OPTIMAL MODEL SELECTION IN HETEROSCEDASTIC REGRESSION USING STRONGLY LOCALISED BASES

Adrien Saumard

Optimal model selection in heteroscedastic regression using strongly localised bases

Adrien Saumard*
CIMFAV, Universidad de Valparaíso

May 2, 2015

Abstract

We investigate optimality of model selection procedures in regard to the least-squares loss in a heteroscedatic with random design regression context. For the selection of some linear models endowed with a localized basis, as for some Haar expansions, we show the optimality of a data-driven penalty calibration procedure, the so-called slope heuristics. By doing so, we exhibit a minimal penalty being half of the optimal one. The optimal penalty shape being unknown in general, we also propose a hold-out penalization procedure and show that the latter is asymptotically optimal.

Keywords: nonparametric regression, heteroscedastic noise, random design, model selection, slope heuristics, hold-out.

1 Introduction

The slope heuristics [11] is a recent calibration method of penalization procedures in model selection : from the knowledge of a (good) penalty shape it allows to calibrate a penalty that performs an accurate model selection. It is based on the existence of a minimal penalty, around which there is a drastic change in the behavior of the model selection procedure. Moreover, the optimal penalty is simply linked to the minimal one by a factor two. The slope heuristics is thus a general method for the selection of M-estimators [6] and it has been successfully applied in various methodological studies surveyed in [8].

However, there is a gap between the wide range of applicability of the slope heuristics and its theoretical justification. Indeed, there are only a few studies, in quite restrictive frameworks, that theoretically describe the optimality of this penalty calibration procedure. First, Birgé and Massart [11] have shown the validity of the slope heuristics in a generalized linear Gaussian model setting, including the case of homescedastic regression with fixed design. Then, Arlot and Massart [6] validated the slope heuristics in a heteroscedastic with random design regression framework, for the selection of linear models of histograms. These result has been extended to the case of piecewise polynomial functions in [18]. Lerasle [12, 13] has shown the optimality of the slope heuristics in least-squares density estimation for the selection of some linear models for both independent and dependent data. Finally, it has been shown in [15] that the slope heuristics is valid for the selection of histograms in maximum likelihood density estimation.

In the present paper, we extend previous results related to heteroscedastic regression by showing the optimality of the slope heuristics for the selection of more general linear models. More precisely, the linear models that we discuss are endowed with an orthonormal basis achieving a good enough control of the sup-norms of its elements with respect to their quadratic norms, together with a control of the number of intersections of the support of the elements of the basis, see Section 3.1. This assumption on the analytical structure of the models is in particular closely related to the assumption of *localized basis* introduced by Birgé and Massart in [10] to derive accurate exponential bounds on the excess risk of general bounded M-estimators on sieves. It

*Research partly supported by the french Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration) and by post-doctoral Fondecyt grant 3140600.

allows us in particular to recover models of piecewise polynomial functions exposed in [18] and to treat, for the first time in the context of the slope heuristics, models made of Haar expansions.

If the noise is homoscedastic, then the shape of the ideal penalty is known, and is linear in the dimension of the models as in the case of Mallows' C_p. However, if the noise is heteroscedastic, then Arlot [5] showed that the ideal penalty is not in general a function of the linear dimensions of the models. Hence, a suitable estimator of this shape is needed. As emphasized by Arlot [3, 4], V-fold and resampling penalties are good, natural candidates for this task. In this paper, we show that a hold-out penalty is indeed asymptotically optimal under very mild conditions on the data split, extending to more general models previous results established in [18]. As a matter of fact, a half-and-half split leads to an optimal penalization.

The paper is organized as follows. In Section 2, we describe the statistical framework. The linear models are presented in Section 3. The slope heuristics is validated in Section 4, and the hold-out penalization is considered in Section 5. The proofs, that build upon previous results obtained in [18], are exposed in Section 6.

2 Statistical framework

Let us take n independent observations $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ with common distribution P. The feature space $\mathcal{X}$ is a subset of $\mathbb{R}^d$ and in most of the examples we will take $\mathcal{X} = [0,1]$. The marginal distribution of X_i is denoted by P^X. We assume that the data satisfy the following relation

$$Y_i = s_* (X_i) + \sigma (X_i) \varepsilon_i \ , \tag{1}$$

where $s_* \in L_2 \left(P^X\right)$. Conditionally to X_i, the residual ε_i is assumed to have zero mean and variance equal to one. The function $\sigma : \mathcal{X} \to \mathbb{R}_+$ is the unknown heteroscedastic noise level. A generic random variable with distribution P, independent of the sample $(\xi_1, ..., \xi_n)$, is denoted by $\xi = (X, Y)$.

It follows from (1) that s_* is the unknown regression function of Y with respect to X. Our aim is to estimate s_* from the sample. To do so, we are given a finite collection of models $\mathcal{M}_n$, with cardinality depending on the sample size n. Each model $M \in \mathcal{M}_n$ is assumed to be a finite-dimensional vector space. We denote by D_M the linear dimension of M. The models to be considered in this paper are introduced in details in Section 3 below.

