1,M},r_M,\alpha} > 0$ such that, by setting

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

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

$$\Omega_2 = \left\{ \max_{k\in\{1,\ldots,D\}^2} \left| (P_n - P)\left(\varphi_k\cdot\varphi_l\right)\right| \leq L^{(2)}_{\alpha,r_M} \min\left\{ \|\varphi_k\|_{\infty} ; \|\varphi_l\|_{\infty} \right\} \sqrt{\frac{\ln n}{n}} \right\} \ ,$$

where $L^{(2)}_{\alpha,r_M}$ is defined in Lemma 12. By Lemma 12, we have that for all $n \geq n_0(A_+)$, $\mathbb{P}(\Omega_2) \geq 1 - n^{-\alpha}$ and so, for all $n \geq n_0(A_+)$,

$$\mathbb{P}\left(\Omega_1 \bigcap \Omega_2 \right) \geq 1 - 2n^{-\alpha} \ . \tag{58}$$

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

$$\begin{aligned}
& \mathbb{P}\left(\|s_n - s_M\|_{\infty} > C \right) \\
\leq \ & \mathbb{P}\left(\inf_{s\in\mathcal{F}^{\infty}_{>C}} P_n(Ks - Ks_M) \leq \inf_{s\in\mathcal{F}^{\infty}_{C}} P_n(Ks - Ks_M) \right) \\
= \ & \mathbb{P}\left(\sup_{s\in\mathcal{F}^{\infty}_{>C}} P_n(Ks_M - Ks) \geq \sup_{s\in\mathcal{F}^{\infty}_{C}} P_n(Ks_M - Ks) \right) \\
\leq \ & \mathbb{P}\left(\left\{ \sup_{s\in\mathcal{F}^{\infty}_{>C}} P_n(Ks_M - Ks) \geq \sup_{s\in\mathcal{F}^{\infty}_{C/2}} P_n(Ks_M - Ks) \right\} \bigcap \Omega_1 \bigcap \Omega_2 \right) + 2n^{-\alpha} \ .
\end{aligned} \tag{59}$$

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

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

we have

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

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

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

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

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

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

$$P_n\left(Ks_M - Ks\right) = h_n\left(\beta\right) \ . \tag{60}$$

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

$$|\beta|_{M,\infty} = r_M\sum_{i=1}^{p}\sqrt{A_i}\max_{k\in\Pi_i}|\beta_k| \ . \tag{61}$$

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 $B\in\mathbb{R}^{D\times D}$,

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

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

$$\begin{aligned} \|B\|_M \; = \; & \sup_{\beta\in\mathbb{R}^D,\ |\beta|_{M,\infty}=1}\left\{r_M\sum_{i=1}^{p}\sqrt{A_i}\max_{k\in\Pi_i}\left|\sum_{l=1}^{D}B_{k,l}\beta_l\right|\right\} \\ = \; & \sup_{\beta\in\mathbb{R}^D,\ |\beta|_{M,\infty}=1}\left\{r_M\sum_{i=1}^{p}\sqrt{A_i}\max_{k\in\Pi_i}\left|\sum_{j=1}^{p}\sum_{l\in\Pi_j}B_{k,l}\beta_l\right|\right\} \\ = \; & \sup_{\beta\in\mathbb{R}^D,\ |\beta|_{M,\infty}=1}\left\{\sum_{i=1}^{p}\sqrt{A_i}\max_{k\in\Pi_i}\left\{r_M\sum_{j=1}^{p}\sqrt{A_j}\max_{l\in\Pi_j}|\beta_l|\left(\sqrt{A_j^{-1}}\sum_{l\in\Pi_j}|B_{k,l}|\right)\right\}\right\} \\ = \; & \sum_{i=1}^{p}\sqrt{A_i}\max_{k\in\Pi_i}\left\{\max_{j\in\{1,...,p\}}\left\{\sqrt{A_j^{-1}}\sum_{l\in\Pi_j}|B_{k,l}|\right\}\right\} \ . \end{aligned} \tag{62}$$

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

$$\mathcal{F}_{>C}^{\infty}\subset\left\{s\in M\ ;\ s - s_M = \sum_{k=1}^{D}\beta_k\varphi_k\ \ \&\ \ |\beta|_{M,\infty}\geq C\right\} \tag{63}$$

and

$$\mathcal{F}_{C/2}^{\infty}\supset\left\{s\in M\ ;\ s - s_M = \sum_{k=1}^{D}\beta_k\varphi_k\ \ \&\ \ |\beta|_{M,\infty}\leq C/2\right\} \ . \tag{64}$$

Hence, from (59), (60) (63) and (64) we deduce that if we find on $\Omega_1 \bigcap \Omega_2$ a value of C such that

$$\sup_{\beta \in \mathbb{R}^D,\ |\beta|_{M,\infty} \geq C} h_n(\beta) < \sup_{\beta \in \mathbb{R}^D,\ |\beta|_{M,\infty} \leq C/2} h_n(\beta) \ , \tag{65}$$

then inequality (13) 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 $(k,l) \in \{1,...,D\}^2$ and $\beta = (\beta_i)_{i=1}^D \in \mathbb{R}^D$,

$$\frac{\partial h_n}{\partial \beta_k}(\beta) = R_{n,k}^{(1)} - 2\sum_{i=1}^{D} \beta_i R_{n,k,i}^{(2)} - 2\beta_k \tag{66}$$

