ADAPTIVE ESTIMATION IN AUTOREGRESSION OR β-MIXING REGRESSION VIA MODEL SELECTION

By Y. Baraud, F. Comte, G. Viennet

Ecole Normale Supérieure, DMA, Université Paris VI and Université Paris VII

We study the problem of estimating some unknown regression function in a β-mixing dependent framework. To this end, we consider some collection of models which are finite dimensional spaces. A penalized least-squares estimator (PLSE) is built on a data driven selected model among this collection. We state non asymptotic risk bounds for this PLSE and give several examples where the procedure can be applied (autoregression, regression with arithmetically β-mixing design points, regression with mixing errors, estimation in additive frameworks, estimation of the order of the autoregression ...). In addition we show that under a weak moment condition on the errors, our estimator is adaptive in the minimax sense simultaneously over some family of Besov balls.

1. Introduction We consider the problem of estimating the unknown function f, from $\mathbb{R}^k$ into $\mathbb{R}$, based on the observation of n (possibly) dependent data $(Y_i, \vec{X}_i)$, $1 \leq i \leq n$, arising from the model

$$Y_i = f(\vec{X}_i) + \varepsilon_i. \tag{1.1}$$

We assume that $(\vec{X}_i)_{1\leq i\leq n}$ is a stationary sequence of random vectors in $\mathbb{R}^k$ and we denote by μ the common law of the $\vec{X}_i$'s. The ε_i's are unobservable identically distributed centered random variables admitting a finite variance denoted by σ_2^2. Throughout the paper we assume that σ_2^2 is a known quantity (or that a bound on it is known). In this introduction, we assume that the ε_i's are independent random variables. As an example of model (1.1), consider the case of a random design set $\vec{X}_i = X_i$ with values in $[0, 1]$ with a regression function f assumed to satisfy some Hölderian regularity condition

$$\sup_{0\leq x<y\leq 1} \frac{|f(x) - f(y)|}{(y-x)^\alpha} = |f|^{(\alpha)} < +\infty \tag{1.2}$$

for some $\alpha \in (0, 1]$. Another possible illustration is a linear autoregressive model

$$X_i = \sum_{j=1}^{k'} \beta_j X_{i-j} + \varepsilon_i \tag{1.3}$$

where k' is an integer smaller than k. This means that $Y_i = X_i$, $\vec{X}_i = (X_{i-1}, \ldots, X_{i-k})'$ and $f(u_1, ..., u_k) = \sum_{j=1}^{k'} \beta_j u_j$. Such models have been extensively studied in the past under the conditions that α or k' are known. There

AMS 1991 subject classifications. Primary 62G05; Secondary 62J02.

Key words and phrases. Nonparametric regression, Least-squares estimator, Model selection, Adaptive estimation, Autoregression order, Additive framework, Time series, Mixing processes

have been some generalizations to the cases of unknown α and k', but then the results are typically given in an asymptotic form (as $n \to +\infty$).

In this paper, the aim is to introduce an estimation procedure for Model (1.1) which, when applied to some Hölderian function f satisfying (1.2) with unknown values of α and $|f|^{(\alpha)}$, will perform almost as well as a procedure based on the knowledge of those two parameters. This is what is usually called adaptation. In the same way, our procedure will result in estimation of Model (1.3) with an unknown value of k' ($k' \leq k$, k known) which is almost as good as if k' were known. Moreover, the results will be given in the form of non asymptotic bounds for the risk of our estimators. Many other examples can be treated by the same method. One could, for instance, replace the regularity conditions (1.2) by more sophisticated ones and Model (1.3) by a nonlinear analogue.

In order to explain the main idea underlying the approach, let us turn back to the two previous examples. Model (1.3) says that f belongs to some specific k'-dimensional linear space $S_{k'}$ of functions from $\mathbb{R}^{k'}$ to $\mathbb{R}$. When k' is known, a classical estimator of f is the least squares estimator over $S_{k'}$. Dealing with an unknown k' therefore amounts to choosing a "good" value $\hat{k}$ of k' from the data. By "good", we mean here that the estimation procedure based on $\hat{k}$ should perform almost as well as the procedure based on the true value of k'.

The treatment of Model (1.1) when f satisfies a condition of type (1.2) is actually quite similar. Let us expand f in some suitable orthonormal basis $\{\phi_j\}_{j\geq 1}$ of $\mathbb{L}^2([0,1], dx)$ (the Haar basis for instance). Then (1.1) can be written as $Y_i = \sum_{j=1}^{\infty} \beta_j \phi_j(X_i) + \varepsilon_i$ and a classical procedure for estimating f is as follows: define S_J to be the J-dimensional linear space generated by $\phi_1, \ldots, \phi_J$ and $\hat{f}_J$ to be the least squares estimator on S_J, that is the least squares estimator for Model (1.1) when f is supposed to belong to S_J. The problem is to determine from the data set some $\hat{J}$ in such a way that the least squares estimator $\hat{f}_{\hat{J}}$ performs almost as well as the best least-squares estimator of f, i.e. the one which achieves the minimum of the risk.

In order to give a further explanation of the procedure, we need to be precise as to the "risk" we are dealing with. Throughout the paper we consider least-squares estimators of f, obtained by minimizing over a finite dimensional linear subspace $S \subset \mathbb{L}^2(\mathbb{R}^k, dx)$ the (least squares) contrast function γ_n defined by

$$\forall t \in \mathbb{L}^2(\mathbb{R}^k, dx), \quad \gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - t(\vec{X}_i))^2. \tag{1.4}$$

A minimizer of γ_n in S, $\hat{f}_S$, always exists but might not be unique. Indeed, in common situations the minimization of γ_n over S leads to an affine space of possible solutions and then it becomes impossible to consider the $\mathbb{L}^2(\mathbb{R}^k, dx)$-quadratic risk of "the least-squares estimator" of f in S. In contrast, the (random) $\mathbb{R}^n$-vector $(\hat{f}_S(\vec{X}_1), ..., \hat{f}_S(\vec{X}_n))'$ is always uniquely defined; this is the reason we consider the risk of $\hat{f}_S$ based on the design points, i.e.

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(f(\vec{X}_i) - \hat{f}_S(\vec{X}_i)\right)^2\right] = \mathbb{E}\left[\|f - \hat{f}_S\|_n^2\right].$$

In addition, under suitable assumptions on the design set and the ε_i's, the risk of $\hat{f}_S$ can be decomposed in a classical way into a bias and a variance term. More precisely, we have

$$\mathbb{E}[\|f - \hat{f}_S\|_n^2] \leq \mathrm{d}_\mu^2(f, S) + \sigma_2^2 \frac{\dim(S)}{n}, \tag{1.5}$$

where for $t, s \in \mathbb{L}^2(\mathbb{R}^k, \mu)$, $\mathrm{d}_\mu^2(s,t)$ denotes $\mathbb{E}[(t(\vec{X}_1) - s(\vec{X}_1))^2]$ and $\mathrm{d}_\mu^2(f,S) = \inf_{s \in S} \mathrm{d}_\mu^2(f,s)$. The inequality (1.5) is usually sharp; note that equality occurs when the $\vec{X}_i$'s and the ε_i's are independent for instance.

Coming back to Model (1.1) we see that the quadratic risk $\mathbb{E}[\|f - \hat{f}_J\|_n^2]$ is of order

$$\mathrm{d}_\mu^2(f, S_J) + \sigma_2^2 \frac{J}{n}, \tag{1.6}$$

for S_J generated by the Haar basis $(\phi_j)_{1 \leq j \leq J}$ as above. Then (1.2) standardly implies that $\mathrm{d}_\mu(f, S_J) \leq C|f|^{(\alpha)} J^{-\alpha}$ whatever μ. When α and $|f|^{(\alpha)}$ are known, it is possible to determine the value of J that minimizes (1.6). If α and $|f|^{(\alpha)}$ are unknown, the problem of adaptation, that is doing almost as well as if they were known, clearly amounts to choosing an estimation procedure $\hat{J}$ based on the data, such that the estimator based on $\hat{J}$ is almost as good as the estimator based on the optimal value of J. The analogy with the study of Model (1.3) then becomes obvious and we have shown that the problem of adaptation to some unknown smoothness for Hölderian regression functions amounts to what is generally called a problem of *model selection*, that is finding a procedure solely based on the data to choose one statistical model among a (possibly large) family of such models, the aim being to choose automatically a model which is close to optimal in the family for the problem at hand. Let us now describe this procedure.

We start with a finite collection of possible models $\{S_m, m \in \mathcal{M}_n\}$ for f, each S_m being a finite-dimensional linear subspace of $\mathbb{L}^2(\mathbb{R}^k)$. The family of models usually depends on n and the function f may or may not belong to one of them. Let us denote by $\hat{f}_m$ the least squares estimator for Model (1.1) based on the model class S_m. We look for a model selection procedure $\hat{m}$ with values in $\mathcal{M}_n$, based solely on the data and not on any *prior* assumption on f, such that the risk of the resulting procedure $\hat{f}_{\hat{m}}$ is almost as good as the risk of the best least squares estimator in the family. Therefore an ideal selection procedure $\hat{m}$ should look for an optimal trade-off between the bias term $\mathrm{d}_\mu^2(f, S_m)$ and the variance term $\sigma_2^2 \dim(S_m)/n$. Our aim is to find $\hat{m}$ such that

$$\mathbb{E}[\|f - \hat{f}_{\hat{m}}\|_n^2] \leq C \min_{m \in \mathcal{M}_n} \left\{ \mathrm{d}_\mu^2(f, S_m) + \sigma_2^2 \frac{\dim(S_m)}{n} \right\}, \tag{1.7}$$

which means that, up to the constant C, our estimator chooses an optimal model.

It is important to notice that an estimator which satisfies (1.7) has many interesting properties provided that the family of models S_m has been suitably chosen. In particular this estimator is adaptive in the minimax sense with respect to many well-known classes of smoothness. The connections between adaptation and model selection and the nice properties of any estimator $\hat{f}_{\hat{m}}$ satisfying (1.7) have been developed at length in Barron, Birgé and Massart (1999) Chapter 5 and many illustrations of potential applications of our results can be found there and in Birgé and Massart (1997). We shall content ourselves in the sequel with a limited number of applications and we refer the interested reader to those papers.

Our model selection criterion is closely related to the classical C_p criterion of Mallows (1973). For each model m we compute the normalized residual sum of squares: $\gamma_n(\hat{f}_m) = n^{-1}\sum_{i=1}^{n}[Y_i - \hat{f}_m(\vec{X}_i)]^2$ and we choose $\hat{m}$ in order to minimize among all models $m \in \mathcal{M}_n$ the penalized residual sum of squares $\gamma_n(\hat{f}_m) + \mathrm{pen}(m)$. Mallows' C_p criterion corresponds to $\mathrm{pen}(m) = 2\sigma_2^2 \dim(S_m)/n$. In this paper, we want to see how one needs to modify Mallows' C_p when the errors or the covariates are correlated.

There have been many studies concerning model selection based on Mallows' C_p or related penalized criteria like Akaike's or the BIC criterion for regressive and autoregressive models (see Akaike (1973; 1974), Shibata (1976; 1981), Li (1987), Polyak and Tsybakov (1992), among many others ...). A common characteristic of these results is their asymptotic character. Extensions of these penalized criteria for data-driven model selection procedures have been done in Barron (1991; 1993), Barron and Cover (1991) and Rissanen (1984). More recently, a general approach to model selection for various statistical frameworks including density estimation and regression has been developed in Birgé and Massart (1997) and Barron, Birgé and Massart (1999), with many applications to adaptive estimation. An original characteristic of their viewpoint is its non asymptotic feature. Unfortunately, their general approach imposes restrictions on the regression model (1.1), (e.g. the regression function needs to be bounded by some known quantity) which makes it unattractive for practical issues. We relax such restrictions and also obtain non asymptotic results. Our approach is inspired from Baraud's (2000) work. Although there have been many results concerning adaptation for the classical regression model with independent variables, not much is known to our knowledge concerning general adaptation methods for regression involving dependent variables. It is not within the scope of this paper to make an historical review for the case of independent variables.

Concerning dependent variables, it is worth mentioning the work of Modha and Masry (1996) which deals with Model (1.1) when the process $(\vec{X}_i, Y_i)_{i\in\mathbb{Z}}$ is strongly mixing and when the function f satisfies some Fourier-transform-type representation. In Modha and Masry (1998), the problem of one step ahead prediction of real valued stationary exponentially strongly mixing processes is considered. Minimum complexity regression estimators based on Legendre polynomials are used to estimate both the model memory and the predictor function. In the particular case of an autoregressive model their approach does not lead to optimal rates of convergence. In the case of a one dimensional first order autoregressive model, Neumann and Kreiss (1998) and Hoffmann (1999) study the behavior of nonparametric adaptive estimators (local polynomials and wavelet thresholding estimators) by approximating an AR(1) autoregression experiment by a regression experiment with independent variables.

Our estimation procedure is the same as that proposed by Baraud (2000) in the case of a regression framework with deterministic design points and i.i.d. errors. Thus, we show that the procedure is robust (to a certain extent) to possible dependency between the data $(\vec{X}_i, Y_i)$'s. More precisely, we assume that the data are β-mixing (for a precise definition of β-mixing, see Kolmogorov and Rozanov (1960)) and we show that under an adequate condition on the decay of the β-mixing coefficients (for instance arithmetical or geometrical decay) the estimation procedure is still relevant. Of course, this robustness with respect to dependency is obvious when the sequences of $\vec{X}_i$'s and ε_i's are independent and when the ε_i's are i.i.d.. Indeed, the result can merely be obtained by arguing as follows. We start from Inequality (11) in Baraud (2000, Corollary 3.1) which gives the result conditionally to the variables $\vec{X}_i$'s. Then, by integrating with respect to those, one gets (1.7). We emphasize that the result holds under mild assumptions on the statistical framework (an adequate moment condition on the i.i.d. errors and stationarity of the distribution of the $\vec{X}_i$'s). Consequently, we shall only consider either the case where the sequences of $\vec{X}_i$'s and ε_i's are dependent or the case where the ε_i's are dependent.

The case of β-mixing data is natural in the autoregression context, where, in addition, the above condition on the β-mixing coefficients is usually met. This makes the procedure of particular interest in this case.

Our techniques of proof are based on the work of Baraud (2000). Unfortunately, the possible dependency of the $\vec{X}_i$'s prevents us from directly using classical inequalities on product measures like Talagrand (1996) 's concentration inequalities. Taking advantage of the β-mixing assumptions, we instead use coupling techniques derived from Berbee's Lemma (1979) and inspired from Viennet's (1997) work in order to approximate the original sequence $(\vec{X}_i)_{1\leq i\leq n}$ by a new sequence built on independent blocks.

Lastly, we mention that the results presented in this paper can be extended to the case where the variance σ_2^2 of the errors is unknown, which is the practical case, by estimating it by residual least-squares. For further details we refer to Baraud's (1998) PhD thesis, where a previous version of this work is available.

The paper is organized as follows: the estimation procedure and the main assumptions are given in Section 2. We apply the procedure to various statistical frameworks in Section 3. In each of these frameworks, we state non asymptotic risk bounds, the proofs of those results being delayed to Section 6. Section 4 is devoted to the main result (treating the case of independent errors), Section 5 to an extension to the case of dependent errors. The proof of those results are given in Sections 7 to 10.

2. The estimation procedure and the assumptions We observe pairs $(Y_i, \vec{X}_i), i = 1, \ldots, n$ arising from Model (1.1) and our aim is to estimate the unknown function f from $\mathbb{R}^k$ into $\mathbb{R}$, on some (compact) subset $A \subset \mathbb{R}^k$.