We denote by $\|s\|_2 = \left(\int_{\mathcal{X}} s^2 dP^X\right)^{1/2}$ the usual norm in $L_2 \left(P^X\right)$ and by s_M the linear projection of s_* onto M in the Hilbert space $\left(L^2 \left(P^X\right), \|\cdot\|_2\right)$. For a function $f \in L_1 (P)$, we write $P(f) = Pf = \mathbb{E}[f(\xi)]$. By setting $K : L_2 \left(P^X\right) \to L_1 (P)$ the least-squares contrast, defined by

$$K(s) : (x, y) \mapsto (y - s(x))^2 \ , \qquad s \in L_2 \left(P^X\right) \ , \tag{2}$$

the regression function s_* satisfies

$$s_* = \arg \min_{s \in L_2(P^X)} P(K(s)) \ . \tag{3}$$

For the linear projections s_M we get

$$s_M = \arg \min_{s \in M} P(K(s)) \ . \tag{4}$$

For each model $M \in \mathcal{M}_n$, we consider a least-squares estimator $s_n (M)$ (possibly non unique), satisfying

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

where $P_n = n^{-1} \sum_{i=1}^n \delta_{\xi_i}$ is the empirical measure built from the data.

In order to avoid cumbersome notations, we will often write Ks in place of $K(s)$ for the image of a suitable function s by the contrast K. We measure the performance of the least-squares estimators by their excess loss,

$$\ell (s_*, s_n (M)) := P (Ks_n (M) - Ks_*) = \|s_n (M) - s_*\|_2^2 \ .$$

We have the following decomposition,

$$\ell\left(s_*, s_n\left(M\right)\right) = \ell\left(s_*, s_M\right) + \ell\left(s_M, s_n\left(M\right)\right) \ ,$$

where

$$\ell\left(s_*, s_M\right) := P\left(Ks_M - Ks_*\right) = \left\|s_M - s_*\right\|_2^2 \quad \text{and} \quad \ell\left(s_M, s_n\left(M\right)\right) := P\left(Ks_n\left(M\right) - Ks_M\right) \geq 0 \ .$$

The quantity $\ell\left(s_*, s_M\right)$ is called the bias of the model M and $\ell\left(s_M, s_n\left(M\right)\right)$ is the excess loss of the least-squares estimator $s_n\left(M\right)$ on the model M. By the Pythagorean identity, we have

$$\ell\left(s_M, s_n\left(M\right)\right) = \left\|s_n\left(M\right) - s_M\right\|_2^2 \ .$$

Given the collection of models $\mathcal{M}_n$, an oracle model M_* is defined as a minimizer of the losses - or equivalently excess losses - of the estimators at hand,

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n\left(M\right)\right)\right\} \ . \tag{5}$$

The associated oracle estimator $s_n\left(M_*\right)$ thus achieves the best performance in terms of excess loss among the collection $\left\{s_n\left(M\right); M \in \mathcal{M}_n\right\}$. The oracle model is a random quantity because it depends on the data and it is also unknown as it depends on the distribution P of the data. We propose to estimate the oracle model by a penalization procedure.

Given some known penalty pen, that is a function from $\mathcal{M}_n$ to $\mathbb{R}$, we consider the following data-dependent model, also called selected model,

$$\widehat{M} \in \arg\min_{M \in \mathcal{M}_n} \left\{P_n\left(Ks_n\left(M\right)\right) + \text{pen}\left(M\right)\right\} \ . \tag{6}$$

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

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

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

3 Strongly localized bases

We define here the analytic constraints that we need to put on the models in order to derive our model selection results. We also provide examples of such models.

3.1 Definition

Let us take a finite-dimensional model M with linear dimension $D = D_M$ and orthonormal basis $\left(\varphi_k\right)_{k=1}^D$. The family $\left(\varphi_k\right)_{k=1}^D$ is called a *strongly localized basis* if the following assumption is satisfied:

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

$$\sum_{i=1}^p \sqrt{A_i} \leq r_M \sqrt{D} \ , \tag{7}$$

and

$$\text{for all } \beta = \left(\beta_k\right)_{k=1}^D \in \mathbb{R}^D, \ \left\|\sum_{k=1}^D \beta_k \varphi_k\right\|_\infty \leq r_M \sum_{i=1}^p \sqrt{A_i} \max_{k \in \Pi_i} \left|\beta_k\right| \ . \tag{8}$$

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 $j \in \{1,...,p\}$,

$$\max_{k \in \Pi_i} \text{Card}\,(\Pi_{j,k}) \leq A_c \left(A_j A_i^{-1} \vee 1 \right) \ . \tag{9}$$

It is worth noting that a strongly localized basis is a localized basis in the sense of Birgé and Massart [10]. 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

$$\text{for all } \beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D, \ \left\| \sum_{k=1}^D \beta_k \varphi_k \right\|_\infty \leq r_\varphi \sqrt{D} \max_{k \in \{1,...,D\}} |\beta_k| \ . \tag{10}$$

Now, (7) and (8) imply (10). Moreover, we require in (9) a control of the number of intersections between the supports of the elements of the considered orthonormal basis to be strongly localized.

3.2 Examples

3.2.1 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 [16].

