ESAIM: Probability and Statistics July 2001, Vol. 5, 33–49
URL: http://www.emath.fr/ps/

MODEL SELECTION FOR (AUTO-)REGRESSION WITH DEPENDENT DATA

Yannick Baraud[1], F. Comte[2] and G. Viennet[3]

Abstract. In this paper, we study the problem of non parametric estimation of an unknown regression function from dependent data with sub-Gaussian errors. As a particular case, we handle the autoregressive framework. For this purpose, we consider a collection of finite dimensional linear spaces (*e.g.* linear spaces spanned by wavelets or piecewise polynomials on a possibly irregular grid) and we estimate the regression function by a least-squares estimator built on a data driven selected linear space among the collection. This data driven choice is performed *via* the minimization of a penalized criterion akin to the Mallows' C_p. We state non asymptotic risk bounds for our estimator in some $\mathbb{L}_2$-norm and we show that it is adaptive in the minimax sense over a large class of Besov balls of the form $\mathcal{B}_{\alpha,p,\infty}(R)$ with $p \geq 1$.

Mathematics Subject Classification. 62G08, 62J02.

Received April 15, 1999. Revised July 20, 1999 and May 14, 2001.

1. Introduction

We consider here the problem of estimating the unknown function f from n observations $(Y_i, \vec{X}_i)$, $1 \leq i \leq n$ drawn from the regression model

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

where $(\vec{X}_i)_{1\leq i\leq n}$ is a sequence of possibly dependent random vectors in $\mathbb{R}^k$ and the ε_i's are i.i.d. unobservable real valued centered errors with variance σ^2. In particular, if $Y_i = X_i$ and $\vec{X}_i = (X_{i-1}, \ldots, X_{i-k})'$ we recover the classical autoregressive framework of order k. In this paper, we measure the risk of an estimator *via* the expectation of some random $\mathbb{L}_2$-norm based on the $\vec{X}_i$'s. More precisely, if $\hat{f}$ denotes some estimator of f, we define the risk of $\hat{f}$ by

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

where for any functions s, t, $\mathrm{d}_n^2(s,t)$ denotes the squared random distance $n^{-1}\sum_{i=1}^n (s(\vec{X}_i) - t(\vec{X}_i))^2$. We have in mind to estimate f thanks to some suitable least-squares estimator. For this purpose we introduce some finite

Keywords and phrases: Nonparametric regression, least-squares estimator, adaptive estimation, autoregression, mixing processes.

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

[2] Laboratoire de Probabilités et Modèles Aléatoires, Boîte 188, Université Paris 6, 4 place Jussieu, 75252 Paris Cedex 05, France.

[3] Laboratoire de Probabilités et Modèles Aléatoires, Boîte 7012, Université Paris 7, 2 place Jussieu, 75251 Paris Cedex 05, France.

collection of finite dimensional linear spaces $\{S_m, m \in \mathcal{M}_n\}$ (in the sequel, the S_m's are called models) and we associate to each S_m, the least-squares estimator $\hat{f}_m$ of f on it. Under suitable assumptions (in particular if the $\vec{X}_i$'s and the ε_i's are independent sequences) the risk of $\hat{f}_m$ is equal to

$$\mathbb{E}\left[\mathrm{d}_n^2(f, S_m)\right] + \frac{\dim(S_m)}{n}\sigma^2.$$

The aim of this paper is to propose some suitable data driven selection procedure to select some $\hat{m}$ among $\mathcal{M}_n$ in such a way that the least-squares estimator $\hat{f}_{\hat{m}}$ performs almost as well as the best $\hat{f}_m$ over the collection (*i.e.* the one which has the smallest risk). The selection procedure that is considered is a penalized criterion of the following form:

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

where pen is a penalty function mapping $\mathcal{M}_n$ into $\mathbb{R}_+$. Of course the major problem is to determine such a penalty function in order to obtain a resulting estimator $\tilde{f} = \hat{f}_{\hat{m}}$ that performs almost as well as the best $\hat{f}_m$ *i.e.* such that the risk of $\tilde{f}$ achieves, up to a constant, the minimum of the risks over the collection $\mathcal{M}_n$. More precisely we show that one can find a penalty function such that

$$\mathbb{E}\left[\mathrm{d}_n^2(f, \tilde{f})\right] \leq C \inf_{m \in \mathcal{M}_n} \left[\mathbb{E}\left[\mathrm{d}_n^2(f, S_m)\right] + \frac{\dim(S_m) L_m}{n}\sigma^2 \right] \tag{1.2}$$

where the L_m's are related to the collection of models. If the collection of models is not too "rich" then the L_m's can be chosen to be constants independent of n and the right-hand side of (1.2) turns out to be the minimum of the risks (up to a multiplicative constant) among the collection of least-squares estimators that are considered. In most cases the L_m's are either constants or of order $\ln(n)$.