Our estimation procedure is the following one. We consider a finite family of linear subspaces $\{S_m\}_{m\in\mathcal{M}_n}$ of $(\mathbb{L}^2(A, dx), \| \|)$. We assume that the S_m's are finite dimensional linear spaces consisting of A-compactly supported functions. Hereafter, D_m denotes the dimension of S_m and f_m the $\mathbb{L}^2(\mathbb{R}^k, \mu)$-projection of f onto S_m. We associate to each S_m a least-squares estimator $\hat{f}_m$ of f which minimizes among

$t \in S_m$ the empirical least-squares contrast function γ_n defined by (1.4). Note that such a minimizer might not be unique as an element of S_m but the $\mathbb{R}^n$-vector $(\hat{f}_m(\vec{X}_1), ..., \hat{f}_m(\vec{X}_n))'$ is uniquely defined. We select our estimator $\tilde{f}$ among the family of least-squares estimators $\{\hat{f}_m\}_{m \in \mathcal{M}_n}$ in the following way: given a nonnegative penalty function $\text{pen}(\cdot)$ on $\mathcal{M}_n$, we define $\hat{m}$ as the minimizer among $\mathcal{M}_n$ of the penalized criterion

$$\gamma_n(\hat{f}_m) + \text{pen}(m)$$

and we set $\tilde{f} = \hat{f}_{\hat{m}} \in S_{\hat{m}}$. The choice of the penalty function is the main concern of this paper.

The main assumptions used in the paper are listed below. Assumptions $(\mathbf{H}_\varepsilon)$ and $(\mathbf{H}_{X,\varepsilon})$ will be weakened in Section 5:

$(\mathbf{H}_X)$ The sequence $(\vec{X}_i)_{i \geq 0}$ is identically distributed with common law μ admitting a density h_X w.r.t. the Lebesgue measure which is bounded from below and above, i.e.

$$0 < h_0 \leq h_X(u) \leq h_1 \quad \forall u \in A.$$

$(\mathbf{H}_\varepsilon)$ The sequence ε_i's are i.i.d. centered random variables admitting a finite variance denoted by σ_2^2.

$(\mathbf{H}_{X,Y})$ The sequence of the $(\vec{X}_i, Y_i)$'s is β-mixing.

$(\mathbf{H}_{X,\varepsilon})$ For all $i \in \{1, ..., n\}$, ε_i is independent of the sequence $(\vec{X}_j)_{j \leq i}$.

$(\mathbf{H}_S)$ There exists a constant Φ_0 such that for any pair $(m, m') \in \mathcal{M}_n^2$, and any $t \in S_m + S_{m'}$

$$\|t\|_\infty \leq \Phi_0 \sqrt{\dim(S_m + S_{m'})} \|t\|. \tag{2.8}$$

Comments. Assumption $(\mathbf{H}_{X,Y})$ is equivalent to the β-mixing of the sequence of the $(\vec{X}_i, \varepsilon_i)$'s, which is the property which is used in the proof. As mentioned in the introduction, if the sequences $(\vec{X}_i)_{1 \leq i \leq n}$ and $(\varepsilon_i)_{1 \leq i \leq n}$ are independent and the ε_i's are i.i.d., then the result can be obtained under milder conditions. In particular, except stationarity, no other assumption on the distribution of the $\vec{X}_i$'s is required. Condition $(\mathbf{H}_S)$ is most easily fulfilled when the collection of models is nested, i.e. is an increasing sequence (for inclusion) of linear spaces and when there exists some Φ_0 such that for each $m \in \mathcal{M}_n$

$$\|t\|_\infty \leq \Phi_0 \sqrt{\dim(S_m)} \|t\|, \quad \forall t \in S_m. \tag{2.9}$$

This connection between the sup-norm and the $\mathbb{L}^2(A, dx)$-norm is satisfied for numbers of collections of models of interest. Birgé and Massart (1998), Lemma 6, have shown that for any $\mathbb{L}^2(A, dx)$-orthonormal basis $(\phi_\lambda)_{\lambda \in \Lambda(m)}$ of S_m:

$$\left\| \sum_{\lambda \in \Lambda(m)} \phi_\lambda^2 \right\|_\infty^{1/2} = \sup_{t \in S_m, t \neq 0} \frac{\|t\|_\infty}{\|t\|}. \tag{2.10}$$

Hence (2.9) holds if and only if there exists an orthonormal basis $(\phi_\lambda)_{\lambda\in\Lambda(m)}$ of S_m such that

$$\left\|\sum_{\lambda\in\Lambda(m)}\phi_\lambda^2\right\|_\infty^{1/2}\leq\Phi_0\sqrt{\dim(S_m)},\tag{2.11}$$

and then the result is true for any orthonormal basis of S_m.

3. Examples

3. Examples In the section we apply our estimation procedure to various statistical frameworks. In each framework, we give an example of a collection of models $\{S_m,\ m\in\mathcal{M}_n\}$ and for some $x>1$, choose the penalty term to be equal to

$$\text{pen}(m)=x^3\frac{D_m}{n}\sigma_2^2,\quad\forall m\in\mathcal{M}_n,$$

except in Section 3.3 where the penalty term is chosen in a different way. In each case, we give sufficient conditions for $\tilde{f}=\hat{f}_{\hat{m}}$ to achieve the best trade-off (up to a constant) between the bias and the variance term among the collection of estimators $\{\hat{f}_m,\ m\in\mathcal{M}_n\}$. Namely, we show that for any ρ in $]1,x[$

$$\mathbb{E}\left[\|f\mathbf{I}_A-\tilde{f}\|_n^2\right]\leq\left(\frac{x+\rho}{x-\rho}\right)^2\inf_{m\in\mathcal{M}_n}\left[\|f\mathbf{I}_A-f_m\|_\mu^2+2x^3\frac{D_m}{n}\sigma_2^2\right]+\frac{R}{n},\tag{3.12}$$

for some constant $R=R(\rho)$ to be specified. With no loss of generality we shall assume that $A=[0,1]^k$. Those results, proved in Section 6, derive from our main theorems which are to be found in Sections 4 and Section 5.

3.1. Autoregression framework

3.1. *Autoregression framework* We deal with a particular feature of the regression framework (1.1), the autoregression framework of order 1 given by

$$Y_i=X_i=f(X_{i-1})+\varepsilon_i,\quad i=1,...,n.\tag{3.13}$$

The process is initialized with some real valued random variable X_0.

We assume the following:

> $(\mathbf{H}_{AR1})$: The random variable X_0 is independent of the ε_i's. The ε_i's are i.i.d. centered random variables admitting a density, h_ε, with respect to the Lebesgue measure and satisfying $\sigma_2^2=\mathbb{E}[|\varepsilon_1|^2]<\infty$. The density h_ε is a positive bounded and continuous function and the function f satisfies for some $0\leq a<1$ and $b\in\mathbb{R}$
>
> $$\forall u\in\mathbb{R},\quad|f(u)|\leq a|u|+b.\tag{3.14}$$
>
> The sequence of the random variables X_i's is stationary of common law μ.

The existence of a stationary law μ derives from the assumptions on the ε_i's and f. To estimate f we use the collection of models given below.

Collection of piecewise polynomials: Let r be some positive integer and $m(n)$ the largest integer such that $r2^{m(n)}\leq n/\ln^3(n)$ i.e. $m(n)=\text{int}[\ln(n/\ln^3(n))/(r\ln(2))]$

(int$[u]$ denotes the integer part of u). Let $\mathcal{M}_n$ be the set of integers $\{0,\dots,m(n)\}$, for each $m\in\mathcal{M}_n$ we define S_m as the linear span of piecewise polynomials of degree less than r based on the dyadic grid $\{j/2^m,\ j=0,\dots,2^m-1\}\subset[0,1]$.

The result on $\tilde f$ is the following one.

Proposition 1. *Consider the autoregression framework (3.13) and assume that* $(\mathbf{H}_{AR1})$ *holds. If* $\sigma_p^p=\mathbb{E}[|\varepsilon_i|^p]<\infty$ *for some* $p>6$ *then (3.12) holds for some constant* R *that depends on* $p,x,\rho,h_\varepsilon,\sigma_p^2,r,\|f\mathbf{I}_A-\int f\mathbf{I}_A dx\|_\infty$.

To obtain results in probability on $\|f\mathbf{I}_A-\tilde f\|_n^2$, it is actually enough to assume $\mathbb{E}[|\varepsilon_i|^p]<\infty$ for some $p>2$, we refer to (4.29) and the comment given there.

3.2. *Regression framework* We give an illustration of Theorem 1 in case of regression with arithmetically β-mixing design points. Of course the case of autoregression with arithmetically β-mixing X_i's can be treated similarly. Let us consider the regression model

$$Y_i=f(X_i)+\varepsilon_i\quad i=1,...,n. \tag{3.15}$$

In this section, we consider a sequence ε_i for $i\in\mathbb{Z}$ and we take the X_i's to be generated by a standard time series model:

$$X_i=\sum_{k=0}^{+\infty}a_k\varepsilon_{i-1-2k}. \tag{3.16}$$

Then we make the following assumption

> $(\mathbf{H}_{Reg})$**:** The ε_i's are i.i.d. Gaussian random variables. The a_j's are such that $a_0=1$, $\sum_{j=0}^{+\infty}a_jz^{2j}\neq 0$ for all z with $|z|\le 1$ and for all $j\ge 1$, $|a_j|\le Cj^{-d}$ for some constants $C>0$ and $d>17$.

The value 17 as bound for d is certainly not sharp. The model (3.16) for the X_i's together with the assumptions on the coefficients a_j aim at ensuring that $(\mathbf{H}_{X,Y})$ is fulfilled with arithmetically β-mixing variables. Of course, any other model implying the same property would suit.
We introduce the following collection of models.

Collection of wavelets: For any integer J, let $\Lambda(j)=\{(j,k)/\ k=1,...,2^j\}$ and let

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

be an $\mathbb{L}^2([0,1],dx)$-orthonormal system of compactly supported wavelets of regularity r built by Cohen, Daubechies and Vial (1993). For some positive integer $J_n>J_0$, let $\mathcal{S}_n$ be the space spanned by the $\phi_{J_0,k}$'s for $(J_0,k)\in\Lambda(J_0)$ and by the $\{\varphi_{j,k}$'s for $(j,k)\in\cup_{J=J_0}^{J_n-1}\Lambda(J)\}$. The integer J_n is chosen in such a way that $\dim(\mathcal{S}_n)=2^{J_n}$ is of order $n^{4/5}/\ln(n)$. We set $\mathcal{M}_n=\{J_0,...,J_n-1\}$ and for each

$m \in \mathcal{M}_n$ we define S_m as the linear span of the $\phi_{J_0,k}$'s for $(J_0, k) \in \Lambda(J_0)$ and the $\varphi_{j,k}$'s for $(j,k) \in \cup_{J=J_0}^{m} \Lambda(J)$.

For a precise description and use of these wavelet systems, see Donoho and Johnstone (1998). These new functions derive from Daubechies' wavelets (1992) at the interior of $[0,1]$ and are boundary corrected at the "edges".

PROPOSITION 2. *Assume that* $\|f\mathbf{1}_A\|_\infty < \infty$ *and that for all* $m \in \mathcal{M}_n$*, the constant functions belong to* S_m*. If* $(\mathbf{H}_{Reg})$ *is satisfied, then (3.12) holds true for some constant* R *depending on* $x, \rho, h_0, h_1, \sigma_2^2, C, d, \|f\mathbf{1}_A - \int f\mathbf{1}_A dx\|_\infty$.

3.3. *Regression with dependent errors* We consider the regression framework

$$\begin{cases} Y_i = f(\vec{X}_i) + \varepsilon_i, & i = 1, ..., n \\ \varepsilon_i = a\varepsilon_{i-1} + u_i & i = 1, ..., n. \end{cases} \tag{3.17}$$

We observe the pairs $(Y_i, \vec{X}_i)$ for $i = 1, ..., n$.
We assume that

> $(\mathbf{H}_{Rd})$**:** The real number a satisfies $0 \le a < 1$, and the u_i's are i.i.d. centered random variables admitting a common finite variance. The law of the ε_i's is assumed to be stationary admitting a finite variance σ_2^2. The sequence of the $\vec{X}_i$'s is geometrically β-mixing (i.e. satisfying (6.31)) and the sequences of the $\vec{X}_i$'s and the ε_i's are independent.

Geometrically β-mixing $\vec{X}_i$'s can be generated by an autoregressive model with a regression function g and errors η_i satisfying an assumption of the same kind as $(\mathbf{H}_{AR1})$ in Section 3.1.

The main difference between this framework and the previous one lies in the dependency between the ε_i's. To deal with it, we need to modify the penalty term:

PROPOSITION 3. *Assume that* $\|f\mathbf{1}_A\|_\infty < \infty$*, that* $(\mathbf{H}_X)$ *and* $(\mathbf{H}_{Rd})$ *hold and that* $\mathbb{E}[|\varepsilon_1|^p] < \infty$ *for some* $p > 6$*. Let* $x > 1$*, if the penalty term* pen *satisfies*

$$\forall m \in \mathcal{M}_n, \quad \mathrm{pen}(m) \ge x^3 \left(1 + \frac{2a}{1-a}\right) \frac{D_m}{n} \sigma_2^2, \tag{3.18}$$

then by using the collection of piecewise polynomials described in Section 3.1 *and applying the estimation procedure given in Section 2 we have that the estimator* $\tilde{f}$ *satisfies for any* $\rho \in]1, x[$,

$$\mathbb{E}\left[\|f\mathbf{1}_A - \tilde{f}\|_n^2\right] \le \left(\frac{x+\rho}{x-\rho}\right)^2 \inf_{m \in \mathcal{M}_n} \left[\|f\mathbf{1}_A - f_m\|_\mu^2 + 2\mathrm{pen}(m)\right] + \frac{R}{n}, \tag{3.19}$$

where R *depends on* $a, p, \sigma_p, \|f\mathbf{1}_A - \int f\mathbf{1}_A dx\|_\infty, x, \rho, h_0, h_1, \Gamma, \theta$.

In contrast with the results of the previous examples, we cannot give a choice of a penalty term which would work for any value of a. An unknown lower bound for the choice of the penalty term seems to be the price to pay when the ε_i's are no longer independent. This example shows how this lower bound varies with respect to unknown number a, this number quantifying in some sense a discrepancy to independence (the independence corresponds to $a = 0$). We also see that a choice of the penalty term of the form

$$\text{pen}(m) = \kappa \frac{D_m}{n} \sigma_2^2$$

with κ large is safer than a choice of κ close to 1. This should be kept in mind every time the independence of the ε_i's is debatable (we refer the reader to the comments following Theorem 2).

3.4. *Additive models* We consider the additive regression models, widely used in Economics, described by

$$Y_i = e_f + f_1(X_i^{(1)}) + f_2(X_i^{(2)}) + \ldots + f_k(X_i^{(k)}) + \varepsilon_i \tag{3.20}$$

where the ε_i's are i.i.d. and e_f denotes a constant. Model (3.20) follows from Model (1.1) with $\vec{X}_i = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(k)})'$ and the additive function f: $f(x_1, \ldots, x_k) = e_f + f_1(x_1) + \ldots + f_k(x_k)$. For identifiability, we assume that $\int_{[0,1]} f_i(x)dx = 0$, for $i = 1, ..., k$. Such a model assumes that the effects on Y of the variables $X^{(j)}$ are additive. Our aim is to estimate f on $A = [0,1]^k$. The estimation method allows to build estimators of $f_1, \ldots, f_k$ in different spaces.

Let ℓ be some integer. We define $S_\ell^{(1)}$ as the linear space of piecewise polynomials t of degree less that r, $r \geq 1$, based on the dyadic grid $\{j/2^\ell, j = 0, \ldots, 2^\ell\} \subset [0,1]$, satisfying $\int_{[0,1]} t(x)\ dx = 0$ and $S_\ell^{(2)}$ as the linear span of the functions $\psi_{2j-1}(x) = \sqrt{2}\cos(2\pi jx)$ and $\psi_{2j}(x) = \sqrt{2}\sin(2\pi jx)$ for $j = 1, ..., 2^\ell$. Now we set $m_1(n)$ ($m_2(n)$ respectively) the largest integers ℓ such that $\dim(S_\ell^{(1)})$ ($\dim(S_\ell^{(2)})$ respectively) is smaller than $\sqrt{n}/\ln^3(n)$. Lastly $\mathcal{M}_n^{(1)}$ and $\mathcal{M}_n^{(2)}$ denote respectively the sets of integer $\{0, ..., m_1(n)\}$ and $\{0, ..., m_2(n)\}$.
We propose to estimate the f_i's either by piecewise or trigonometric polynomials. To do so, we introduce the choice function g from $\{1, ..., k\}$ into $\{1, 2\}$ and consider the following collections of models.