Lemma 1 *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 $c_{M,P}$,*

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} P^X(I)} \geq c_{M,P} > 0 \ . \tag{11}$$

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,

$$\text{for all } \beta = (\beta_I)_{I \in \mathcal{P}} \in \mathbb{R}^D, \quad \left\| \sum_{I \in \mathcal{P}} \beta_I \varphi_I \right\|_\infty \leq c_{M,P}^{-1} \sqrt{D}\, |\beta|_\infty \ . \tag{12}$$

By Lemma 1, we deduce that if the partition $\mathcal{P}$ satisfies the assumption of lower regularity given in (11) then inequality (8) is satisfied for M, with $p = 1$ and $r_M = c_{M,P}^{-1} > 0$. Moreover, notice that for all $(i,j) \in \{1,...,D\}^2$,

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

and in this case (**Aslb**) is straightforwardly satisfied.

3.2.2 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

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

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 [16] 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 [16].

Lemma 2 *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* $I \in \mathcal{P}$,

$$\max_{j \in \{0,...,r\}} \left\|\varphi_{I,j}\right\|_\infty \leq L_{r,c_{\min}} \frac{1}{\sqrt{\mathrm{Leb}(I)}} \ . \tag{13}$$

As a consequence, if it holds

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \mathrm{Leb}(I)} \geq c_{M,\mathrm{Leb}} > 0 \ , \tag{14}$$

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

$$\left\|\sum_{I,j} \beta_{I,j} \varphi_{I,j}\right\|_\infty \leq L_{r,c_{\min},c_{M,\mathrm{Leb}}} \sqrt{D}\,|\beta|_\infty \ , \tag{15}$$

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

Lemma 2 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 (13). Moreover, if we assume the lower regularity of the partition with respect to Leb then the orthonormal basis is localized.

It is worth noticing that in the case of histograms developed in Section 3.2.1 above, we do not need to assume the existence of a density for P^X or to restrict ourselves to the unit interval.

Finally, under assumptions of Lemma 2, the property of strongly localized basis is satisfied ($p = 1$ and $A_c = r+1$ are convenient).

3.2.3 Haar expansions

Let $\mathcal{X} = [0,1]$, $m \in \mathbb{N}$. We set for every integers $i, j, l \geq 0$, satisfying $i \leq j$ and $1 \leq l \leq 2^i$,

$$\Lambda(j) = \left\{(j,k) \ ; \ 1 \leq k \leq 2^j\right\} \ , \tag{16}$$

$$\Lambda(j,i,l) = \left\{(j,k) \ ; \ 2^{j-i}(l-1) + 1 \leq k \leq 2^{j-i} l\right\} \ . \tag{17}$$

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

$$M = \text{Span}\left\{\rho_\lambda\ ;\ \lambda \in \Lambda_m\right\}\ . \tag{18}$$

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 3 *Let $m \in \mathbb{N}$. Assume that $\mathcal{X} = [0,1]$ and let M be the model of dimension D given by (18). Then*

$$D = \text{Card}(\Lambda_m) = 2^{m+1}\ . \tag{19}$$

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

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

$$\varphi_{j,k} : x \in [0,1] \mapsto \frac{1}{\sqrt{p_{j,k,+}^2 p_{j,k,-} + p_{j,k,-}^2 p_{j,k,+}}}\left(p_{j,k,+}\mathbf{1}_{[2^{-j}(k-1),2^{-j}(k-1/2)]} - p_{j,k,-}\mathbf{1}_{\left(2^{-j}\left(k-\frac{1}{2}\right),2^{-j}k\right]}\right)\ . \tag{20}$$

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 $j \geq 0$, $1 \leq k \leq 2^j$,

$$\left\|\varphi_{j,k}\right\|_\infty \leq \sqrt{\frac{2}{c_{\min}}}2^{j/2}\ . \tag{21}$$

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

$$\sum_{j=-1}^{m} \sqrt{A_j} \leq \left(\sqrt{2}+1\right)\sqrt{D} \tag{22}$$

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

$$\left\|\sum_{\lambda \in \Lambda_m} \beta_\lambda \varphi_\lambda\right\|_\infty \leq \sqrt{\frac{2}{c_{\min}}}\sum_{j=-1}^{m}\sqrt{A_j}\max_{k \in \Lambda_j}|\beta_k|\ . \tag{23}$$

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

$$\max_{\mu \in \Lambda_i} \text{Card}\left(\Lambda_{j,\mu}\right) \leq A_j A_i^{-1} \vee 1\ . \tag{24}$$

By Lemma 3, which proof is straightforward and 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 (18) 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.

4 The slope heuristics

4.1 Assumptions and comments

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 strongly localized bases: 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

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

The set of assumptions (**SA**) can be divided into three groups. Firstly, assumptions (**P1**), (**P2**), (**P3**) and (**Ap**$_u$) are linked to properties of the collection of models $\mathcal{M}_n$. Secondly, assumptions (**An**) and (**Ab**) give some constraints on the general regression relation stated in (1). Thirdly, assumption (**Auslb**) specifies some quantities related to the choice of models with strongly localized bases. More precisely, the latter assumption ensures uniformity along the collection of models of the constants defining the strongly localized bases.

Assumption (**P1**) states that the collection of models has a "small" complexity, more precisely a polynomially increasing one with respect to the amount of data. For this kind of complexities, if one wants to design a good model selection procedure for prediction, the chosen penalty should estimate the mean of the ideal one on each model. Indeed, as Talagrand's type concentration inequalities for the empirical process are exponential, they allow to neglect the deviations of the quantities of interest from their mean, uniformly over the collection of models. This is not the case for large collections of models, where one has to put an extra-log factor inside the penalty, depending on the complexity of the collection of models, see for instance [9, 7].