We look now at the set of solutions β of the following system,

$$\frac{\partial h_n}{\partial \beta_k}(\beta) = 0 \ , \ \forall k \in \{1,...,D\} \ . \tag{67}$$

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 (66), the system given in (67) can be written

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

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 $\left\|R_n^{(2)}\right\|_M$, in order to show that the matrix $I_D + R_n^{(2)}$ is nonsingular. On Ω_2 we have

$$\begin{aligned}
\left\|R_n^{(2)}\right\|_M &= \sum_{i=1}^{p} \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \max_{j \in \{1,...,p\}} \left\{ \sqrt{A_j^{-1}} \sum_{l \in \Pi_j} \left| R_{n,k,l}^{(2)} \right| \right\} \right\} \\
&= \sum_{i=1}^{p} \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \max_{j \in \{1,...,p\}} \left\{ \sqrt{A_j^{-1}} \sum_{l \in \Pi_{j,k}} \left| R_{n,k,l}^{(2)} \right| \right\} \right\} \\
&\leq \sum_{i=1}^{p} \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \max_{j \in \{1,...,p\}} \left\{ \sqrt{A_j^{-1}} \left|\Pi_{j,k}\right| \max_{l \in \Pi_j} \left| (P_n - P)(\varphi_k \cdot \varphi_l) \right| \right\} \right\} \\
&\leq A_c L_{\alpha, r_M}^{(2)} \sqrt{\frac{\ln n}{n}} \sum_{i=1}^{p} \max_{j \in \{1,...,p\}} \left\{ \sqrt{\frac{A_i}{A_j}} \left(\frac{A_j}{A_i} \vee 1 \right) \sqrt{\min\{A_i; A_j\}} \right\}
\end{aligned} \tag{68}$$

We deduce from (8) and (68) that on Ω_2,

$$\left\|R_n^{(2)}\right\|_M \leq L_{A_c, \alpha, r_M} \cdot p \sqrt{\frac{A_p \ln n}{n}} \ . \tag{69}$$

Hence, from (69) 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 Ω_2,

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

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

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

Now, on Ω_1 we have by (8),

$$\left|R_n^{(1)}\right|_{M,\infty} \leq r_M \left(\sum_{i=1}^{p} \sqrt{A_i}\right) \max_{k\in\{1,\dots,D\}} \left|(P_n - P)\left(\psi_{1,M}\cdot\varphi_k\right)\right| \leq r_M L^{(1)}_{A_{1,M},r_M,\alpha}\sqrt{\frac{D\ln n}{n}} \tag{70}$$

and we deduce that for all $n_0\left(A_+, A_c, r_M, \alpha\right)$, it holds on $\Omega_2 \bigcap \Omega_1$,

$$\left|\beta^{(n)}\right|_{M,\infty} \leq \frac{1}{2}\left\|\left(I_d + R_n^{(2)}\right)^{-1}\right\|_M \left|R_n^{(1)}\right|_{M,\infty} \leq r_M L^{(1)}_{A_{1,M},r_M,\alpha}\sqrt{\frac{D\ln n}{n}} \ . \tag{71}$$

Moreover, by the formula (60) we have

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

and we thus see that h_n is concave. Hence, for all $n_0\left(A_+, A_c, r_M, \alpha\right)$, we get that on Ω_2, $\beta^{(n)}$ is the unique maximum of h_n and on $\Omega_2 \bigcap \Omega_1$, by (71), concavity of h_n and uniqueness of $\beta^{(n)}$, we get

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

with $C = 2r_M L^{(1)}_{A_{1,M},r_M,\alpha}\sqrt{\frac{D\ln n}{n}}$, which concludes the proof. ■

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] S. Boucheron and P. Massart. A high dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 2010. To appear.

[6] 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.

[7] David L. Donoho. Asymptotic minimax risk for sup-norm loss: Solution via optimal recovery. *Probability Theory and Related Fields*, 99:145–170, 1994.

[8] I.A. Ibragimov and R.Z. Khasminskii. Bounds for the risks of non-parametric regression estimates. *Theory Probab. Appl.*, 27:84–89, 1982.

[9] I.A. Ibragimov and R.Z. Khasminskii. Exact asymptotically minimax estimator for nonparametric regression in uniform norm. *Theory Probab. Appl.*, 38:775–782, 1993.

[10] R. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Annals of Probability*, 1:63–87 (electronic), 2005.

[11] Matthieu Lerasle. Optimal model selection in density estimation, 2009. arXiv:0910.1654.

[12] P. Massart. *Concentration Inequalities and Model Selection*. Springer-Verlag, 2007.

[13] 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.

[14] Adrien Saumard. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression, August 2010. hal-00512304, v1.

[15] Adrien Saumard. *Regular Contrast Estimation and the Slope Heuristics*. PhD thesis, University Rennes 1, October 2010. oai:tel.archives-ouvertes.fr:tel-00569372.

[16] Adrien Saumard. The slope heuristics in heteroscedastic regression, August 2010. hal-00512306, v1.

[17] C.J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

[18] Michel Talagrand. *The Generic Chaining*. Springer Monographs in Mathematics. Springer-Verlag, New York, 2005. Upper and lower bounds of stochastic processes.

[19] 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.