Mixed additive collection of models: We set $\mathcal{M}_n = \mathcal{M}_{k,n} = \{m = (k, m_1, ..., m_k),\ m_j \in \mathcal{M}_n^{(g(j))}\}$ and for each $m = (k, m_1, ..., m_k) \in \mathcal{M}_n$ we define

$$S_m = \left\{ t(x_1, ..., x_k) = a + \sum_{i=1}^{k} t_i(x_i),\ \ (a, t_1, ..., t_k) \in \mathbb{R} \times \prod_{i=1}^{k} S_{m_i}^{(g(i))} \right\}.$$

The performance of $\tilde{f}$ is given by the following result

PROPOSITION 4. *Assume that $\|f\mathbf{I}_A\|_\infty < \infty$, that the sequence of the $(\vec{X}_i, Y_i)$ is geometrically β-mixing, i.e. satisfies (6.31) and that $(\mathbf{H}_X)$, $(\mathbf{H}_\varepsilon)$ and $(\mathbf{H}_{X,\varepsilon})$ are fulfilled. Consider the additive regression framework (3.20) with the above collection of models. If $\sigma_p^p = \mathbb{E}[|\varepsilon|^p] < \infty$ for some $p > 6$, then $\tilde{f}$ satisfies (3.12) for some constant R depending on $k, p, \sigma_p, \|f\mathbf{I}_A - \int f\mathbf{I}_A dx\|_\infty, x, h_0, h_1, \Gamma, \theta$.*

We can deduce from Proposition 4 that our procedure is adaptive in the minimax sense. The point of interest is that the additive framework avoids the curse of dimensionality in the rate of convergence i.e. we can derive similar rates of convergence for $k \geq 2$ as for $k = 1$.

Let $\alpha > 0$ and $l > 2$, we recall that a function f from $[0, 1]$ into $\mathbb{R}$ belongs to the Besov space $\mathcal{B}_{\alpha,l,\infty}$ if it satisfies

$$|f|_{\alpha,l} = \sup_{y>0} y^{-\alpha} w_d(f, y)_l < +\infty, \quad d = [\alpha] + 1,$$

where $w_d(f, y)_l$ denotes the modulus of smoothness. For a precise definition of those notions we refer to DeVore and Lorentz (1993), Chapter 2, Section 7. Since for $l \geq 2$, $\mathcal{B}_{\alpha,l,\infty} \subset \mathcal{B}_{\alpha,2,\infty}$, we now restrict ourselves to the case where $l = 2$. In the sequel, for any $L > 0$ $\mathcal{B}_{\alpha,2,\infty}(L)$ denotes the set of functions which belong to $\mathcal{B}_{\alpha,2,\infty}$ and satisfy $|f|_{\alpha,2} \leq L$. Then the following result holds.

Proposition 5. *Consider Model (3.20) with $k \geq 2$. Let $L > 0$, assume that $\|f\mathbf{1}_A\|_\infty \leq L$ and that for all $i = 1, ..., k$, $f_i \in \mathcal{B}_{\alpha_i,2,\infty}(L)$ for some $\alpha_i > 1/2$. Assume that for all $i = 1, ..., k$ such that $g(i) = 1$, $\alpha_i \leq r$. Set $\alpha = \min\{\alpha_1, ..., \alpha_k\}$, if $\mathbb{E}[|\varepsilon_1|^p] < \infty$ for some $p > 6$ then under the assumptions of Proposition 4,*

$$\mathbb{E}\left[\|f\mathbf{1}_A - \tilde{f}\|_n^2\right] \leq C(k, L, \alpha, R) n^{-\frac{2\alpha}{2\alpha+1}}. \tag{3.21}$$

Comments.

- In the case where $k = 1$, by using the collection of piecewise polynomials described in Section 3.1, (3.21) holds under the weaker assumption that $\alpha > 0$, we refer the reader to the proof of Proposition 5.

- A result of the same flavor can be established in probability, this would require a weaker moment condition on the ε_i's. Namely, using (4.29) we show similarly that for any $\eta > 0$, there exists a positive constant $C(\eta)$ (also depending on k, L, α and R) such that

  $$\|f\mathbf{1}_A - \tilde{f}\|_n \leq C(\eta) n^{-\frac{\alpha}{2\alpha+1}},$$

  with probability greater or equal to $1 - \eta$, as soon as $\mathbb{E}[|\varepsilon_1|^p] < \infty$ for some $p > 2$.

3.5. *Estimation of the order of an additive autoregression* Consider an additive autoregression framework,

$$X_i = e_f + f_1(X_{i-1}) + f_2(X_{i-2}) + \ldots + f_k(X_{i-k}) + \varepsilon_i \tag{3.22}$$

where the ε_i's are i.i.d. and e_f denotes a constant. Under suitable assumptions ensuring that the $\vec{X}_i = (X_{i-1}, ..., X_{i-k})'$'s are stationary and geometrically β-mixing, the estimation of f_1,...,f_k can be handled in the same way as in the previous section. The aim of this section is to provide an estimator of the order of autoregression, i.e. an estimator of the integer k_0 ($k_0 \leq k$, k being known) satisfying $f_{k_0} \neq 0$ and $f_i = 0$ for all $i > k_0$. To do so, let $\mathcal{M}_n = \bigcup_{j=0}^{k} \mathcal{M}_{j,n}$ (we use the notations introduced in

Section 3.4) and consider the collection of models $\{S_m,\ m \in \mathcal{M}_n\}$. We estimate k_0 by $\hat{k}_0 = \hat{k}_0(x)$ defined as the first coordinate of $\hat{m}$, $\hat{m}$ being given by

$$\hat{m} = \arg\min_{m \in \mathcal{M}_n} \left[\gamma_n(\hat{f}_m) + x^3 \frac{D_m}{n} \sigma_2^2 \right].$$

We measure the performance of $\hat{k}_0$ via that of $\tilde{f} = \hat{f}_{\hat{m}}$, the latter being known, under the assumptions of Theorem 1, to achieve the best trade-off (up to a constant) between the bias term and the variance term among the collections of least-squares estimators $\{\hat{f}_m,\ m \in \mathcal{M}_n\}$.

4. The main result In this section, we give our main result concerning the estimation of a regression function from dependent data. Although this result considers the case of particular collections of models, extension including very general collections are to be found in the comments following the theorem.

4.1. *The main theorem* Let $\mathcal{S}_n$ be some finite dimensional linear subspace of A-supported functions of $\mathbb{L}^2(\mathbb{R}^k, dx)$. Let $\{\phi_\lambda\}_{\lambda \in \Lambda_n}$ be an orthonormal basis of $\mathcal{S}_n \subset \mathbb{L}^2(A, dx)$ and set $D_n = |\Lambda_n| = \dim(\mathcal{S}_n)$. We assume that there exists some positive constant $\Phi_1 \geq 1$ such that for all $\lambda \in \Lambda_n$

$$(\mathbf{H}_{\mathcal{S}_n}) \quad \|\phi_\lambda\|_\infty \leq \Phi_1 \sqrt{D_n} \quad \text{and} \quad |\{\lambda' / \ \|\phi_{\lambda'}\phi_\lambda\|_\infty \neq 0\}| \leq \Phi_1.$$

The second condition means that for each λ, the supports of ϕ_λ and $\phi_{\lambda'}$ are disjoint except for at most Φ_1 functions $\phi_{\lambda'}$'s. We shall see in Section 10 that those conditions imply that (2.9) holds with $\Phi_0^2 = \Phi_1^3$. In addition we assume some constraint on the dimension of $\mathcal{S}_n$

$(\mathbf{H}_{D_n})_{(\Psi,b)}$ There exists an increasing function Ψ mapping $\mathbb{R}_+$ into $\mathbb{R}_+$ satisfying for some $K > 0$ and $b \in]0, 1/4[$

$$\forall u \geq 1, \qquad \ln(u) \vee 1 \leq \Psi(u) \leq K u^b,$$

such that

$$D_n \leq \frac{n}{\Psi(n)\ln(n)}. \tag{4.23}$$

Theorem 1. *Let us consider Model (1.1) with f an unknown function from $\mathbb{R}^k$ into $\mathbb{R}$ such that $\|f\mathbf{I}_A\|_\infty < \infty$ and where Conditions $(\mathbf{H}_X)$, $(\mathbf{H}_\varepsilon)$ and $(\mathbf{H}_{X,\varepsilon})$ are fulfilled. Consider a family $\{S_m\}_{m \in \mathcal{M}_n}$ of linear subspaces of $\mathcal{S}_n$. Assume that $\{S_m\}_{m \in \mathcal{M}_n}$ satisfies $(\mathbf{H}_S)$ and that $\mathcal{S}_n$ satisfies $(\mathbf{H}_{\mathcal{S}_n})$ and $(\mathbf{H}_{D_n})_{(\Psi,b)}$. Suppose that $(\mathbf{H}_{X,Y})$ is fulfilled for a sequence of β-mixing coefficients satisfying*

$$\forall q \geq 1, \ \ \beta_q \leq M \left[\Psi^{-1}(Bq)\right]^{-3}, \tag{4.24}$$

for some $M > 0$ and for some constant B (given by (7.46)). For any $x > 1$, let pen *be a penalty function such that*

$$\forall m \in \mathcal{M}_n, \ \ \text{pen}(m) \geq x^3 \frac{D_m}{n} \sigma_2^2.$$

Let $\rho \in]1, x[$, for any $\bar{p} \in]0, 1]$, if there exists $p > p_0 = 2(1 + 2\bar{p})/(1 - 4b)$ such that $\sigma_p^p = \mathbb{E}[|\varepsilon_1|^p] < \infty$, we have that the PLSE $\tilde{f}$ defined by

$$\tilde{f} = \arg\min_{m \in \mathcal{M}_n} \left\{ \gamma_n(\hat{f}_m) + \mathrm{pen}(m) \right\} \text{ with } \quad \gamma_n(g) = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - g(\vec{X}_i) \right]^2 \tag{4.25}$$

satisfies

$$\begin{aligned} &\left(\mathbb{E} \left[\| f \mathbf{I}_A - \tilde{f} \|_n^{2\bar{p}} \right] \right)^{1/\bar{p}} \\ &\leq \left(\frac{x + \rho}{x - \rho} \right)^2 \inf_{m \in \mathcal{M}_n} \left[| f \mathbf{I}_A - f_m \|_\mu^2 + 2\mathrm{pen}(m) \right] + C \frac{R_n}{n} \end{aligned} \tag{4.26}$$

where C is a constant depending on $p, x, \rho, \bar{p}, \Phi_0, h_0, h_1, M, K$ and R_n is given by

$$R_n^{\bar{p}} = \sigma_p^{2\bar{p}} \left[\sum_{m \in \mathcal{M}_n} D_m^{-p/2 + \bar{p}} + \frac{|\mathcal{M}_n|}{n^{(1/4 - b)(p - p_0)}} + \frac{\| f \mathbf{I}_A \|_\infty^{2\bar{p}}}{\sigma_p^{2\bar{p}}} \right]. \tag{4.27}$$

Comments.

1. The functions Ψ of particular interest are either of the form $\Psi(u) = \ln(u)$ or $\Psi(u) = u^c$ with $0 < c < 1/4$. In the first case, (4.24) is equivalent to a geometric decay of the β-mixing coefficients (then, we say that the variables are geometrically β-mixing), in the second case (4.24) is equivalent to an arithmetic decay (the sequence is then arithmetically β-mixing).

2. A choice of D_n small in front of n allows to deal with stronger dependency between the $(Y_i, \vec{X}_i)$'s. In return, choosing D_n too small may lead to a serious drawback with regard to the performance of the PLSE. Indeed, in the case of nested models, the smaller D_n the smaller the collection of models and the poorer the performance of $\tilde{f}$.

3. Assumption $(\mathbf{H}_{\mathcal{S}_n})$ is fulfilled when $\mathcal{S}_n$ is generated by piecewise polynomials of degree r on $[0, 1]$ (in that case $\Phi_1 = 2(r + 1)$ suits) or by wavelets as those described in Section 3.2 (a suitable basis is obtained by rescaling the father wavelets $\phi_{J_0,k}$'s).

4. We shall see in Section 10 that the result of Theorem 1 holds for a larger class of linear spaces $\mathcal{S}_n$ (i.e. for $\mathcal{S}_n$'s which do not verify $(\mathbf{H}_{\mathcal{S}_n})$), provided that (4.23) is replaced by

$$D_n^2 \leq \frac{n}{\ln(n)\Psi(n)}. \tag{4.28}$$

5. Take $\bar{p} = 1$, the main term involved in the right-hand side of (4.26) is usually

$$\inf_{m \in \mathcal{M}_n} \left[\| f \mathbf{I}_A - f_m \|_\mu^2 + 2\mathrm{pen}(m) \right].$$

   It is worth noticing that the constant in front of this term, i.e.

$$C_1(x, \rho) = \left(\frac{x + \rho}{x - \rho} \right)^2$$

only depends on x and ρ, and not on unpleasant quantities such as h_0, h_1. If Theorem 1 gives no precise recommendation on the choice of x to optimize the performance of the PLSE, it suggests, in contrast, that a choice of x close to 1 is certainly not a good choice since it makes the constant $C_1(x,\rho)$ blow up (we recall that ρ must belong to $]1,x[$). Fix ρ, we see that $C_1(x,\rho)$ decreases to 1 as x becomes large; the negative effect of choosing x large being that it increases the value of the penalty term.

6. Why does Theorem 1 give a result for values of $\bar{p} \neq 1$? By using Markov's inequality, we can derive from (4.26) a result in probability saying that for any $\tau > 0$,
$$\mathbb{P}\left[\|f\mathbf{1}_A - \tilde{f}\|_n^2 > \tau\left(\inf_{m\in\mathcal{M}_n}\left[\|f\mathbf{1}_A - f_m\|_\mu^2 + 2\mathrm{pen}(m)\right] + \frac{R_n}{n}\right)\right] \leq \frac{C'}{\tau^{\bar{p}}} \tag{4.29}$$
where C' depends on $x, \rho, \bar{p}, C$. If $\mathbb{E}[|\varepsilon_1|^p] < \infty$ for some $p > 2$ and if it is possible to choose $\Psi(u)$ of order a power of $\ln(u)$ (this is the case when the $(Y_i, \vec{X}_i)$'s are geometrically β-mixing) then one can choose both b in $(\mathbf{H}_{D_n})_{(\Psi,b)}$ and $\bar{p}$ small enough to ensure that $p > 2(1+2\bar{p})/(1-4b)$. Consequently we get that (4.29) holds true under the weak assumption that $\mathbb{E}[|\varepsilon_1|^p] < \infty$ for some $p > 2$. Lastly we mention that an analogue of (4.29) where $\|f\mathbf{1}_A - \tilde{f}\|_n^2$ is replaced by $\|f\mathbf{1}_A - \tilde{f}\|_\mu^2$ can be obtained. This can be derived from the fact that, under the assumptions of Theorem 1, the (semi)norms $\|\ \|_\mu$ and $\|\ \|_n$ are equivalent on $\mathcal{S}_n$ on a set of probability close to 1 (we refer to the proof of Theorem 1 and for further details to Baraud (2001)).
7. For adequate collection of models, the quantity R_n remains bounded by some number R not depending on n. In addition, if for all $m \in \mathcal{M}_n$, the constants belong to S_m, then the quantity $\|f\mathbf{1}_A\|_\infty$ involved in R_n can be replaced by the smaller one $\|f\mathbf{1}_A - \int f\mathbf{1}_A\|_\infty$.

5. Generalization of Theorem 1 In this Section we give an extension of Theorem 1 by relaxing the independence of the ε_i's and by weakening Assumption $(\mathbf{H}_{X,\varepsilon})$. In particular, the next result shows that the procedure is robust to possible dependency (to some extent) of the ε_i's.
We assume that

(H'$_\varepsilon$) The ε_i's satisfy for some positive number ϑ
$$\sup_{t, \|t\|_\mu \leq 1} \mathbb{E}\left[\left(\sum_{i=1}^q \varepsilon_i t(\vec{X}_i)\right)^2\right] \leq q\vartheta \tag{5.30}$$
for any $1 \leq q \leq n$.