We assume in (**P3**) that the collection of models contains a model M_0 of reasonably large dimension and a model M_1 of high dimension, which is necessary since we prove the existence of a jump between high and reasonably large dimensions. One can notice that in practice, the parameter β_+, which depends on the bias of the model is not known and so the existence of M_0 is not straightforward. However, it suffices for the statistician to take at least one model per dimension lower than the chosen upper bound to ensure the existence of M_0 and M_1.

Assumption ($\mathbf{Ap}_u$) states that the models have good enough approximation properties in terms of the quadratic loss. Furthermore, assumption (**Ab**) is rather restrictive, since it excludes Gaussian noise. However, the assumption of bounded noise is somehow classical when dealing with M-estimation and related procedures. Indeed, a central tool in this field is empirical process theory and more especially, concentration inequalities for the supremum of the empirical process. We used the classical inequalities of Bousquet, and Klein and Rio in [17] and [18]. As a matter of fact, we do not know yet if an adaptation of our proofs (including results established in [17]) by using extensions of the latter inequalities to some unbounded cases - as for instance in Adamczak's concentration inequalities [1] - would be possible.

The noise restriction stated in (**An**) is needed to derive our results which are optimal to the first order. It is quite common in this context since it is also needed in the work of Arlot and Massart [6] concerning the validation of the slope heuristics for histogram models and in [18] for piecewise polynomials.

4.2 Statement of the theorems

We are now able to state our main results leading to the slope heuristics. They describe the behavior of the penalization procedure defined in (6).

Theorem 4 *Take a positive penalty: for all $M \in \mathcal{M}_n$, $\mathrm{pen}(M) \geq 0$. Suppose that the assumptions (**SA**) of Section 4.1 hold, and furthermore suppose that for $A_{\mathrm{pen}} \in [0,1)$ and $A_p > 0$ the model M_1 of assumption (**P3**) satisfies*

$$0 \leq \mathrm{pen}(M_1) \leq A_{\mathrm{pen}} \mathbb{E}\left[P_n\left(Ks_{M_1} - Ks_n(M_1)\right)\right] \ , \tag{25}$$

*with probability at least $1 - A_p n^{-2}$. Then there exist a constant $A_1 > 0$ only depending on constants in (**SA**), as well as an integer n_0 and a positive constant A_2 only depending on A_{pen} and on constants in (**SA**) such that, for all $n \geq n_0$, it holds with probability at least $1 - A_1 n^{-2}$,*

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \frac{n^{\beta_+/(1+\beta_+)}}{(\ln n)^3} \inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n(M)\right)\right\} \ , \tag{26}$$

*where $\beta_+ > 0$ is defined in assumption ($\mathbf{Ap}_u$) of (**SA**).*

Theorem 4 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 loss of the selected estimator is much larger than the excess loss of the oracle. Moreover, this level is given by the mean of the empirical excess loss of the least-squares estimator on each model. Let us also notice that the lower bound given in (26) gets worse as β_+ increases. This is due to the fact that when β_+ increases, the approximation properties of the models improve and the performances in terms of excess loss for the oracle estimator also improve.

Theorem 5 *Suppose that the assumptions (**SA**) of Section 4.1 hold, and furthermore suppose that for some $\delta \in [0,1)$ and $A_p, A_r > 0$, there exists an event of probability at least $1 - A_p n^{-2}$ on which, for every model $M \in \mathcal{M}_n$ such that $D_M \geq A_{\mathcal{M},+}(\ln n)^3$, it holds*

$$\left|\mathrm{pen}(M) - 2\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right]\right| \leq \delta\left(\ell\left(s_*, s_M\right) + \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right]\right) \tag{27}$$

together with

$$\left|\mathrm{pen}(M)\right| \leq A_r \left(\frac{\ell\left(s_*, s_M\right)}{(\ln n)^2} + \frac{(\ln n)^3}{n}\right) \ . \tag{28}$$

Then, for any $\eta \in \left(0, \beta_+ / \left(1+\beta_+\right)\right)$*, there exist an integer* n_0 *only depending on* η, δ *and* β_+ *and on constants in (**SA**), a positive constant* A_3 *only depending on* $c_{\mathcal{M}}$ *given in (**SA**) and on* A_p*, two positive constants* A_4 *and* A_5 *only depending on constants in (**SA**) and on* A_r *and a sequence*

$$\theta_n \leq \frac{A_4}{(\ln n)^{1/4}} \tag{29}$$

such that it holds for all $n \geq n_0$*, with probability at least* $1 - A_3 n^{-2}$*,*

$$D_{\widehat{M}} \leq n^{\eta + 1/\left(1+\beta_+\right)}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{(1-\delta)^2}\right) \ell\left(s_*, s_n\left(M_*\right)\right) + A_5 \frac{(\ln n)^3}{n} \ . \tag{30}$$

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

$$A_{\mathcal{M},+} (\ln n)^3 \leq D_{\widehat{M}} \leq n^{\eta + 1/\left(1+\beta_+\right)} \tag{31}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{(1-\delta)^2}\right) \ell\left(s_*, s_n\left(M_*\right)\right) \ . \tag{32}$$

Theorem 5 states that if the penalty is close to twice the minimal one, then the selected estimator satisfies a pathwise oracle inequality with constant almost one, and so the model selection procedure is approximately optimal. Moreover, the dimension of the selected model is of reasonable dimension, bounded by a power less than one of the sample size.