There have been many studies concerning model selection based on Mallows' [22] C_p or related penalization criteria like Akaike's or the BIC criterion for regressive models (see Akaike [1,2], Shibata [28,29], Li [20], Polyak and Tsybakov [27], among many others ...). A common characteristic of all their results is their asymptotic feature. More recently, a general approach to model selection for various statistical frameworks including density estimation and regression has been developed in Barron *et al.* [7] with many applications to adaptive estimation. An original feature of their viewpoint is its non asymptotic character. Unfortunately, their general approach imposes such restrictions to the regression Model (1.1) that it is hardly usable in practice. Following their ideas, Baraud [4, 5] has extended their results to more attractive situations involving realistic assumptions. Baraud [4] is devoted to the study of fixed design regression while Baraud [5] considers Model (1.1) when all random variables $\vec{X}_i$'s and ε_i's are independent, the ε_i's being i.i.d. with a moment of order $p > 2$. Then Baraud *et al.* [6] relaxed the assumption of independence on the $(\vec{X}_i)$'s and the ε_i's as well. Our approach here as well as in the previous papers remains non asymptotic. Although there have been many results concerning adaptation for the classical regression model with independent variables, to our knowledge, not much is known concerning general adaptation methods for non parametric 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, Modha and Masry [24] deal with the model given by (1.1) when the process $(\vec{X}_i, Y_i)_{i \in \mathbb{Z}}$ is strongly mixing. Their approach leads to sub-optimal rates of convergence. It is worth mentioning, for a one dimensional first order autoregressive model, the works of Neumann and Kreiss [26] and Hoffmann [16] which rely on the approximation of an AR(1) autoregression experiment by a regression experiment with independent variables. They study here various non parametric adaptive estimators such as local polynomials and wavelet thresholding estimators. Modha and Masry [25] consider the problem of one step ahead prediction of real valued stationary exponentially strongly mixing processes. Minimum complexity regression estimators based on Legendre polynomials are used to estimate both the model memory and the predictor function. Again

their approach does not lead to optimal rates of convergence, at least in the particular case of an autoregressive model.

Of course, this paper must be compared with our previous work (Baraud *et al.* [6]), where we had milder moment conditions on the errors (the ε_i's must admit moments of order $p > 2$) but stronger condition on the collection of models. Now we require the ε_i's to be sub-Gaussian (typically, the ε_i's are Gaussian or bounded) but we do not impose any assumption on our family of models (except for finiteness); it can be in particular as large as desired. Moreover, we no longer allow any dependency between the ε_i's, but we can provide results for more general types of dependency for the $\vec{X}_i$'s, typically when some norm connections are fulfilled (*i.e.* on the set Ω_n defined by (3.6)). Any kind of dependency is permitted on the $\vec{X}_i$'s as soon as the $\vec{X}_i$'s and the ε_i's are independent sequences of random variables. In the autoregressive framework, they are possibly arithmetically or geometrically β-mixing (the definitions are recalled below). Note that Baraud [5] gave the same kind of results in the independent framework under even milder conditions but assuming that the errors are Gaussian. The techniques involved are appreciably different. We can also refer to Birgé and Massart [8] for a general study of the fixed design regression with Gaussian errors.

Let us now present our results briefly. One can find collections of models such that the estimator $\hat{f}_{\hat{m}}$ is adaptive in the minimax sense over some Besov balls $\mathcal{B}_{\alpha,p,\infty}(R)$ with $p \geq 1$. Furthermore, in various statistical contexts, we also show that the estimator achieves the minimax rate of convergence although the underlying distribution of the $\vec{X}_i$'s is not assumed to be absolutely continuous with respect to the Lebesgue measure. For other estimators and in the case of independent data, such a result has been established by Kohler [18].

The paper is organized as follows: the general statistical framework is described in Section 2, and the main results are given under an Assumption $\mathbf{(H_\mu)}$ in Section 3. Section 4 gives applications to minimax adaptive estimation in the case of wavelets basis. Section 5 is devoted to the study of condition $\mathbf{(H_\mu)}$ in the case of independent sequences $\vec{X}_i$'s and ε_i's or in the case of dependent sequences and (β-mixing) variables $\vec{X}_i$'s. Most proofs are gathered in Sections 6 to 9.