In addition Assumption $(\mathbf{H}_{X,\varepsilon})$ is replaced by the milder one

(H'$_{X,\varepsilon}$) For all $i \in \{1,...,n\}$, $\vec{X}_i$ and ε_i are independent.

Then the following result holds

THEOREM 2. *Consider the assumptions of Theorem 1 and replace* $(\mathbf{H}_\varepsilon)$ *by* $(\mathbf{H'}_\varepsilon)$ *and* $(\mathbf{H}_{X,\varepsilon})$ *by* $(\mathbf{H'}_{X,\varepsilon})$*. For any* $x > 1$*, let* pen *be a penalty function such that*

$$\forall m \in \mathcal{M}_n, \quad \mathrm{pen}(m) \geq x^3 \frac{D_m}{n} \vartheta.$$

Then, the result (4.26) of Theorem 1 holds for a constant C *that also depends on* ϑ*.*

COMMENTS.

- In the case of i.i.d. ε_i's and under Assumption $(\mathbf{H}_{X,\varepsilon})$ (which clearly implies $(\mathbf{H'}_{X,\varepsilon})$), it is straightforward that (5.30) holds with $\vartheta = \sigma_2^2$. Indeed under Condition $(\mathbf{H}_{X,\varepsilon})$, for all $t \in \mathbb{L}^2(\mathbb{R}^k, \mu)$

$$\mathbb{E}\left[\left(\sum_{i=1}^q \varepsilon_i t(\vec{X}_i)\right)^2\right] = \sum_{i=1}^q \mathbb{E}\left[\varepsilon_i^2 t^2(\vec{X}_i)\right] + 0 = q\sigma_2^2 \|t\|_\mu^2.$$

  Then, we recover Theorem 1.

- Assume that the sequences $(\vec{X}_i)_{i=1,\ldots,n}$ and $(\varepsilon_i)_{i=1,\ldots,n}$ are independent (which clearly implies $(\mathbf{H'}_{X,\varepsilon})$) and that the ε_i's are β-mixing. Then, we know from Viennet (1997) that there exists a function d_β depending on the β-mixing coefficients of the ε_i's such that for all $t \in \mathbb{L}^2(\mathbb{R}^k, \mu)$

$$\mathbb{E}\left[\left(\sum_{i=1}^q \varepsilon_i t(\vec{X}_i)\right)^2\right] \leq q\mathbb{E}\left[\varepsilon_1^2 d_\beta(\varepsilon_1)\right] \|t\|_\mu^2,$$

  which amounts to taking $\vartheta = \vartheta(\beta) = \mathbb{E}\left[\varepsilon_1^2 d_\beta(\varepsilon_1)\right]$ in (5.30). Roughly speaking $\vartheta(\beta)$ is close to σ_2^2 when the β-mixing coefficients of the ε_i's are close to 0 which corresponds to the independence of the ε_i's. Thus, in this context the result of Theorem 2 can be understood as a result of robustness, since $\vartheta(\beta)$ is unknown. Indeed, the penalized procedure described in Theorem 1 with a penalty term satisfying, for some $\kappa > 1$,

$$\forall m \in \mathcal{M}_n, \quad \mathrm{pen}(m) \geq \kappa \frac{D_m}{n} \sigma_2^2,$$

  still works if $\vartheta(\beta) < \kappa\sigma_2^2$. This also means that if the independence of the ε_i's is debatable, it is safer to increase the value of the penalty term.

6. Proof of the propositions of Section 3

6.1. *Proof of Proposition 1.* The result is a consequence of Theorem 1. Let us show that under $(\mathbf{H}_{AR1})$ the assumptions of Theorem 1 are fulfilled. Condition $(\mathbf{H}_\varepsilon)$ is direct. Under (3.14) it is clear that $\|f\mathbf{1}_{[0,1]}\|_\infty < \infty$ holds true. We now set $\mathcal{S}_n = S_{m(n)}$ and $\Psi(x) = \ln^2(x)$. Since

$$\dim(\mathcal{S}_n) = D_n \leq \frac{n}{\ln^3(n)},$$

$(\mathbf{H}_{D_n})_{(\Psi,b)}$ holds for any $b > 0$ and for some constant $K = K(b)$. As to Conditions $(\mathbf{H}_S)$ and $(\mathbf{H}_{\mathcal{S}_n})$, they hold with $\Phi_0 = r$ (we refer to Birgé and Massart (1998)). Under Condition (3.14), we know from Duflo (1997) that the process $(X_i)_{i\in\mathbb{N}}$ admits a stationary law μ. Furthermore, we know that if the ε_i's admit a positive bounded continuous density with respect to the Lebesgue measure then so does μ. This can easily be deduced from the connection between h_X and h_ε given by

$$h_X(y) = \int h_\varepsilon(y - f(x))h_X(x)dx \quad \forall y \in \mathbb{R}.$$

Then we can derive the existence of positive numbers h_1 and h_0 bounding the density h_X from above and below on $[0,1]$ and thus $(\mathbf{H}_X)$ is true. In addition we know from Doukhan (1994) that under (3.14) the X_i's are geometrically β-mixing i.e. there exist two positive constant Γ, θ such that

$$\beta_q \le \Gamma e^{-\theta q}, \quad \forall q \ge 1. \tag{6.31}$$

Since $\Psi^{-1}(u) = \exp(\sqrt{u})$, clearly there exists some constant $M = M(\Gamma,\theta) > 0$ such that

$$\beta_q \le \Gamma e^{-\theta q} \le M e^{-3\sqrt{Bq}}, \quad \forall q \ge 1.$$

Lastly, the ε_i's being independent of the sequence $(X_j)_{j<i}$, $(\mathbf{H}_{X,\varepsilon})$ is true and we know that the β-mixing coefficients of both sequences $(X_{i-1},\varepsilon_i)_{i=1,\dots,n}$ and $(X_{i-1})_{i=1,\dots,n}$ are the same. Consequently, Condition $(\mathbf{H}_{X,Y})$ holds and (4.24) is fulfilled. By choosing $\bar{p} = 1$, Theorem 1 can be applied if $\mathbb{E}[|\varepsilon_i|^p] < \infty$ for some $p > 6/(1-4b)$. This is true for b small enough and then (3.12) follows from (4.26) with

$$\begin{aligned} R_n &= \sigma_p^2\left[\sum_{m\in\mathcal{M}_n} D_m^{-p/2+1} + \frac{|\mathcal{M}_n|}{n^{(1/4-b)(p-6/(1-4b))}} + \frac{\|f\mathbf{1}_{[0,1]}\|_\infty^2}{\sigma_p^2}\right] \\ &\le \sigma_p^2\left[\sum_{m=0}^{+\infty}(r2^m)^{-2} + \sup_{n\ge1}\frac{\ln(n)}{n^{(1/4-b)(p-6/(1-4b))}} + \frac{\|f\mathbf{1}_{[0,1]}\|_\infty^2}{\sigma_p^2}\right] \\ &= R'. \end{aligned}$$

Take $R = CR'$ where C is the constant involved in (4.26) to finish the proof of Proposition 1. □

6.2. *Proof of Proposition 2.* Conditions $(\mathbf{H}_S)$ and $(\mathbf{H}_{\mathcal{S}_n})$ are fulfilled (we refer to Birgé and Massart (1998)). Next we check that $(\mathbf{H}_{X,Y})$ holds true and more precisely that the sequence $(\varepsilon_i, X_i)_{1\le i\le n}$ is arithmetically β-mixing with β-mixing coefficients satisfying

$$\forall q \in \{1,...,n\}, \quad \beta_q \le \Gamma q^{-\theta}, \tag{6.32}$$

for some constants $\Gamma > 0$ and $\theta > 15$. For that purpose, simply write $(\varepsilon_t, X_t)' = \sum_{j=0}^{\infty} A_j e(t-j)$ with $e(t-j) = (\varepsilon_{t-2j}, \varepsilon_{t-1-2j})'$, for $j \ge 0$, A_0 is the 2×2-identity matrix and $A_j = \begin{pmatrix} 0 & 0 \\ 0 & a_j \end{pmatrix}$. Then Pham and Tran's (1985) Theorem 2.1

implies under $(\mathbf{H}_{Reg})$, that (ε_t, X_t) is absolutely regular with coefficients $\beta_n \leq K\sum_{j=n}^{+\infty}\left(\sum_{k\geq j}|a_k|\right) \leq (KC)/((d-1)(d-2))n^{-d+2}$. This implies (6.32) with $\theta = d-2 > 15$. In addition, it can be proved that if $a_j = j^{-d}$ then $\beta_n \geq C(d)n^{-d}$, which shows that we do not reach the geometrical rate of mixing.
Clearly the other assumptions of Theorem 1 are satisfied and it remains to apply it with $p = 30$ (a moment of order 30 exists since the ε_i's are gaussian), $\Psi(u) = u^{1/5}$ and $\bar{p} = 1$. An upper bound for R_n which is does not depend on n can be established in the same way as in the proof of Proposition 1. □

6.3. *Proof of Proposition 3.* The line of proof is similar to that of Proposition 1, the difference lying in the fact that we need to check the assumptions of Theorem 2. Most of them are clearly fulfilled, we only check $(\mathbf{H}_{X,Y})$ and $(\mathbf{H'}_{\varepsilon})$. We note that the pairs $(\vec{X}_i, Y_i)$'s are geometrically β-mixing (which shows that $(\mathbf{H}_{X,Y})$ holds true) since both sequences X_i's and ε_i's are geometrically β-mixing (since the ε_i's are drawn from a "nice" autoregression model, we refer to Section 3.1) and are independent. Next we show that $(\mathbf{H'}_{\varepsilon})$ holds true with $\vartheta = (1+2a/(1-a))\sigma_2^2$. This will end the proof of Proposition 3. For all $t \in \mathbb{L}^2(\mathbb{R}^k, \mu)$,

$$\mathbb{E}\left[\left(\sum_{i=1}^{q}\varepsilon_i t(\vec{X}_i)\right)^2\right] \leq \sum_{i=1}^{q}\|t\|_\mu^2\sigma_2^2 + 2\sum_{i<j}\mathbb{E}[\varepsilon_i\varepsilon_j]\mathbb{E}[t(\vec{X}_i)t(\vec{X}_j)].$$

For $i < j$,

$$\begin{aligned}\mathbb{E}[\varepsilon_i\varepsilon_j] &= \mathbb{E}\left[\varepsilon_i(u_j + ... + a^k u_{j-k} + ... + a^{j-i-1}u_{i-1} + a^{j-i}\varepsilon_i)\right]\\ &= 0 + a^{j-i}\sigma_2^2,\end{aligned}$$

thus

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{q}\varepsilon_i t(\vec{X}_i)\right)^2\right] &\leq q\|t\|_\mu^2\sigma_2^2 + 2\sum_{i<j}a^{j-i}\mathbb{E}[t(\vec{X}_i)t(\vec{X}_j)]\sigma_2^2\\ &\leq \left(q + 2\sum_{1\leq i<j\leq q}a^{j-i}\right)\|t\|_\mu^2\sigma_2^2,\end{aligned}$$

by Cauchy-Schwarz's inequality. Therefore, we obtain

$$\mathbb{E}\left[\left(\sum_{i=1}^{q}\varepsilon_i t(\vec{X}_i)\right)^2\right] \leq q\left(1 + \frac{2a}{1-a}\right)\|t\|_\mu^2\sigma_2^2,$$

which gives the result. □

6.4. *Proof of Proposition 4.* Proposition is a consequence of Theorem 1. It is enough to apply it with $\bar{p} = 1$. In the sequel, we check that the assumptions of the theorem are fulfilled and we bound R_n (given by (4.27)) by some constant that does not depend on n. To bound the β-mixing coefficients of the sequence of the (Y_i, X_i)'s, we argue as in the proof of Proposition 1, with $\mathcal{S}_n = S_{(m_{g(1)}(n),\ldots,m_{g(k)}(n))}$,

$\dim(S_{(m_{g(1)}(n),\ldots,m_{g(k)}(n))}) \le \sqrt{n}/\ln^3(n)$ and $\Psi(n) = \ln^2(n)$. Inequality (4.28) is verified (thus condition $(\mathbf{H}_{\mathcal{S}_n})$ can be omitted). Let us now check $(\mathbf{H}_S)$. Since for all $m, m' \in \mathcal{M}_n$, $S_m + S_{m'}$ and $\mathcal{S}_n$ belong to the collection of models $\{S_m,\ m \in \mathcal{M}_n\}$, the assumption $(\mathbf{H}_S)$ holds true if we prove that (2.9) is satisfied for any S_m, $m \in \mathcal{M}_n$. Now note that for each $m \in \mathcal{M}_n$, the following decomposition in $\mathbb{L}^2([0,1]^k, dx_1...dx_k)$ holds

$$S_m = \mathbb{R}.\mathbf{1} \overset{\perp}{\oplus} S_m^{(1)} \overset{\perp}{\oplus} ... \overset{\perp}{\oplus} S_m^{(k)},$$

where $S_m^{(i)} = \{t \in S_m,\ t(x_1,...,x_k) = t_i(x_i)\}$ and $\mathbf{1}$ denotes the constant function on $[0,1]^k$. Clearly, $S_m^{(i)}$ satisfies (2.9) if and only if $S_{m_i}^{(g(i))}$ does, which is true. Now the fact that the S_m's satisfy (2.9) is a consequence of this lemma

Lemma 1. *Let $S^{(1)}$,..., $S^{(k)}$ be k linear spaces which are piecewise orthogonal in $\mathbb{L}^2([0,1]^k, dx_1 \ldots dx_k)$. If for each $i = 1, \ldots, k$, $S^{(i)}$ satisfies (2.9), then so does $S = S^{(1)} + ... + S^{(k)}$.*

Proof. The result follows from a Cauchy Schwarz argument: for all $t_i \in S^{(i)}$, $i = 1, ..., k$,

$$\|\sum_{i=1}^k t_i\|_\infty \le \sum_{i=1}^k \|t_i\|_\infty \le \Phi_0 \left(\sum_{i=1}^k \sqrt{\dim(S^{(i)})}\|t_i\|\right)$$
$$\le \Phi_0 \left(\sum_{i=1}^k \dim(S^{(i)})\right)^{1/2} \left(\sum_{i=1}^k \|t_i\|^2\right)^{1/2} = \Phi_0 \sqrt{\dim(S)}\|\sum_{i=1}^k t_i\|.$$

□

To finish the proof of Proposition 4 we bound R_n by some constant R which does not depend on n. Note that $|\mathcal{M}_n|$ is of order a power of $\ln(n)$ so the point is to show that $\sum_{m \in \mathcal{M}_n} D_m^{-2}$ (we recall that $\bar{p} = 1$ and $p > 6$) remains bounded by some quantity which does not depend on n. Now for each $m = (k, m_1, ..., m_k) \in \mathcal{M}_n$ we have that D_m is of order $2^{m_1} + ... + 2^{m_k}$, thus by using the convexity inequality $k^{-1}(a_1 + ... + a_k) \ge (a_1 ... a_k)^{1/k}$ which holds for any positive numbers $a_1, ..., a_k$, we obtain that $\sum_{m \in \mathcal{M}_n} D_m^{-2}$ is bounded (up to a constant) by

$$\sum_{m_1=0}^\infty ... \sum_{m_k=0}^\infty (2^{m_1} + ... + 2^{m_k})^{-2} \le \sum_{m_1=0}^\infty ... \sum_{m_k=0}^\infty 2^{-2(m_1+...+m_k)/k}$$
$$= \left(\sum_{j=0}^\infty 2^{-2j/k}\right)^k = R < \infty.$$

□

6.5. *Proof of Proposition 5.* Let $k \geq 2$. We start from (3.12) and bound the bias term. Let $f_m^{'}$ the $\mathbb{L}^2([0,1], dx)$ projection of f onto S_m, we have that $\|f\mathbf{1}_A - f_m\|_\mu^2 \leq \|f\mathbf{1}_A - f_m^{'}\|_\mu^2 \leq h_1\|f - f_m^{'}\|^2$ by $(\mathbf{H}_X)$ and for each $m = (k, m_1, ..., m_k)$,

$$\|f - f_m^{'}\|^2 = \sum_{i=1}^{k} \int_{[0,1]} \left(f_i(x) - f_{m,i}^{'}(x)\right)^2 dx,$$

where $f_{m,i}^{'}$ denotes the $\mathbb{L}^2([0,1], dx)$ projection of f_i onto $S_{m_i}^{(g(i))}$. Lastly we use standard results of approximation theory (see Barron, Birgé and Massart (1999), Lemma 13 or DeVore and Lorentz (1993)) which ensure that $\int_{[0,1]} \left(f_i(x) - f_{m,i}^{'}(x)\right)^2 dx \leq C(\alpha_i, L)D_{m_i}^{-2\alpha_i}$ (if $g(i) = 1$, this holds true in the case of piecewise polynomials since $r \geq \alpha_i$). We obtain (3.21) by taking for each $i = 1, ..., k$, $m_i \in \mathcal{M}_n^{(i)}$ such that D_{m_i} is of order $n^{1/(2\alpha_i+1)}$ which is possible since $\alpha_i > 1/2$ and therefore $n^{1/(2\alpha_i+1)} \leq D_n$ (at least for n large enough). In the one dimensional case, by considering the piecewise polynomials described in Section 3.1, D_n is of order $n/\ln^3(n)$ (such a choice is possible in this case) and then a choice of m among $\mathcal{M}_n$ such that D_m is of order $n^{1/(2\alpha+1)}$ is possible for any $\alpha > 0$. □

7. Proofs of Theorems 1 and 2 The proof of Theorem 2 is clear from the proof of Theorem 1. Indeed the assumptions $(\mathbf{H}_{X,\varepsilon})$ and $(\mathbf{H}_\varepsilon)$ are only needed in (8.53) and (8.56). For the rest of the proof assuming $(\mathbf{H'}_{X,\varepsilon})$ is enough. It remains to notice that an analogue of (8.53) and (8.56) is easily obtained from Assumption $(\mathbf{H'}_\varepsilon)$.
Now we prove Theorem 1. The proof is divided in consecutive claims.