Condition (**Ap**) allows to remove the remainder terms from the oracle inequality (30) by ensuring that the selected model is of dimension not too small, as stated in (31). Assumption (**Ap**) is the conjunction of assumption ($\mathbf{Ap}_u$) with a polynomial lower bound of the bias of the models. On histogram models, Arlot showed in Section 8.10 of [2] that this lower bound is satisfied for non constant α-Hölder, $\alpha \in (0, 1]$, regression functions and for regular partitions.

Finally, from Theorems 4 and 5, we identify the minimal penalty with the mean of the empirical excess loss on each model,

$$\mathrm{pen}_{\min}\left(M\right) = \mathbb{E}\left[P_n\left(K s_M - K s_n\left(M\right)\right)\right] \ ,$$

thus generalizing the results obtained in [6] and [18] to the case of models endowed with strongly localized bases.

5 Hold-out penalization

The conditions on the penalty given in Theorems 4 and 5 can not be directly checked in practice. Indeed, they are expressed in terms of the mean of the empirical excess loss on each model, which is an unknown quantity in general.

In the case where the noise level is homoscedastic but unknown, Mallows' penalty, which is known to be asymptotically optimal, is only known through a constant, the noise level, which can be estimated *via* the slope heuristics (for practical issues about the slope heuristics, see Baudry et al. [8]). But in the common situation where the noise level is sufficiently heteroscedastic, the shape of the ideal penalty is not linear in the dimension of the models and not even a *function* of the linear dimensions. In such a case, Arlot [5] proved that

any calibration of a linear penalty leads to a suboptimal procedure, but yet can achieve an oracle inequality with a leading constant more than one.

In order to achieve a nearly optimal selection procedure in the general situation, it remains to estimate the ideal penalty or, thanks to the slope heuristics, the shape of the ideal penalty. This section is devoted to this task. We propose a hold-out type penalty that automatically adapts to heteroscedasticity. Let us now detail our hold-out penalization procedure.

The ideal penalty is defined by

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

for all $M \in \mathcal{M}_n$. A natural idea is to divide the data into two groups, indexed by I_1 and I_2, satisfying $I_1 \cap I_2 = \emptyset$ and $I_1 \cup I_2 = \{1, ..., n\}$ and to propose the following hold-out type penalty,

$$\mathrm{pen}_{ho,C}(M) := C\left(P_{n_2}(Ks_{n_1}(M)) - P_{n_1}(Ks_{n_1}(M))\right) \ ,$$

where $P_{n_i} = 1/n_i \sum_{j \in I_i} \delta_{\xi_j}$, $n_i = \mathrm{Card}(I_i)$, for $i = 1, 2$, $s_{n_1}(M) \in \arg\min_{s \in M} P_{n_1}(Ks)$ and $C > 0$ is a constant to be determined. Indeed, if n_1 is not too small, $P_{n_1}(Ks_{n_1}(M))$ is likely to vary like $P_n(Ks_n(M))$ and $P_{n_2}(Ks_{n_1}(M))$ is, conditionally to $\left(\xi_j\right)_{j \in I_1}$, an unbiased estimate of $P(Ks_{n_1}(M))$, which again is likely to vary like $P(Ks_n(M))$. Moreover, we see from Theorem 10 in [17] that when the model M is fixed, the quantities $P_n(Ks_n(M))$ and $P(Ks_n(M))$ are almost inversely proportional to n, so a good constant in front of the hold-out penalty should be $C_{opt} = n_1/n$.

The previous observation is justified by the following theorem, where for the sake of clarity we fixed $n_1 = n_2 = n/2$. We set

$$\mathrm{pen}_{ho}(M) = \frac{1}{2}\left(P_{n_2}(Ks_{n_1}(M)) - P_{n_1}(Ks_{n_1}(M))\right) \quad \text{and} \quad \widehat{M}_{1/2} \in \arg\min_{M \in \mathcal{M}_n} \left\{P_n(Ks_n(M)) + \mathrm{pen}_{ho}(M)\right\} \ . \tag{33}$$

Theorem 6 *Consider the procedure defined in (33), with $n_1 = n_2 = n/2$. Suppose that the assumptions (**SA**) of Section 4.1 hold. Then, for any $\eta \in \left(0, \beta_+/\left(1+\beta_+\right)\right)$, there exist an integer n_0 only depending on η and on constants in (**SA**), a positive constant A_6 only depending on $c_{\mathcal{M}}$ given in (**SA**), two positive constants A_7 and A_8 only depending on constants in (**SA**) and a sequence $\theta_n \leq A_7 (\ln n)^{-1/4}$ such that it holds for all $n \geq n_0$, with probability at least $1 - A_6 n^{-2}$,*

$$D_{\widehat{M}_{1/2}} \leq n^{\eta + 1/\left(1+\beta_+\right)}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{1/2}\right)\right) \leq (1+\theta_n)\,\ell\left(s_*, s_n(M_*)\right) + A_8 \frac{(\ln n)^3}{n} \ . \tag{34}$$