Claim 1. $\forall m \in \mathcal{M}_n$,

$$\|f\mathbf{1}_A - \tilde{f}\|_n^2 \leq \|f\mathbf{1}_A - f_m\|_n^2 + \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i(\tilde{f} - f_m)(\vec{X}_i) + \text{pen}(m) - \text{pen}(\hat{m}). \tag{7.33}$$

Proof of Claim 1. By definition of $\tilde{f}$ we know that for all $m \in \mathcal{M}_n$ and $t \in S_m$

$$\gamma_n(\tilde{f}) + \text{pen}(\hat{m}) \leq \gamma_n(t) + \text{pen}(m).$$

In particular this holds for $t = f_m$ and algebraic computations lead to

$$\|f - \tilde{f}\|_n^2 \leq \|f - f_m\|_n^2 + \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i(\tilde{f} - f_m)(\vec{X}_i) + \text{pen}(m) - \text{pen}(\hat{m}). \tag{7.34}$$

Note that the relation

$$\|f - t\|_n^2 = \|f\mathbf{1}_A - t\|_n^2 + \|f - f\mathbf{1}_A\|_n^2$$

is satisfied for any A-supported function t. Applying this identity respectively to $t = \tilde{f}$ and $t = f_m$ (those functions being A-supported as elements of $\bigcup_{m' \in \mathcal{M}_n} S_{m'}$), we derive (7.33) from (7.34). □

Claim 2. *Let q_n, $q_{n,1}$ be integers such that $0 \leq q_{n,1} \leq q_n/2$, $q_n \geq 1$. Set $u_i = (\varepsilon_i, \vec{X}_i)$, $i = 1, ..., n$, then there exist random variables $u_i^* = (\varepsilon_i^*, \vec{X}_i^*)$, $i = 1, ..., n$ satisfying the following properties:*

- *For $\ell = 1, ..., \ell_n = [n/q_n]$, the random vectors*
$$\vec{U}_{\ell,1} = \left(u_{(\ell-1)q_n+1}, ..., u_{(\ell-1)q_n+q_{n,1}}\right)' \text{ and } \vec{U}^*_{\ell,1} = \left(u^*_{(\ell-1)q_n+1}, ..., u^*_{(\ell-1)q_n+q_{n,1}}\right)'$$
have the same distribution, and so have the random vectors
$$\vec{U}_{\ell,2} = \left(u_{(\ell-1)q_n+q_{n,1}+1}, ..., u_{\ell q_n}\right)' \text{ and } \vec{U}^*_{\ell,2} = \left(u^*_{(\ell-1)q_n+q_{n,1}+1}, ..., u^*_{\ell q_n}\right)'.$$
- *For $\ell = 1, ..., \ell_n$,*
$$\mathbb{P}\left[\vec{U}_{\ell,1} \neq \vec{U}^*_{\ell,1}\right] \leq \beta_{(q_n - q_{n,1})} \text{ and } \mathbb{P}\left[\vec{U}_{\ell,2} \neq \vec{U}^*_{\ell,2}\right] \leq \beta_{q_{n,1}}. \tag{7.35}$$
- *For each $\delta \in \{1, 2\}$, the random vectors $\vec{U}^*_{1,\delta}, ..., \vec{U}^*_{\ell_n,\delta}$ are independent.*

Proof of Claim 2. The claim is a corollary of Berbee's coupling lemma (1979) (see Doukhan *et al.* (1995)) together with $(\mathbf{H}_{X,Y})$. For further details about the construction of the u_i^*'s we refer to Viennet (1997), see Proposition 5.1 and its proof p. 484. □

We set
$$A_0 = h_0^2(1 - 1/\rho)^2/(80\Phi_1^4 h_1), \tag{7.36}$$
and we choose $q_n = \text{int}[A_0\Psi(n)/4] + 1 \geq 1$ ($\text{int}[u]$ denotes the integer part of u) and $q_{n,1} = q_{n,1}(x)$ to satisfy $\sqrt{q_{n,1}/q_n} + \sqrt{1 - q_{n,1}/q_n} \leq \sqrt{x}$, namely $q_{n,1}$ of order $((x-1)^2 \wedge 1)q_n/2$ works. For the sake of simplicity, we assume q_n to divide n i.e. $n = \ell_n q_n$ and we introduce the sets Ω^* and Ω_ρ defined as follows:
$$\Omega^* = \left\{(\varepsilon_i, \vec{X}_i) = (\varepsilon_i^*, \vec{X}_i^*)/\ i = 1, ..., n\right\}$$
and for $\rho \geq 1$,
$$\Omega_\rho = \left\{ \|t\|_\mu^2 \leq \rho\|t\|_n^2, \quad \forall t \in \bigcup_{m,m' \in \mathcal{M}_n} S_m + S_{m'} \right\}.$$
We denote by Ω_ρ^* the set $\Omega^* \cap \Omega_\rho$. From now on, the index m denotes a minimizer of the quantity $\|f\mathbf{I}_A - f_{m'}\|_\mu^2 + \text{pen}(m')$ for $m' \in \mathcal{M}_n$. Therefore, m is fixed and, for the sake of simplicity, the index m is omitted in the three following notations. Let $B(m', \mu)$ be the unit ball in $S(m') = S_{m'} + S_m$ with respect to $\| \ \|_\mu$, i.e.
$$B(m', \mu) = \left\{ t \in S_{m'} + S_m/\ \|t\|_\mu^2 = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n t^2(\vec{X}_i)\right] \leq 1 \right\}.$$
For each $m' \in \mathcal{M}_n$, we set $D(m') = \dim(S(m'))$.

CLAIM 3. *Let x, ρ be numbers satisfying $x > \rho > 1$. If* pen *is chosen to satisfy*

$$\text{pen}(m') \geq x^3 \frac{D_{m'}}{n} \sigma_2^2, \tag{7.37}$$

then

$$\begin{aligned} &\|f\mathbf{I}_A - \tilde{f}\|_n^2 \mathbf{I}_{\Omega_\rho^*} \\ &\quad \leq C_1(x,\rho)\left[\|f\mathbf{I}_A - f_m\|_n^2 + 2\text{pen}(m)\right] + \frac{x(x+\rho)}{x-\rho} n^{-2} W_n(\hat{m}), \end{aligned} \tag{7.38}$$

where $W_n(m')$ is defined by

$$W_n(m') = \left(\left(\sup_{t \in B(m',\mu)} \sum_{i=1}^n \varepsilon_i^* t(\vec{X}_i^*)\right)^2 - x^2 n D(m') \sigma_2^2\right)_+,$$

for $m' \in \mathcal{M}_n$ and where $C_1(x,\rho) = (x+\rho)^2/(x-\rho)^2 > 1$.

PROOF OF CLAIM 3. The following inequalities hold on Ω_ρ^*. Starting from (7.33) we get

$$\begin{aligned} \|f\mathbf{I}_A - \tilde{f}\|_n^2 &\leq \|f\mathbf{I}_A - f_m\|_n^2 + \frac{2}{n}\|\tilde{f} - f_m\|_\mu \sum_{i=1}^n \varepsilon_i^* \frac{(\tilde{f} - f_m)(\vec{X}_i^*)}{\|\tilde{f} - f_m\|_\mu} + \text{pen}(m) - \text{pen}(\hat{m}) \\ &\leq \|f\mathbf{I}_A - f_m\|_n^2 + \frac{2}{n}\|\tilde{f} - f_m\|_\mu \sup_{t \in B(\hat{m},\mu)} \sum_{i=1}^n \varepsilon_i^* t(\vec{X}_i^*) + \text{pen}(m) - \text{pen}(\hat{m}). \end{aligned}$$

Using the elementary inequality $2ab \leq xa^2 + x^{-1}b^2$, which holds for any positive numbers a, b, we have

$$\begin{aligned} \|f\mathbf{I}_A - \tilde{f}\|_n^2 \leq \|f\mathbf{I}_A - f_m\|_n^2 + x^{-1}\|\tilde{f} - f_m\|_\mu^2 + n^{-2}x\left(\sup_{t \in B(\hat{m},\mu)} \sum_{i=1}^n \varepsilon_i^* t(\vec{X}_i^*)\right)^2 \\ +\text{pen}(m) - \text{pen}(\hat{m}). \end{aligned}$$

On $\Omega_\rho^* \subset \Omega_\rho$, we know that for all $t \in \bigcup_{m' \in \mathcal{M}_n} S_m + S_{m'}$, $\|t\|_\mu^2 \leq \rho\|t\|_n^2$, hence

$$\begin{aligned} \|f\mathbf{I}_A - \tilde{f}\|_n^2 &\leq \|f\mathbf{I}_A - f_m\|_n^2 + x^{-1}\rho\|\tilde{f} - f_m\|_n^2 + n^{-2}x\left(\sup_{t \in B(\hat{m},\mu)} \sum_{i=1}^n \varepsilon_i^* t(\vec{X}_i^*)\right)^2 \\ &\quad +\text{pen}(m) - \text{pen}(\hat{m}) \\ &\leq \|f\mathbf{I}_A - f_m\|_n^2 + x^{-1}\rho\left(\|\tilde{f} - f\mathbf{I}_A\|_n + \|f\mathbf{I}_A - f_m\|_n\right)^2 \\ &\quad +n^{-2}x\left(\sup_{t \in B(\hat{m},\mu)} \sum_{i=1}^n \varepsilon_i^* t(\vec{X}_i^*)\right)^2 + \text{pen}(m) - \text{pen}(\hat{m}), \end{aligned}$$

by the triangular inequality. Since for all $y > 0$ (y is chosen at the end of the proof)

$$\left(\|\tilde{f} - f\mathbf{I}_A\|_n + \|f\mathbf{I}_A - f_m\|_n\right)^2 \leq (1+y)\|\tilde{f} - f\mathbf{I}_A\|_n^2 + (1+y^{-1})\|f\mathbf{I}_A - f_m\|_n^2,$$

we obtain

$$
\begin{aligned}
&\|f\mathbf{1}_A-\tilde f\|_n^2(1-\rho\frac{1+y}{x})\\
&\leq \|f\mathbf{1}_A-f_m\|_n^2(1+\rho\frac{1+y^{-1}}{x})+n^{-2}x\left(\sup_{t\in B(\hat m,\mu)}\sum_{i=1}^n\varepsilon_i^*t(\vec X_i^*)\right)^2+\mathrm{pen}(m)-\mathrm{pen}(\hat m)\\
&\leq \|f\mathbf{1}_A-f_m\|_n^2(1+\rho\frac{1+y^{-1}}{x})+\mathrm{pen}(m)+x^3\frac{D_m+D_{\hat m}}{n}\sigma_2^2\\
&\qquad -\mathrm{pen}(\hat m)+\frac{x}{n^2}\left(\left(\sup_{t\in B(\hat m,\mu)}\sum_{i=1}^n\varepsilon_i^*t(\vec X_i^*)\right)^2-x^2nD(\hat m)\sigma_2^2\right)_+,
\end{aligned}
$$

using that $D(\hat m)\leq D_{\hat m}+D_m$. Since the penalty function pen satisfies (7.37) for all $m'\in\mathcal{M}_n$, we obtain that on Ω_ρ^*

$$
\|f\mathbf{1}_A-\tilde f\|_n^2(1-\rho\frac{1+y}{x})\leq\|f\mathbf{1}_A-f_m\|_n^2(1+\rho\frac{1+y^{-1}}{x})+2\mathrm{pen}(m)+xn^{-2}W_n(\hat m),
$$

which gives the claim by choosing $y=(x-\rho)/(x+\rho)$. □

CLAIM 4. *For $p\geq 2(1+2\bar p)/(1-4b)$ we have,*

$$
\begin{aligned}
&\mathbb{E}\left[\|f\mathbf{1}_A-\tilde f\|_n^{2\bar p}\mathbf{1}_{\Omega_\rho^*}\right]\\
&\leq C_1^{\bar p}(x,\rho)\left[\|f\mathbf{1}_A-f_m\|_\mu^2+2\mathrm{pen}(m)\right]^{\bar p}\\
&\quad+\frac{C}{n^{\bar p}}\left(\Phi_0h_0^{-1/2}\right)^p\sigma_p^{2\bar p}\left[\sum_{m'\in\mathcal{M}_n}D_{m'}^{-p/2+\bar p}+(2K)^p\frac{|\mathcal{M}_n|}{n^{(1-4b)(p-2(1+2\bar p)/(1-4b))}}\right],
\end{aligned}
$$

where C is a constant that depends on $x,\rho,p,\bar p$.

PROOF OF CLAIM 4. By taking the power $\bar p\leq 1$ of the right- and left-hand side of (7.38) we obtain

$$
\begin{aligned}
&\|f\mathbf{1}_A-\tilde f\|_n^{2\bar p}\mathbf{1}_{\Omega_\rho^*}\\
&\leq C_1^{\bar p}(x,\rho)\left[\|f\mathbf{1}_A-f_m\|_n^2+2\mathrm{pen}(m)\right]^{\bar p}+\left(\frac{x(x+\rho)}{n^2(x-\rho)}\right)^{\bar p}W_n^{\bar p}(\hat m)\\
&\leq C_1^{\bar p}(x,\rho)\left[\|f\mathbf{1}_A-f_m\|_n^2+2\mathrm{pen}(m)\right]^{\bar p}+\left(\frac{x(x+\rho)}{n^2(x-\rho)}\right)^{\bar p}\sum_{m'\in\mathcal{M}_n}W_n^{\bar p}(m').
\end{aligned}
$$

By taking the expectation on both sides of the inequality and using Jensen's inequality we obtain that

$$
\begin{aligned}
&\mathbb{E}\left[\|f\mathbf{1}_A-\tilde f\|_n^{2\bar p}\mathbf{1}_{\Omega_\rho^*}\right]\\
&\leq C_1^{\bar p}(x,\rho)\left[\|f\mathbf{1}_A-f_m\|_\mu^2+2\mathrm{pen}(m)\right]^{\bar p}+\left(\frac{x(x+\rho)}{n^2(x-\rho)}\right)^{\bar p}\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[W_n^{\bar p}(m')\right].
\end{aligned}\tag{7.39}
$$

We now use the following result,

Proposition 6. *Under the assumptions of Theorem 1,*

$$
\begin{aligned}
&C(p,\bar{p})^{-1}\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[W_n^{\bar{p}}(m')\right]\\
&\le C(p,\bar{p})^{-1}\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\left(\sup_{t\in B(m',\mu)}\sum_{i=1}^{n}\varepsilon_i^* t(\vec{X}_i^*)\right)^2 - x\left(\sqrt{\frac{q_{n,1}}{q_n}}+\sqrt{1-\frac{q_{n,1}}{q_n}}\right)^2 nD(m')\sigma_2^2\right)_+^{\bar{p}}\right]\\
&\le x^{\bar{p}/3}\left(x^{1/3}-1\right)^{\bar{p}-p} n^{\bar{p}}\left(\Phi_0 h_0^{-1/2}\right)^p \sigma_p^{2\bar{p}}\left[\sum_{m'\in\mathcal{M}_n} D_{m'}^{-p/2+\bar{p}}+\frac{q_n^p|\mathcal{M}_n|}{n^{p(p-2)/[4(p-1)]-\bar{p}}}\right].
\end{aligned}
$$

The proof of the second inequality is delayed to Section 8, the first one is a straightforward consequence of our choice of $q_{n,1}$.
Using Proposition 6 we derive from (7.39) that

$$
\begin{aligned}
\mathbb{E}\left[\|f\mathbf{I}_A-\tilde{f}\|_n^{2\bar{p}}\mathbf{I}_{\Omega_\rho^*}\right] &\le C_1^{\bar{p}}(x,\rho)\left[\|f\mathbf{I}_A-f_m\|_\mu^2+2\mathrm{pen}(m)\right]^{\bar{p}}\\
&\quad+\frac{C(x,p,\bar{p})}{n^{\bar{p}}}\left(\Phi_0 h_0^{-1/2}\right)^p\sigma_p^{2\bar{p}}\left[\sum_{m'\in\mathcal{M}_n} D_{m'}^{-p/2+\bar{p}}+\frac{q_n^p|\mathcal{M}_n|}{n^{p(p-2)/[4(p-1)]-\bar{p}}}\right]. \qquad (7.40)
\end{aligned}
$$

Since $A_0\le 1$ and $1\le\Psi(n)\le Kn^b$ we have

$$q_n^p\le 2^p\Psi(n)^p\le(2K)^p n^{bp}$$

hence by using that $p(p-2)/[4(p-1)]\ge(p-2)/4$ we get

$$\frac{q_n^p|\mathcal{M}_n|}{n^{p(p-2)/[4(p-1)]-\bar{p}}}\le(2K)^p\frac{|\mathcal{M}_n|}{n^{(1/4-b)(p-2(1+2\bar{p})/(1-4b))}}. \qquad (7.41)$$

Note that the power of n, $(1/4-b)(p-2(1+2\bar{p})/(1-4b))$ is positive for $p>2(1+2\bar{p})/(1-4b)$. The result follows by gathering (7.40) and (7.41). □

Claim 5. *Under the assumptions of Theorem 1, we have*

$$\mathbb{P}\left[\Omega_\rho^{*c}\right]\le 2(M+e^{16/A_0})n^{-2} \qquad (7.42)$$

and

$$\mathbb{E}\left[\|f\mathbf{I}_A-\tilde{f}\|_n^{2\bar{p}}\mathbf{I}_{\Omega_\rho^{*c}}\right]\le(2(M+e^{16/A_0}))^{1-2\bar{p}/p}\left(\|f\mathbf{I}_A\|_\infty^{2\bar{p}}+\sigma_p^{2\bar{p}}\right)n^{-\bar{p}}. \qquad (7.43)$$

Proof of Claim 5. For the proof of (7.43) we refer to Baraud (2000) (see proof of Theorem 6.1, (49) with $q=\bar{p}$ and $\beta=2$) noticing that $p\ge 2(1+2\bar{p})/(1-4b)>4\bar{p}/(2-\bar{p})$ ($\bar{p}\le 1$). By examining the proof, it is easy to check that if the constants belong to the S_m's then $\|f\mathbf{I}_A\|_\infty$ can be replaced by $\|f\mathbf{I}_A-\int f\mathbf{I}_A\|_\infty$.
To prove (7.42) we use the following Proposition, which is proved in Section 9:

Proposition 7. *Under the assumptions of Theorem 1, for all $\rho > 1$,*

$$\mathbb{P}\left[\Omega_\rho^{*c}\right] \leq 2n^2 \exp\left[-A_0 \frac{\Psi(n)\ln(n)}{q_n}\right] + 2n\beta_{q_{n,1}}. \tag{7.44}$$

Since $q_n = \text{int}[A_0\Psi(n)/4] + 1 \leq A_0\Psi(n)/4 + 1$ we have

$$\begin{aligned} 2n^2 \exp\left[-A_0 \frac{\Psi(n)\ln(n)}{q_n}\right] &\leq 2n^2 \exp\left[4\ln(n)\left(-1 + \frac{4}{A_0\Psi(n)+4}\right)\right] \\ &\leq \frac{2}{n^2} e^{16/A_0}, \end{aligned} \tag{7.45}$$

$\Psi(n)$ being larger than $\ln(n)$. Now, set

$$B = [A_0((x-1)^2 \wedge 1)/8]^{-1} = [h_0^2((x-1)^2 \wedge 1)(1-1/\rho)^2/(640\Phi_0^3 h_1)]^{-1}. \tag{7.46}$$

Since $q_n \geq A_0\Psi(n)/4$, under Condition (4.24) we have

$$\begin{aligned} 2n\beta_{q_{n,1}} &\leq 2nM\left[\Psi^{-1}\left(((x-1)^2 \wedge 1)\frac{Bq_n}{2}\right)\right]^{-3} \\ &\leq 2nM\left[\Psi^{-1}\left(\Psi(n)\right)\right]^{-3} = \frac{2M}{n^2}. \end{aligned} \tag{7.47}$$

Claim 5 is proved by gathering (7.45) and (7.47). □

The proof of Theorem 1 is completed by combining Claim 4 and Claim 5.

8. Proof of Proposition 6

We decompose the proof into two steps:

Step 1. For all $m' \in \mathcal{M}_n$,

$$\begin{aligned} &\mathbb{E}\left[\left(\sup_{t\in B(m',\mu)} \sum_{i=1}^n \varepsilon_i^* t(\vec{X}_i^*) - \left(\sqrt{\frac{q_{n,1}}{q_n}} + \sqrt{1-\frac{q_{n,1}}{q_n}}\right)\sqrt{nD(m')}\sigma_2\right)_+^p\right] \\ &\leq C(p)\sigma_p^p\left[n^{p/2} + (\Phi_0 h_0^{-1/2})^p q_n^p D(m')^{p/2} n^{p^2/[4(p-1)]}\right]. \end{aligned} \tag{8.48}$$

Proof. Using the result of Claim 2, we have the following decomposition:

$$\sum_{i=1}^n \varepsilon_i^* t(\vec{X}_i^*) = \sum_{\ell=1}^{\ell_n}\left(\sum_{i\in I_\ell^{(1)}} \varepsilon_i^* t(\vec{X}_i^*) + \sum_{i\in I_\ell^{(2)}} \varepsilon_i^* t(\vec{X}_i^*)\right)$$

where for $\ell = 1, ..., \ell_n$, $I_\ell^{(1)} = \{(\ell-1)q_n + 1, ..., (\ell-1)q_n + q_{n,1}\}$ and $I_\ell^{(2)} = \{(\ell-1)q_n + q_{n,1} + 1, ..., \ell q_n = (\ell-1)q_n + q_{n,1} + q_n - q_{n,1}\}$. Denoting $\mathbb{E}_1^* = \sqrt{\ell_n q_{n,1} D(m')}\sigma_2$ and $\mathbb{E}_2^* = \sqrt{\ell_n(q_n - q_{n,1})D(m')}\sigma_2$ we have

$$\mathbb{E}\left[\left(\sup_{t\in B(m',\mu)} \sum_{i=1}^n \varepsilon_i^* t(\vec{X}_i^*) - \mathbb{E}_1^* - \mathbb{E}_2^*\right)_+^p\right] \leq 2^{p-1}\mathbb{E}\left[\left(\sup_{t\in B(m',\mu)} \sum_{\ell=1}^{\ell_n}\sum_{i\in I_\ell^{(1)}} \varepsilon_i^* t(\vec{X}_i^*) - \mathbb{E}_1^*\right)_+^p\right]$$

$$+\ 2^{p-1}\mathbb{E}\left[\left(\sup_{t\in B(m',\mu)}\sum_{\ell=1}^{\ell_n}\sum_{i\in I_\ell^{(2)}}\varepsilon_i^* t(\vec{X}_i^*)-\mathbb{E}_2^*\right)_+^p\right].$$

Since the two terms can be bounded in the same way, we only show how to bound the first one. To do so, we use a moment inequality proved in Baraud (2000, Theorem 5.2 p. 478): consider the sequence of independent random vectors of $\left(\mathbb{R}\times\mathbb{R}^k\right)^{q_{n,1}}$, $\vec{U}_1^*, ..., \vec{U}_{\ell_n}^*$ defined by $\vec{U}_\ell^* = \left(\varepsilon_i^*, \vec{X}_i^*\right)'_{i\in I_\ell^{(1)}}$ for $\ell = 1, ..., \ell_n$, and consider $\mathcal{G}_{m'} = \{g_t/\ t\in B(m',\mu)\}$ the set of functions g_t mapping $\left(\mathbb{R}\times\mathbb{R}^k\right)^{q_{n,1}}$ into $\mathbb{R}$ defined by

$$g_t\left((e_1,\vec{x}_1), ..., (e_{q_{n,1}}, \vec{x}_{q_{n,1}})\right) = \sum_{i=1}^{q_{n,1}} e_i t(\vec{x}_i).$$

By applying the moment inequality with the $\vec{U}_\ell^*$'s and the class of functions $\mathcal{G}_{m'}$ we find for all $p\geq 2$

$$C(p)^{-1}\mathbb{E}\left[\left(\sup_{t\in B(m',\mu)}\sum_{\ell=1}^{\ell_n}\sum_{i\in I_\ell^{(1)}}\varepsilon_i^* t(\vec{X}_i^*)-\mathbb{E}_1^*\right)_+^p\right]$$

$$\leq \mathbb{E}\left[\sup_{t\in B(m',\mu)}\sum_{\ell=1}^{\ell_n}\left|\sum_{i\in I_\ell^{(1)}}\varepsilon_i^* t(\vec{X}_i^*)\right|^p\right] + \mathbb{E}^{p/2}\left[\sup_{t\in B(m',\mu)}\sum_{\ell=1}^{\ell_n}\left(\sum_{i\in I_\ell^{(1)}}\varepsilon_i^* t(\vec{X}_i^*)\right)^2\right]$$

$$(8.49)\quad = V_p + V_2^{p/2}$$

provided that

$$\mathbb{E}\left[\sup_{t\in B(m',\mu)}\sum_{\ell=1}^{\ell_n}\sum_{i\in I_\ell^{(1)}}\varepsilon_i^* t(\vec{X}_i^*)\right] \leq \mathbb{E}_1^* = \sqrt{\ell_n q_{n,1} D(m')}\sigma_2. \tag{8.50}$$

Throughout this section, we denote by $G_\ell(t)$ the random process

$$G_\ell(t) = \sum_{i\in I_\ell^{(1)}}\varepsilon_i^* t(\vec{X}_i^*)$$

which is repeatedly involved in our computations. It is worth noticing that it is linear with respect to the argument t.

We first show that (8.50) is true. Let φ_j, $j = 1, ..., D(m')$ be an orthonormal basis of $S_m + S_{m'} \subset \mathbb{L}^2(A,\mu)$. For each $t\in B(m',\mu)$ we have the following decomposition

$$t = \sum_{j=1}^{D(m')} a_j\varphi_j,\quad \sum_{j=1}^{D(m')} a_j^2 \leq 1. \tag{8.51}$$

By Cauchy-Schwarz's inequality we know that

$$\sum_{\ell=1}^{\ell_n} G_\ell(t) = \sum_{j=1}^{D(m')} a_j \left(\sum_{\ell=1}^{\ell_n} G_\ell(\varphi_j) \right) \leq \left[\sum_{j=1}^{D(m')} \left(\sum_{\ell=1}^{\ell_n} G_\ell(\varphi_j) \right)^2 \right]^{1/2}.$$

Thus, by using Jensen's inequality we obtain

$$\mathbb{E}\left[\sup_{t \in B(m',\mu)} \sum_{\ell=1}^{\ell_n} G_\ell(t) \right] \leq \left[\sum_{j=1}^{D(m')} \mathbb{E}\left(\sum_{\ell=1}^{\ell_n} G_\ell(\varphi_j) \right)^2 \right]^{1/2}$$

$$= \left[\sum_{j=1}^{D(m')} \sum_{\ell=1}^{\ell_n} \mathbb{E}[G_\ell^2(\varphi_j)] \right]^{1/2} \tag{8.52}$$

the random variables $(G_\ell(\varphi_j))_{\ell=1,\dots,\ell_n}$ being independent and centered for each $j = 1, ..., D(m')$. Now, for each $\ell = 1, ..., \ell_n$, we know that the laws of the vectors $(\varepsilon_i^*, \vec{X}_i^*)_{i \in I_\ell^{(1)}}$ and $(\varepsilon_i, \vec{X}_i)_{i \in I_\ell^{(1)}}$ are the same, therefore under Condition $\mathbf{H}_{(X,\varepsilon)}$

$$\mathbb{E}\left[G_\ell^2(\varphi_j)\right] = \mathbb{E}\left[\left(\sum_{i \in I_{\ell^{(1)}}} \varepsilon_i \varphi_j(\vec{X}_i) \right)^2 \right] \leq \sum_{i \in I_\ell^{(1)}} \mathbb{E}[\varepsilon_i^2] \mathbb{E}[\varphi_j^2(\vec{X}_i)] = q_{n,1}\sigma_2^2, \tag{8.53}$$

which together with (8.52) proves (8.50).

Let us now bound V_p and V_2 respectively.