*Assume that in addition (**Ap**) holds (see Theorem 5). Then it holds for all $n \geq n_0\left((\boldsymbol{SA}), C_-, \beta_-, \eta\right)$, with probability at least $1 - A_6 n^{-2}$,*

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{1/2}\right)\right) \leq (1+\theta_n) \inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n(M)\right)\right\} \ . \tag{35}$$

Theorem 6 shows the asymptotic optimality of the hold-out penalization procedure, for a half-and-half split of the data. This is a remarkable fact compared to the classical hold-out, defined by

$$\widehat{M}_{ho} \in \arg\min_{M \in \mathcal{M}_n} \left\{P_{n_2}(Ks_{n_1}(M))\right\} \ . \tag{36}$$

Indeed, the choice $n_1 = n/2$ in (36) is likely to lead to an asymptotically suboptimal procedure, as the criterion is close in expectation to $P\left(Ks_{n/2}(M)\right)$, and so is close to the oracle, but for $n/2$ data points. The hold-out

penalization allows us to overcome this difficulty. Arlot [3, 4] described similar advantages for resampling and V-fold penalties.

Notice also that the random hold-out penalty proposed by Arlot [4] is proportional to the mean along the splits of our hold-out penalty, providing thus a "stabilization effect" in practice. This should bring some improvement compared to our unique split, at the price of increased computational cost. However, the stabilization effect seems more difficult to study mathematically, and our results provide a first step toward the study of the more complicated resampling penalties.

6 Proofs

We first recall from [18], Section 5, that Theorems 4, 5 and 6 are valid under the following general set of assumptions (i.e. by replacing (**SA**) by **(GSA)** in the statement of the theorems):

General set of assumptions: (GSA)

Assume (**P1**), (**P2**), (**P3**), (**Ab**), (**An**) and (**Ap**$_u$) of (**SA**). Furthermore suppose that,

(Alb) there exists a constant $r_{\mathcal{M}}$ such that for each $M \in \mathcal{M}_n$ one can find an orthonormal basis $(\varphi_k)_{k=1}^{D_M}$ satisfying, for all $(\beta_k)_{k=1}^{D_M} \in \mathbb{R}^{D_M}$,

$$\left\| \sum_{k=1}^{D_M} \beta_k \varphi_k \right\|_\infty \leq r_{\mathcal{M}} \sqrt{D_M} \left| \beta \right|_\infty \ ,$$

where $\left|\beta\right|_\infty = \max\left\{ \left|\beta_k\right| ; k \in \{1, ..., D_M\} \right\}$.

(Ac$_\infty$) a positive integer n_1 exists such that, for all $n \geq n_1$, there exist a positive constant A_{cons} and an event Ω_∞ of probability at least $1 - n^{-2-\alpha_{\mathcal{M}}}$, on which for all $M \in \mathcal{M}_n$,

$$\left\| s_n\left(M\right) - s_M \right\|_\infty \leq A_{cons} \sqrt{\frac{D_M \ln n}{n}} \ . \tag{37}$$

As assumption (**Alb**) in (**GSA**) is satisfied under assumption (**Auslb**) of the set of assumptions (**SA**) (see section 3.1), it remains to prove the convergence in sup-norm (**Ac**$_\infty$) from assumption (**Auslb**) and the proofs of Theorems 4, 5 and 6 will be complete. This is done *via* the following theorem.

Theorem 7 *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 $\left|\psi_{1,M}\left(X, Y\right)\right| \leq A_{1,M}$ a.s.*

If there exists $A_+ > 0$ such that

$$\max\left\{ D, p^2 A_p \right\} \leq A_+ \frac{n}{\left(\ln n\right)^2} \ , \tag{38}$$

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

$$\mathbb{P}\left(\left\| s_n - s_M \right\|_\infty \geq L_{A_{1,M}, r_M, \alpha} \sqrt{\frac{D \ln n}{n}} \right) \leq n^{-\alpha} \ . \tag{39}$$

Notice that assumption (**Ab'**) in Theorem 7 is included in assumption (**Ab**) of (**GSA**). Before stating the proof of Theorem 7, we need two preliminary lemmas.

Lemma 8 *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

$$\mathbb{P}\left(\max_{k \in \{1,...,D\}^2} |(P_n - P)(\varphi_k \cdot \varphi_l)| \geq L^{(2)}_{\alpha, r_M} \min\left\{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \right\} \sqrt{\frac{\ln n}{n}} \right) \leq n^{-\alpha} \ . \tag{40}$$

Proof of Lemma 8. 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 [14]) and we get, for all $\gamma > 0$,

$$\mathbb{P}\left(|(P_n - P)(\varphi_k \cdot \varphi_l)| \geq \min\left\{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \right\} \left(\sqrt{\frac{2\gamma \ln n}{n}} + \frac{r_M \sqrt{D} \gamma \ln n}{3n} \right) \right) \leq 2n^{-\gamma} \ . \tag{41}$$