The connection between $\|.\|_\infty$ and $\|.\|_\mu$ over $S_m + S_{m'}$ allows to write that for all $t \in B(m', \mu)$,

$$\|t\|_\infty \leq \Phi_0 h_0^{-1/2} \sqrt{D(m')} \times 1. \tag{8.54}$$

Thus,

$$\begin{aligned} V_p &= \mathbb{E}\left[\sup_{t \in B(m',\mu)} \sum_{\ell=1}^{\ell_n} \left| \sum_{i \in I_\ell^{(1)}} \varepsilon_i^* t(\vec{X}_i^*) \right|^p \right] \\ &\leq |I_\ell^{(1)}|^{p-1} \mathbb{E}\left[\sup_{t \in B(m',\mu)} \sum_{\ell=1}^{\ell_n} \sum_{i \in I_\ell^{(1)}} |\varepsilon_i^*|^p |t(\vec{X}_i^*)|^p \right] \\ &\leq q_{n,1}^{p-1} \left(\Phi_0 h_0^{-1/2} \sqrt{D(m')} \right)^{p-2} \mathbb{E}\left[\sup_{t \in B(m',\mu)} \sum_{\ell=1}^{\ell_n} \sum_{i \in I_\ell^{(1)}} |\varepsilon_i^*|^p t^2(\vec{X}_i^*) \right]. \end{aligned}$$

Using (8.51) and Cauchy-Schwarz's inequality we get

$$V_p \leq q_{n,1}^{p-1} \left(\Phi_0 h_0^{-1/2} \sqrt{D(m')} \right)^{p-2} \mathbb{E}\left[\sum_{\ell=1}^{\ell_n} \sum_{j=1}^{D(m')} \sum_{i \in I_\ell^{(1)}} |\varepsilon_i^*|^p \varphi_j^2(\vec{X}_i^*) \right]$$

$$\leq q_n^{p-1}(\Phi_0 h_0^{-1/2})^{p-2} n D(m')^{p/2}\sigma_p^p,$$

recalling that $\ell_n q_{n,1} \leq \ell_n q_n \leq n$. Since for $p \geq 2$, $p^2/[4(p-1)] \geq 1$ one also has

$$V_p \leq q_n^p(\Phi_0 h_0^{-1/2})^p \sigma_p^p D(m')^{p/2} n^{p^2/[4(p-1)]}. \tag{8.55}$$

We now bound V_2. A symmetrization argument (see Giné and Zinn (1984)) gives

$$\begin{aligned} V_2 &= \mathbb{E}\left[\sup_{t\in B(m',\mu)} \sum_{\ell=1}^{\ell_n} G_\ell^2(t)\right] \\ &\leq \sup_{t\in B(m',\mu)} \sum_{\ell=1}^{\ell_n} \mathbb{E}\left[G_\ell^2(t)\right] + 4\mathbb{E}\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n} \xi_\ell G_\ell^2(t)\right|\right] \\ &\leq n\sigma_2^2 + 4\mathbb{E}\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n} \xi_\ell G_\ell^2(t)\right|\right], \end{aligned} \tag{8.56}$$

where the random variables ξ_ℓ's are i.i.d. centered random variables independent of the $\vec{X}_i^*$'s and the ε_i^*'s, satisfying $\mathbb{P}[\xi_1 = \pm 1] = 1/2$. It remains to bound the last term in the right-hand side of (8.56). To do so, we use a truncation argument. We set $M_\ell = \max_{i\in I_\ell^{(1)}} |\varepsilon_i^*|$. For any $c > 0$, we have

$$\begin{aligned} \mathbb{E}\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n} \xi_\ell G_\ell^2(t)\right|\right] &\leq \mathbb{E}\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n} \xi_\ell G_\ell^2(t)\mathbf{1}_{M_\ell\leq c}\right|\right] \\ &\quad + \mathbb{E}\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n} \xi_\ell G_\ell^2(t)\mathbf{1}_{M_\ell> c}\right|\right]. \end{aligned} \tag{8.57}$$

We apply a comparison theorem (Theorem 4.12 p. 112 in Ledoux and Talagrand (1991)) to bound the first term of the right-hand side of (8.57): we know that for each $t \in B(m',\mu)$ the random variables $G_\ell(t)\mathbf{1}_{M_\ell\leq c}$'s are bounded by $B = q_{n,1}\Phi_0 h_0^{-1/2}\sqrt{D(m')}c$ (using (8.54)) and are independent of the ξ_l's. The function $x \mapsto x^2$ defined on the set $[-B, B]$ being Lipschitz with Lipschitz constant smaller than $2B$, we obtain ($\mathbb{E}_\xi$ denotes the conditional expectation with respect to the ε_i^*'s and the X_i^*'s)

$$\begin{aligned} \mathbb{E}_\xi\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n} \xi_\ell G_\ell^2(t)\mathbf{1}_{M_\ell\leq c}\right|\right] &\leq 4B\mathbb{E}_\xi\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n} \xi_\ell G_\ell(t)\mathbf{1}_{M_\ell\leq c}\right|\right] \\ &\leq 4B\mathbb{E}_\xi\left[\sum_{j=1}^{D(m')}\left(\sum_{\ell=1}^{\ell_n} \xi_\ell G_\ell(\varphi_j)\mathbf{1}_{M_\ell\leq c}\right)^2\right]^{1/2} \\ &\leq 4B\left(\sum_{j=1}^{D(m')}\sum_{\ell=1}^{\ell_n} G_\ell^2(\varphi_j)\right)^{1/2}. \end{aligned}$$

We now decondition with respect to the random variables ε_i^*'s and $\vec{X}_i^*$ and using (8.53) we get

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n}\xi_\ell G_\ell^2(t)\mathbf{1}_{M_\ell\le c}\right|\right] &\le 4q_{n,1}\Phi_0 h_0^{-1/2}D(m')\sigma_2\sqrt{n}c\\ &\le 4q_{n,1}^2\Phi_0^2 h_0^{-1}D(m')\sigma_p\sqrt{n}c,\end{aligned}\tag{8.58}$$

noticing that $q_{n,1}$, $\Phi_0 h_0^{-1/2}$ are both greater than 1.

Now, we bound the second term of the right-hand side of (8.57). We have

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n}\xi_\ell G_\ell^2(t)\mathbf{1}_{M_\ell> c}\right|\right] &\le \mathbb{E}\left[\sup_{t\in B(m',\mu)}\sum_{\ell=1}^{\ell_n}G_\ell^2(t)\mathbf{1}_{M_\ell>c}\right]\\ &\le \mathbb{E}\left[\sum_{j=1}^{D(m')}\sum_{\ell=1}^{\ell_n}G_\ell^2(\varphi_j)\mathbf{1}_{M_\ell>c}\right]\\ &\le q_{n,1}\mathbb{E}\left[\sum_{j=1}^{D(m')}\sum_{\ell=1}^{\ell_n}M_\ell^2\mathbf{1}_{M_\ell>c}\sum_{i\in I_\ell^{(1)}}\varphi_j^2(\vec{X}_i^*)\right]\\ &\le q_{n,1}c^{2-p}\sum_{\ell=1}^{\ell_n}\mathbb{E}\left[M_\ell^p\sum_{i\in I_\ell^{(1)}}\left(\sum_{j=1}^{D(m')}\varphi_j^2(\vec{X}_i^*)\right)\right]\\ &\le q_{n,1}^2c^{2-p}\Phi_0^2h_0^{-1}D(m')\sum_{\ell=1}^{\ell_n}\mathbb{E}\left[M_\ell^p\right]\end{aligned}$$

using (2.11). Lastly, since $M_\ell^p\le\sum_{i\in I_\ell^{(1)}}|\varepsilon_i^*|^p$ we get

$$\mathbb{E}\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n}\xi_\ell G_\ell^2(t)\mathbf{1}_{M_\ell> c}\right|\right]\le q_{n,1}^2\Phi_0^2h_0^{-1}nD(m')\sigma_p^pc^{2-p}.\tag{8.59}$$

By gathering (8.58) and (8.59) we obtain that for all $c>0$

$$\mathbb{E}\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n}\xi_\ell G_\ell^2(t)\right|\right]\le 4q_{n,1}^2\Phi_0^2h_0^{-1}\sigma_pD(m')\sqrt{n}\left(c+\sqrt{n}\sigma_p^{p-1}c^{2-p}\right).$$

We choose $c=\sigma_pn^{1/(2p-2)}$, and thus from (8.56) we get

$$V_2=\mathbb{E}\left[\sup_{t\in B(m',\mu)}\left|\sum_{\ell=1}^{\ell_n}\xi_\ell G_\ell^2(t)\right|\right]\le n\sigma_2^2+8q_n^2\Phi_0^2h_0^{-1}\sigma_p^2D(m')n^{p/[2(p-1)]},\tag{8.60}$$

which straightforwardly proves Step 1 by combining (8.49), (8.55) and (8.60). □

STEP 2. For all $x > 1$, $m' \in \mathcal{M}_n$, $\bar{p} < 2p$

$$n^{-\bar{p}}\mathbb{E}\left[\left(\left(\sup_{t\in B(m',\mu)}\sum_{i=1}^{n}\varepsilon_i^* t(\vec{X}_i^*)\right)^2 - x\left(\sqrt{\frac{q_{n,1}}{q_n}}+\sqrt{1-\frac{q_{n,1}}{q_n}}\right)^2 nD(m')\sigma_2^2\right)_+^{\bar{p}}\right]$$
$$\leq C(p,x)(\Phi_0 h_0^{-1})^p\sigma_p^p\left[D(m')^{-(p/2-\bar{p})}+q_n^p n^{\bar{p}-p(p-2)/(p-1)}\right].$$

PROOF. We set $Z_n(m') = \sup_{t\in B(m',\mu)}\sum_{i=1}^{n}\varepsilon_i^* t(\vec{X}_i^*) \geq 0$ and

$$\mathbb{E}^* = \left(\sqrt{\frac{q_{n,1}}{q_n}}+\sqrt{1-\frac{q_{n,1}}{q_n}}\right)\sqrt{nD(m')}\sigma_2 \geq \sqrt{nD(m')}\sigma_2.$$

Since $x > 1$, there exists $\eta > 0$ such that $x = (1+\eta)^3$ (i.e. $\eta = x^{1/3}-1$). Thus for all $\tau > 0$

$$\begin{aligned}\mathbb{P}\left[Z_n^2(m') \geq (1+\eta)^3\left(\mathbb{E}^*\right)^2+\tau\right] &\leq \mathbb{P}\left[Z_n^2(m') \geq \left((1+\eta)\mathbb{E}^*+\sqrt{\frac{\tau}{(1+\eta^{-1})}}\right)^2\right]\\ &\leq \mathbb{P}\left[Z_n(m')-\mathbb{E}^* \geq \eta\mathbb{E}^*+\sqrt{\frac{\tau}{(1+\eta^{-1})}}\right]\\ &\leq \mathbb{P}\left[Z_n(m')-\mathbb{E}^* \geq \sqrt{\eta^2(\mathbb{E}^*)^2+\frac{\tau}{(1+\eta^{-1})}}\right]\\ &\leq \left(\eta^2(\mathbb{E}^*)^2+\frac{\tau}{(1+\eta^{-1})}\right)^{-p/2}\mathbb{E}\left[(Z_n(m')-\mathbb{E}^*)_+^p\right]\\ &\leq \left(\frac{x^{1/3}}{x^{1/3}-1}\right)^{p/2}\frac{\mathbb{E}\left[(Z_n(m')-\mathbb{E}^*)_+^p\right]}{\left((x^{1/3}-1)x^{1/3}nD(m')\sigma_2^2+\tau\right)^{p/2}},\end{aligned}$$

using Markov's inequality. Now, for each $\bar{p}$ such that $2\bar{p} < p$, the integration with respect to the variable τ leads to

$$\begin{aligned}&\mathbb{E}\left[\left(Z_n^2(m')-x\left(\mathbb{E}^*\right)^2\right)_+^{\bar{p}}\right]\\ &= \int_0^{+\infty}\bar{p}\tau^{\bar{p}-1}\mathbb{P}\left[Z_n^2(m')-x\left(\mathbb{E}^*\right)^2 \geq \tau\right]d\tau\\ &\leq \left(\frac{x^{1/3}}{x^{1/3}-1}\right)^{p/2}\mathbb{E}\left[(Z_n(m')-\mathbb{E}^*)_+^p\right]\int_0^{+\infty}\frac{\bar{p}\tau^{\bar{p}-1}}{\left((x^{1/3}-1)x^{1/3}nD(m')\sigma_2^2+\tau\right)^{p/2}}d\tau\\ &\leq \frac{p}{p-2\bar{p}}\frac{\left(x^{1/3}(x^{1/3}-1)\right)^{\bar{p}}}{\left(x^{1/3}-1\right)^p}\frac{\mathbb{E}\left[(Z_n(m')-\mathbb{E}^*)_+^p\right]}{(nD(m')\sigma_2^2)^{p/2-\bar{p}}},\end{aligned}$$

and using STEP 1, we get

$$\mathbb{E}\left[\left(Z_n^2(m')-x\left(\mathbb{E}^*\right)^2\right)_+^{\bar{p}}\right]$$

$$\leq C \frac{\big(x^{1/3}(x^{1/3}-1)\big)^{\bar{p}}}{\big(x^{1/3}-1\big)^{p}} (\Phi_0 h_0^{-1/2})^p \sigma_2^{2\bar{p}-p} \sigma_p^p n^{\bar{p}} \left(D(m')^{-(p/2-\bar{p})} + q_n^p D(m')^{\bar{p}} n^{-p(p-2)/[4(p-1)]} \right)$$

$$\leq C \frac{\big(x^{1/3}(x^{1/3}-1)\big)^{\bar{p}}}{\big(x^{1/3}-1\big)^{p}} (\Phi_0 h_0^{-1/2})^p \sigma_p^{2\bar{p}} n^{\bar{p}} \left(D(m')^{-(p/2-\bar{p})} + q_n^p n^{\bar{p}-p(p-2)/[4(p-1)]} \right),$$

since $D(m') = \dim(S_m + S_{m'}) \leq n$. The constant C depends on p and $\bar{p}$. □

It is now easy to prove Proposition 6 by summing up over m' in $\mathcal{M}_n$.

9. Proof of Proposition 7 Since $\mathbb{P}(\Omega_\rho^{*c}) = \mathbb{P}(\Omega_\rho^c \cap \Omega^*) + \mathbb{P}(\Omega^{*c})$ and since it is clear from Claim 2 that

$$\mathbb{P}(\Omega^{*c}) \leq \ell_n \left(\beta_{(q_n - q_{n,1})} + \beta_{q_{n,1}}\right) \leq 2n\beta_{q_{n,1}}, \tag{9.61}$$

the result holds if we prove

$$\mathbb{P}(\Omega_\rho^c \cap \Omega^*) \leq 2n^2 \exp\left(-A_0 \frac{\Psi(n)\ln(n)}{q_n}\right). \tag{9.62}$$

In fact, we prove a more general result, namely:

$$\mathbb{P}(\Omega_\rho^c \cap \Omega^*) \leq 2D_n^2 \exp\left(-\frac{h_0^2(1-1/\rho)^2}{16h_1} \frac{n}{q_n L(\phi)}\right) \tag{9.63}$$

where $L(\phi)$ is a quantity specific to the orthonormal basis $(\phi_\lambda)_{\lambda\in\Lambda_n}$, defined as follows.

Let $(\phi_\lambda)_{\lambda\in\Lambda_n}$ be a $\mathbb{L}^2(dx)$-orthonormal basis of $\mathcal{S}_n$ and as in Baraud (2001) define the quantities:

$$V = \left(\sqrt{\int_A \phi_\lambda^2(x)\phi_{\lambda'}^2(x)dx}\right)_{\lambda,\lambda'\in\Lambda_n\times\Lambda_n}, \quad B = \left(\|\phi_\lambda\phi_{\lambda'}\|_\infty\right)_{\lambda,\lambda'\in\Lambda_n\times\Lambda_n},$$

and for any symmetric matrix $A = (A_{\lambda,\lambda'})$,

$$\bar{\rho}(A) = \sup_{\{a_\lambda\},\sum_\lambda a_\lambda^2=1} \sum_{\lambda,\lambda'} |a_\lambda||a_{\lambda'}||A_{\lambda,\lambda'}|.$$

We set

$$L(\phi) = \max\{\bar{\rho}^2(V), \bar{\rho}(B)\}. \tag{9.64}$$

Then, to finish the proof of Proposition 7, it remains to check that

$$L(\phi) \leq K' \frac{n}{\Psi(n)\ln(n)}, \tag{9.65}$$

for some constant K' independent of n (we shall show the result for $K' = \Phi_1^4$). Under $(\mathbf{H}_{\mathcal{S}_n})$, Lemma 2 in Section 10 ensures that

$$L(\phi) \leq \Phi_1^4 D_n,$$

which together with (4.23) leads to (9.65).