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

$$\begin{aligned} & \mathbb{P}\left(\max_{(k,l) \in \{1,...,D\}^2} |(P_n - P)(\varphi_k \cdot \varphi_l)| \geq \left(\sqrt{2\gamma} + \frac{\gamma r_M}{3} \right) \min\left\{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \right\} \sqrt{\frac{\ln n}{n}} \right) \\ \leq & \sum_{(k,l) \in \{1,...,D\}^2} \mathbb{P}\left(|(P_n - P)(\varphi_k \cdot \varphi_l)| \geq \left(\sqrt{2\gamma} + \frac{\gamma r_M}{3} \right) \min\left\{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \right\} \sqrt{\frac{\ln n}{n}} \right) \\ \leq & \sum_{(k,l) \in \{1,...,D\}^2} \mathbb{P}\left(|(P_n - P)(\varphi_k \cdot \varphi_l)| \geq \min\left\{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \right\} \sqrt{\frac{2\gamma \ln n}{n}} + \frac{r_M \sqrt{D} \gamma \ln n}{3n} \right) \\ \leq & 2D^2 n^{-\gamma} \leq n^{-\gamma+2} \ . \end{aligned} \tag{42}$$

We deduce from (42) that (40) holds with $L^{(2)}_\alpha = \sqrt{2\alpha + 4} + (\alpha + 2)\, r_M/3 > 0$. ■

Lemma 9 *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 $\beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D$,*

$$\left\| \sum_{k=1}^{D} \beta_k \varphi_k \right\|_\infty \leq r_M \sqrt{D}\, |\beta|_\infty \ . \tag{43}$$

*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

$$\mathbb{P}\left(\max_{k\in\{1,\ldots,D\}} \left| (P_n - P)\left(\psi_{1,M}\cdot\varphi_k\right)\right| \geq L^{(1)}_{A_{1,M},r_M,\alpha}\sqrt{\frac{\ln n}{n}} \right) \leq n^{-\alpha} \ . \tag{44}$$

Proof of Lemma 9. Let $\beta > 0$. By Bernstein's inequality, we get by straightforward computations (of the spirit of the proof of Lemma 8) 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 7. Let $C > 0$. 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 9, 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

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

where $L^{(2)}_{\alpha,r_M}$ is defined in Lemma 8. By Lemma 8, we have that for all $n \geq n_0(A_+)$, $\mathbb{P}(\Omega_2) \geq 1 - n^{-\alpha}$ and so, for all $n \geq n_0(A_+)$,

$$\mathbb{P}\left(\Omega_1 \bigcap \Omega_2\right) \geq 1 - 2n^{-\alpha} \ . \tag{45}$$

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

$$\begin{aligned}
& \mathbb{P}\left(\|s_n - s_M\|_{\infty} > C\right) \\
\leq\ & \mathbb{P}\left(\inf_{s\in\mathcal{F}^{\infty}_{>C}} P_n(Ks - Ks_M) \leq \inf_{s\in\mathcal{F}^{\infty}_{C}} P_n(Ks - Ks_M) \right) \\
=\ & \mathbb{P}\left(\sup_{s\in\mathcal{F}^{\infty}_{>C}} P_n(Ks_M - Ks) \geq \sup_{s\in\mathcal{F}^{\infty}_{C}} P_n(Ks_M - Ks) \right) \\
\leq\ & \mathbb{P}\left(\left\{ \sup_{s\in\mathcal{F}^{\infty}_{>C}} P_n(Ks_M - Ks) \geq \sup_{s\in\mathcal{F}^{\infty}_{C/2}} P_n(Ks_M - Ks) \right\} \bigcap \Omega_1 \bigcap \Omega_2 \right) + 2n^{-\alpha} \ .
\end{aligned} \tag{46}$$

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

$$P_n\left(Ks_M - Ks\right) = h_n\left(\beta\right) \ . \tag{47}$$

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

$$|\beta|_{M,\infty} = r_M \sum_{i=1}^{p}\sqrt{A_i}\max_{k\in\Pi_i}|\beta_k| \ . \tag{48}$$

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

$$\begin{aligned} \|B\|_M \; = \; & \sup_{\beta\in\mathbb{R}^D,\ |\beta|_{M,\infty}=1}\left\{r_M\sum_{i=1}^{p}\sqrt{A_i}\max_{k\in\Pi_i}\left|\sum_{l=1}^{D}B_{k,l}\beta_l\right|\right\} \\ = \; & \sup_{\beta\in\mathbb{R}^D,\ |\beta|_{M,\infty}=1}\left\{r_M\sum_{i=1}^{p}\sqrt{A_i}\max_{k\in\Pi_i}\left|\sum_{j=1}^{p}\sum_{l\in\Pi_j}B_{k,l}\beta_l\right|\right\} \\ = \; & \sup_{\beta\in\mathbb{R}^D,\ |\beta|_{M,\infty}=1}\left\{\sum_{i=1}^{p}\sqrt{A_i}\max_{k\in\Pi_i}\left\{r_M\sum_{j=1}^{p}\sqrt{A_j}\max_{l\in\Pi_j}|\beta_l|\left(\sqrt{A_j^{-1}}\sum_{l\in\Pi_j}|B_{k,l}|\right)\right\}\right\} \\ = \; & \sum_{i=1}^{p}\sqrt{A_i}\max_{k\in\Pi_i}\left\{\max_{j\in\{1,...,p\}}\left\{\sqrt{A_j^{-1}}\sum_{l\in\Pi_j}|B_{k,l}|\right\}\right\} \ . \end{aligned} \tag{49}$$

Notice that by inequality (8) of (**Aslb**), it holds

$$\mathcal{F}_{>C}^{\infty} \subset \left\{s\in M \ ; \ s - s_M = \sum_{k=1}^{D}\beta_k\varphi_k \ \ \& \ \ |\beta|_{M,\infty}\geq C\right\} \tag{50}$$

and

$$\mathcal{F}_{C/2}^{\infty} \supset \left\{s\in M \ ; \ s - s_M = \sum_{k=1}^{D}\beta_k\varphi_k \ \ \& \ \ |\beta|_{M,\infty}\leq C/2\right\} \ . \tag{51}$$

Hence, from (46), (47) (50) and (51) we deduce that if we find on $\Omega_1 \bigcap \Omega_2$ a value of C such that

$$\sup_{\beta \in \mathbb{R}^D,\ |\beta|_{M,\infty} \geq C} h_n(\beta) < \sup_{\beta \in \mathbb{R}^D,\ |\beta|_{M,\infty} \leq C/2} h_n(\beta) \ , \tag{52}$$

then inequality (39) follows and Theorem 7 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 $\beta = (\beta_i)_{i=1}^D \in \mathbb{R}^D$,

$$\frac{\partial h_n}{\partial \beta_k}(\beta) = R_{n,k}^{(1)} - 2\sum_{i=1}^{D} \beta_i R_{n,k,i}^{(2)} - 2\beta_k \tag{53}$$

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

$$\frac{\partial h_n}{\partial \beta_k}(\beta) = 0 \ , \ \forall k \in \{1,...,D\} \ . \tag{54}$$

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 (53), the system given in (54) 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 $\left\|R_n^{(2)}\right\|_M$, in order to show that the matrix $I_D + R_n^{(2)}$ is nonsingular. On Ω_2 we have

$$\begin{aligned}
\left\|R_n^{(2)}\right\|_M &= \sum_{i=1}^{p} \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \max_{j \in \{1,...,p\}} \left\{ \sqrt{A_j^{-1}} \sum_{l \in \Pi_j} \left|R_{n,k,l}^{(2)}\right| \right\} \right\} \\
&= \sum_{i=1}^{p} \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \max_{j \in \{1,...,p\}} \left\{ \sqrt{A_j^{-1}} \sum_{l \in \Pi_{j,k}} \left|R_{n,k,l}^{(2)}\right| \right\} \right\} \\
&\leq \sum_{i=1}^{p} \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \max_{j \in \{1,...,p\}} \left\{ \sqrt{A_j^{-1}} \left|\Pi_{j,k}\right| \max_{l \in \Pi_j} \left|\left(P_n - P\right)\left(\varphi_k \cdot \varphi_l\right)\right| \right\} \right\} \\
&\leq A_c L_{\alpha, r_M}^{(2)} \sqrt{\frac{\ln n}{n}} \sum_{i=1}^{p} \max_{j \in \{1,...,p\}} \left\{ \sqrt{\frac{A_i}{A_j}} \left(\frac{A_j}{A_i} \vee 1\right) \sqrt{\min\left\{A_i; A_j\right\}} \right\}
\end{aligned} \tag{55}$$

We deduce from (7) and (55) that on Ω_2,

$$\left\|R_n^{(2)}\right\|_M \leq L_{A_c, \alpha, r_M} \cdot p \sqrt{\frac{A_p \ln n}{n}} \ . \tag{56}$$

Hence, from (56) 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 (7),

$$\left|R_n^{(1)}\right|_{M,\infty} \leq r_M \left(\sum_{i=1}^{p} \sqrt{A_i}\right) \max_{k\in\{1,\ldots,D\}} \left|(P_n - P)\left(\psi_{1,M}\cdot\varphi_k\right)\right| \leq r_M L^{(1)}_{A_{1,M},r_M,\alpha}\sqrt{\frac{D\ln n}{n}} \tag{57}$$

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

$$\left|\beta^{(n)}\right|_{M,\infty} \leq \frac{1}{2}\left\|\left(I_d + R_n^{(2)}\right)^{-1}\right\|_M \left|R_n^{(1)}\right|_{M,\infty} \leq r_M L^{(1)}_{A_{1,M},r_M,\alpha}\sqrt{\frac{D\ln n}{n}} \ . \tag{58}$$

Moreover, by the formula (47) we have

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

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 (58), 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] R. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:1000–1034, 2008.

[2] S. Arlot. *Resampling and Model Selection.* PhD thesis, University Paris-Sud 11, December 2007. oai:tel.archives-ouvertes.fr:tel-00198803_v1.

[3] S. Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566v2.

[4] S. Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624, 2009.

[5] S. Arlot. Choosing a penalty for model selection in heteroscedastic regression, June 2010. arXiv:0812.3141.

[6] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[7] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[8] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470, 2012.

[9] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[10] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[11] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[12] M. Lerasle. Optimal model selection for density estimation of stationary data under various mixing conditions. *Ann. Statist.*, 39(4):1852–1877, 2011.

[13] M. Lerasle. Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(3):884–908, 2012.

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

[15] A. Saumard. Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models, August 2010. hal-00512310.

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

[17] A. Saumard. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. *Electron. J. Statist.*, 6(1-2):579–655, 2012.

[18] A. Saumard. Optimal model selection in heteroscedastic regression using piecewise polynomial functions. *Electron. J. Statist.*, 7:1184–1223, 2013.