Now we prove Inequality (9.63). First note that if $\rho > 1$,

$$\sup_{t\in\mathcal{S}_n/\{0\}} \frac{\|t\|_\mu^2}{\|t\|_n^2} \geq \rho \Leftrightarrow \sup_{t\in\mathcal{S}_n/\{0\}} \left(\frac{-\nu_n(t^2)}{\|t\|_\mu^2}\right) \geq 1-\frac{1}{\rho},$$

where $\nu_n(u) = (1/n)\sum_{i=1}^n (u(\vec{X}_i) - \mathbb{E}_\mu(u))$ denotes the centered empirical process. Then for $\rho > 1$,

$$\mathbb{P}^*\left(\sup_{t\in\mathcal{S}_n/\{0\}} \frac{\|t\|_\mu^2}{\|t\|_n^2} \geq \rho\right) \leq \mathbb{P}^*\left(\sup_{t\in B_n^\mu(0,1)} |\nu_n(t^2)| \geq 1-\frac{1}{\rho}\right)$$

where we denote by $\mathbb{P}^*(A)$ the probability $\mathbb{P}(A\cap\Omega^*)$, and by $B_\mu^n(0,1) = \{t \in \mathcal{S}_n, \|t\|_\mu \leq 1\}$.
For $t \in B_n^\mu(0,1)$, $t = \sum_{\lambda\in\Lambda_n} a_\lambda\phi_\lambda$ with $\sum_{\lambda\in\Lambda_n} a_\lambda^2 \leq h_0^{-1}$, and we have

$$\begin{aligned}\sup_{t\in B_n^\mu(0,1)} |\nu_n(t^2)| &\leq \sup_{\sum_\lambda a_\lambda^2\leq 1} h_0^{-1}\left|\sum_{\lambda,\lambda'\in\Lambda_n^2} a_\lambda a_{\lambda'}\nu_n(\phi_\lambda\phi_{\lambda'})\right| \\ &\leq \sup_{\sum_\lambda a_\lambda^2\leq 1} h_0^{-1}\sum_{\lambda,\lambda'\in\Lambda_n^2} |a_\lambda||a_{\lambda'}||\nu_n(\phi_\lambda\phi_{\lambda'})|\end{aligned}$$

Let $x = h_0^2(1-1/\rho)^2/(16h_1L(\phi))$. Then on the set $\{\forall(\lambda,\lambda') \in \Lambda_n^2/\nu_n(\phi_\lambda\phi_{\lambda'}) \leq 2V_{\lambda,\lambda'}\sqrt{2h_1x} + 2B_{\lambda,\lambda'}x\}$, we have

$$\begin{aligned}\sup_{t\in B_n^\mu(0,1)} |\nu_n(t^2)| &\leq 2h_0^{-1}\left(\sqrt{2h_1x}\bar{\rho}(V) + x\bar{\rho}(B)\right) \\ &\leq (1-1/\rho)\left(\frac{1}{\sqrt{2}}\left(\frac{\bar{\rho}^2(V)}{L(\phi)}\right)^{1/2} + \frac{h_0(1-1/\rho)}{8h_1}\frac{\bar{\rho}(B)}{L(\phi)}\right) \\ &\leq (1-1/\rho)\left(\frac{1}{\sqrt{2}}+\frac{1}{8}\right) \leq (1-1/\rho).\end{aligned}$$

The proof of Inequality (9.63) is then achieved by using the following claim.

Claim 6. *Let $(\phi_\lambda)_{\lambda\in\Lambda_n}$ be an $\mathbb{L}^2(A,dx)$ basis of $\mathcal{S}_n$. Then, for all $x \geq 0$ and all integer q, $1 \leq q \leq n$,*

$$\mathbb{P}^*\left(\exists(\lambda,\lambda') \in \Lambda_n^2/|\nu_n(\phi_\lambda\phi_{\lambda'})| > 2V_{\lambda,\lambda'}\sqrt{2h_1x} + 2B_{\lambda,\lambda'}x\right) \leq 2D_n^2\exp\left(-\frac{nx}{q_n}\right).$$

This implies that

$$\mathbb{P}(\Omega_\rho^c\cap\Omega^*) \leq 2D_n^2\exp\left(-\frac{h_0^2(1-1/\rho)^2}{16h_1}\frac{n}{q_nL(\phi)}\right),$$

and thus Inequality (9.63) holds true. □

Proof of Claim 6. Let $\nu_n^*(\phi_\lambda\phi_{\lambda'}) = \nu_{n,1}^*(\phi_\lambda\phi_{\lambda'}) + \nu_{n,2}^*(\phi_\lambda\phi_{\lambda'})$ be defined by

$$\nu_{n,k}^*(\phi_\lambda\phi_{\lambda'}) = \frac{1}{\ell_n}\sum_{l=0}^{\ell_n-1} Z_{l,k}^*(\phi_\lambda\phi_{\lambda'}),\ \ k = 1, 2$$

where for $0 \leq l \leq \ell_n - 1$,

$$Z^*_{l,k}(\phi_\lambda \phi_{\lambda'}) = \frac{1}{q_n} \sum_{i \in \mathcal{I}_l^{(k)}} \left(\phi_\lambda(\vec{X}_i^*)\phi_{\lambda'}(\vec{X}_i^*) - \mathbb{E}_\mu(\phi_\lambda \phi_{\lambda'}) \right), \quad k = 1, 2.$$

We have

$$\begin{aligned}
&\mathbb{P}\left(|\nu_n(\phi_\lambda \phi_{\lambda'})| > 2V_{\lambda,\lambda'}\sqrt{2h_1 x} + 2B_{\lambda,\lambda'} x \right) \\
&\leq \mathbb{P}^* \left(|\nu^*_{n,1}(\phi_\lambda \phi_{\lambda'})| > V_{\lambda,\lambda'}\sqrt{2h_1 x} + B_{\lambda,\lambda'} x \right) + \mathbb{P}\left(|\nu^*_{n,2}(\phi_\lambda \phi_{\lambda'})| > V_{\lambda,\lambda'}\sqrt{2h_1 x} + B_{\lambda,\lambda'} x \right) \\
&= \mathbb{P}_1 + \mathbb{P}_2.
\end{aligned}$$

Now, we bound $\mathbb{P}_1$ and $\mathbb{P}_2$ by using Bernstein's inequality (see Lemma 8 p.366 in Birgé and Massart (1998)) applied to the independent variables $Z^*_{l,k}$, which satisfy $\|Z^*_{l,k}\|_\infty \leq B_{\lambda,\lambda'}$ and $\mathbb{E}^{1/2}[(Z^*_{l,k})^2] \leq \sqrt{h_1} V_{\lambda,\lambda'}$. Then we obtain $\mathbb{P}_1 + \mathbb{P}_2 \leq 2\exp(-x\ell_n)$, which proves the claim 6. □

10. Constraints on the dimension of $\mathcal{S}_n$ Most elements of the following proof can be found in Baraud (2001), but we recall them for the paper to be self-contained.

Let $\mathcal{S}_n$ be the linear subspace defined at the beginning of Section 4. We recall that $\mathcal{S}_n$ is generated by an orthonormal basis $(\phi_\lambda)_{\lambda \in \Lambda_n}$ and that $D_n = |\Lambda_n|$. In the previous section the conditions on $\mathcal{S}_n$ (given by $(\mathbf{H}_{\mathcal{S}_n})$) and D_n (given by (4.23)) are used to prove (9.65). To obtain (9.65) we proceed into two steps: first, under some particular characteristics of the basis $(\phi_\lambda)_{\lambda \in \Lambda_n}$ (in the case of Theorem 1 these characteristics are given by $(\mathbf{H}_{\mathcal{S}_n})$), we state an upper bound on $L(\phi)$ depending on Φ_1(or Φ_0) and D_n. Secondly, starting from this bound we specify a constraint on D_n for (9.65) to hold. In the next lemma we consider various cases of linear spaces $\mathcal{S}_n$ (including those considered in Theorem 1) and provide upper bounds on $L(\phi)$ according to the characteristics of one of their orthonormal basis.

Lemma 2. *Let $L(\phi)$ be the quantity defined by (9.64).*

1. If $\mathcal{S}_n$ satisfies (2.9) then $L(\phi) \leq \Phi_0^2 D_n^2$.

2. Under $(\mathbf{H}_{\mathcal{S}_n})$, $L(\phi) \leq \Phi_1^4 D_n$. Moreover, (2.9) holds true with $\Phi_0^2 = \Phi_1^3$.

We obtain from *1.* and *2.* that the constraints on D_n given by (4.28) and (4.23) lead to (9.65).

Proof of *1.* On the one hand, by Cauchy-Schwarz's inequality we have that

$$\begin{aligned}
\bar{\rho}^2(V) &\leq \sum_{\lambda,\lambda' \in \Lambda_n} \int \phi_\lambda^2 \phi_{\lambda'}^2 \leq \sum_{\lambda' \in \Lambda_n} \int \left(\sum_{\lambda \in \Lambda_n} \phi_\lambda^2 \right) \phi_{\lambda'}^2 \\
&\leq \| \sum_{\lambda \in \Lambda_n} \phi_\lambda^2 \|_\infty \sum_{\lambda' \in \Lambda_n} \int \phi_{\lambda'}^2 \leq \Phi_0^2 D_n^2,
\end{aligned}$$

using (2.11). On the other hand, by (2.9) we know that $\|\phi_\lambda\|_\infty \leq \Phi_0\sqrt{D_n} \times 1$. Thus, using similar arguments one gets

$$\bar{\rho}(B) \leq \Phi_0^2 D_n^2,$$

which leads to $L(\phi) \leq \Phi_0^2 D_n^2$. □

Proof of *2.*. We now prove that (2.9) holds true in the case *2*. Note that for all x,

$$\sum_{\lambda \in \Lambda_n} \phi_\lambda^2(x) \leq \Phi_1 \|\phi_\lambda\|_\infty^2 \leq \Phi_1^3 D_n.$$

thus, (2.11) holds true with $\Phi_0^2 = \Phi_1^3$.
Under $(\mathbf{H}_{\mathcal{S}_n})$, $\Delta(\lambda) = \{\lambda' \in \Lambda_n \ / \ \phi_\lambda\phi_{\lambda'} \not\equiv 0\}$ satisfies $|\Delta(\lambda)| \leq \Phi_1$ and

$$\forall \lambda \in \Lambda_n,\ \forall \lambda' \in \Delta(\lambda),\ \int \phi_\lambda^2 \phi_{\lambda'}^2 \leq \Phi_1^2 D_n.$$

Therefore,

$$\begin{aligned}
\bar{\rho}(V) &= \sup_{\{(a_\lambda)_\lambda,\ \sum_\lambda a_\lambda^2 = 1\}} \sum_\lambda \sum_{\lambda' \in \Delta(\lambda)} |a_\lambda||a_{\lambda'}| \left(\int \phi_\lambda^2 \phi_{\lambda'}^2 \right)^{1/2} \\
&\leq \sqrt{\Phi_1^2 D_n} \sup_{\{(a_\lambda)_\lambda,\ \sum_\lambda a_\lambda^2 = 1\}} \sum_\lambda |a_\lambda| \sum_{\lambda' \in \Delta(\lambda)} |a_{\lambda'}| \\
&= \sqrt{\Phi_1^2 D_n} W_n.
\end{aligned}$$

Besides, $\forall \lambda \in \Lambda_n, \forall \lambda' \in \Delta(\lambda)$, $\|\phi_\lambda \phi_{\lambda'}\|_\infty \leq \Phi_1^2 D_n$ and thus

$$\bar{\rho}(B) = \sup_{\sum_\lambda a_\lambda^2 = 1} |a_\lambda||a_{\lambda'}| \|\phi_\lambda\phi_{\lambda'}\|_\infty \leq \Phi_1^2 D_n W_n.$$

Lastly,

$$\begin{aligned}
W_n^2 &\leq \sup_{\sum_\lambda a_\lambda^2 = 1} \sum_{\lambda \in \Lambda_n} \left(\sum_{\lambda' \in \Delta(\lambda)} |a_{\lambda'}| \right)^2 \leq \Phi_1 \sup_{\sum_\lambda a_\lambda^2 = 1} \sum_{\lambda \in \Lambda_n} \sum_{\lambda' \in \Delta(\lambda)} a_{\lambda'}^2 \\
&= \Phi_1 \sup_{\sum_\lambda a_\lambda^2 = 1} \sum_{\lambda' \in \Lambda_n} \sum_{\lambda \in \Delta(\lambda')} a_{\lambda'}^2 = \Phi_1 \sup_{\sum_\lambda a_\lambda^2 = 1} \sum_{\lambda' \in \Lambda_n} |\Delta(\lambda')| a_{\lambda'}^2 \\
&\leq \Phi_1^2.
\end{aligned}$$

In other words, $\bar{\rho}(V) \leq \Phi_1^2 \sqrt{D_n}$ and $\bar{\rho}(B) \leq \Phi_1^3 D_n$, which gives the bound $L(\phi) \leq \Phi_1^4 D_n$ since $\Phi_1 \geq 1$. □

Acknowledgments. The authors are deeply grateful to Lucien Birgé for numbers of constructive suggestions and thank Pascal Massart for helpful comments.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings 2nd International Symposium on Information Theory*, P.N. Petrov and F. Csaki (Eds.), Akademia Kiado, Budapest, 267–281.

AKAIKE, H. (1984). A new look at the statistical model identification. *IEEE Trans. on Automatic Control* **19**, 716–723.

BARAUD, Y. (1998). Sélection de modèles et estimation adaptative dans différents cadres de régression. Ph.D. Thesis. # 5158, Université Paris-Sud.

BARAUD, Y. (2001). Model selection for regression on a random design. Preprint 01-10, DMA, Ecole Normale Supérieure, Paris.

BARAUD, Y. (2000). Model selection for regression on a fixed design, *Probab. Theory Relat. Fields* **117**, 467–493.

BARRON, A.R. (1991). Complexity regularization with application to artificial neural networks, *Proceedings NATO Advanced Study Institute on Nonparametric Functional estimation*, G.Roussas, Ed., Dordrecht, The Netherlands: Kluwer, 561–576.

BARRON, A.R. (1993). Universal approximation bounds for superpositions of a sigmoidal function processes. *IEEE Trans. Inform. Theory* **39**, 930–945.

BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risks bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113**, 301–413.

BARRON, A.R. and COVER, T.M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37**, 1034–1054.

BERBEE, H.C.P. (1979). Random walks with stationary increments and renewal theory. *Math. Tracts. Mathematisch Centrum*, Amsterdam, **112**.

BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. In Festschrift for Lucien Lecam: Research Papers in Probability and Statistics (D. Pollard, E. Torgensen and G. Yangs, eds), 55–87, Springer Verlag, New-York.

BIRGÉ, L. and MASSART, P. (1998). Exponential bounds for minimum contrast estimators on sieves. *Bernoulli* **4**, 329–375.

COHEN, A. DAUBECHIES, I. and VIAL, P. (1993). Wavelet and fast wavelet transform on an interval. *Appl. Comp. Harmon. Anal.* **1**, 54–81.

DAUBECHIES, I. (1992). *Ten lectures on wavelets.* SIAM: Philadelphia.

DEVORE, R.A. and LORENTZ, C.G. (1993). *Constructive Approximation.* Springer-Verlag.

DONOHO, D.L. and JOHNSTONE, I.M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879–921.

DOUKHAN, P. (1994). *Mixing properties and examples.* Springer-Verlag.

DOUKHAN, P., MASSART, P. and RIO, E. (1995). Invariance principle for absolutely regular empirical processes. *Ann. Instit. H. Poincaré Probab. Statist.* **31**, 393–427.

DUFLO, M. (1997). *Random Iterative Models.* Springer, Berlin, New-York.

GINÉ, E. and ZINN, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.* **12**, 929–989.

HOFFMANN, M. (1999). On nonparametric estimation in nonlinear AR(1)-models. *Statist. Probab. Lett.* **44**, 29–45.

KOLMOGOROV, A.R. and ROZANOV, Y.A. (1960). On the strong mixing conditions for stationary gaussian sequences. *Theor. Probab. Appl.* **5**, 204–207.

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in banach spaces.* Springer-Verlag.

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

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

MODHA, D.S. and MASRY, E. (1996) Minimum complexity regression estimation with weakly dependent observations. *IEEE Trans. Inform. Theory* **42**, 2133–2145.

MODHA, D.S. and MASRY, E. (1998). Memory-universal prediction of stationary random processes. *IEEE Trans. Inform. Theory* **44**, 117-133.

NEUMANN, M. and KREISS, J.-P. (1998). Regression-type inference in nonparametric autoregression. *Ann. Statist.* **26**, 1570–1613.

PHAM, D.T. and TRAN, L.T. (1985). Some mixing properties of time series models. *Stoc. Proc. Appl.* **19**, 297–303.

POLYAK, B.T. and TSYBAKOV, A. (1992). A family of asymptotically optimal methods for choosing the order of a projective regression estimate. *Theory Probab. Appl.* **37**, 471–481.

RISSANEN, J. (1984). Universal coding, information, prediction and estimation. *IEEE Trans. Inform. Theory* **30**, 629–636.

SHIBATA, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117-126.

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

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

VIENNET, G. (1997). Inequalities for absolutely regular processes: application to density estimation. *Probab. Theory Relat. Fields* **107**, 467–492.

Yannick Baraud
Ecole Normale Supérieure
DMA
45 rue d'Ulm
75230 Paris Cedex 05
FRANCE

Fabienne Comte
Laboratoire
de Probabilités
et Modèles Aléatoires
Boite 188
Université Paris 6
4, place Jussieu
75252 Paris Cedex 05
FRANCE

Gabrielle Viennet
Laboratoire
de Probabilités
et Modèles Aléatoires
Boite 7012
Université Paris 7
2, place Jussieu
75251 Paris Cedex 05
FRANCE