2. The estimation procedure

Let us recall that we observe pairs $(Y_i, \vec{X}_i), i = 1, \ldots, n$ arising from (1.1)

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

The $\vec{X}_i' = (X_{i,1}, \ldots, X_{i,k})$'s are random variables with law μ_i and we set $\mu = n^{-1}\sum_{i=1}^n \mu_i$. The ε_i's are independent centered random variables. The ε_i's may be independent of the $\vec{X}_i$'s or not. In particular, we have in mind to handle the autoregressive case for which $Y_i = X_i$ and $\vec{X}_i = (X_{i-1}, \ldots, X_{i-k})'$. Then the model can be written:

$$X_i = f(X_{i-1}, \ldots, X_{i-k}) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.1}$$

Since we do not assume the ε_i's to be bounded random variables, the law of the $\vec{X}_i$'s is supported by $\mathbb{R}^k$. Nevertheless we aim at providing a "good" estimator of the unknown function $f : \mathbb{R}^k \to \mathbb{R}$ only on some given compact set $A \subset \mathbb{R}^k$.

Let us now describe our estimation procedure. We consider a finite collection of finite dimensional linear spaces $\{S_m\}_{m\in\mathcal{M}_n}$ consisting of A-supported functions belonging to $\mathbb{L}_2(A, \mu)$. In the sequel the linear spaces S_m's are called models. For each $m \in \mathcal{M}_n$, we associate to each model of the collection the least-squares estimator of f, denoted by $\hat{f}_m$, which minimizes over $t \in S_m$ the least-squares contrast function γ_n defined by

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^n \left[Y_i - t(\vec{X}_i)\right]^2. \tag{2.2}$$

Then, given a suitable penalty function pen(·), that is a nonnegative function on $\mathcal{M}_n$ depending only on the data and known parameters, we define $\hat{m}$ as the minimizer over $\mathcal{M}_n$ of $\gamma_n(\hat{f}_m) + \text{pen}(m)$. This implies that the resulting Penalized Least Square Estimator (PLSE for short) $\tilde{f} = \hat{f}_{\hat{m}}$ satisfies for all $m \in \mathcal{M}_n$ and $t \in S_m$

$$\gamma_n(\tilde{f}) + \text{pen}(\hat{m}) \leq \gamma_n(t) + \text{pen}(m). \tag{2.3}$$

The choice of a proper penalty function is the main concern of this paper since it determines the properties of the PLSE.

Throughout this paper, we denote by $\| \ \|$ the Hilbert norm associated to the Hilbert space $\mathbb{L}_2(A,\mu)$ and for each $t \in \mathbb{L}_2(A,\mu)$, $\|t\|_n^2$ denotes the random variable $n^{-1}\sum_{i=1}^n t^2(\vec{X}_i)$. For each $m \in \mathcal{M}_n$, D_m denotes the dimension of S_m and f_m the $\mathbb{L}_2(A,\mu)$-orthogonal projection of f onto S_m. Moreover, we denote by $\mathbb{R}_+^*$ the set of positive real numbers and by ν the Lebesgue measure.

3. Main theorem

Our main result relies on the following assumption on the joint law of the $\vec{X}_i$'s and the ε_i's:

($\mathbf{H}_{X,\varepsilon}$)

(i) The ε_i's are i.i.d. centered random variables that satisfy for all $u \in \mathbb{R}$

$$\mathbb{E}\left[\exp(u\varepsilon_1)\right] \leq \exp\left(\frac{u^2 s^2}{2}\right), \tag{3.1}$$

for some positive s.

(ii) For each $k \in \{1,\ldots,n\}$, ε_k is independent of the σ-field $\mathcal{F}_k = \sigma(\vec{X}_j, 1 \leq j \leq k)$.

Inequality (3.1) is fulfilled as soon as ε_1 is a centered random variable either Gaussian with variance $s^2 = \sigma^2$ or a.s. bounded by s. In the autoregressive model given by (2.1), Condition (ii) is satisfied.

Theorem 3.1. *Let us consider Model (1.1) where f is an unknown function belonging to $\mathbb{L}_2(A,\mu)$ and the random variables ε_i's and $\vec{X}_i$'s satisfy* **($\mathbf{H}_{X,\varepsilon}$)**. *Set $f_A = f\mathbf{1}_A$, let $(L_m)_{m\in\mathcal{M}_n}$ be nonnegative numbers and set*

$$\Sigma_n = \sum_{m\in\mathcal{M}_n} \exp\left(-L_m D_m\right). \tag{3.2}$$

There exists some universal constant ϑ such that if the penalty function is chosen to satisfy

$$\text{pen}(m) \geq \vartheta s^2 \frac{D_m}{n}(1+L_m) \quad \textit{for all} \quad m \in \mathcal{M}_n,$$

then the PLSE $\tilde{f}$ defined by

$$\tilde{f} = \hat{f}_{\hat{m}} \tag{3.3}$$

with

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n}\left\{\frac{1}{n}\sum_{i=1}^n\left[Y_i - \hat{f}_m(\vec{X}_i)\right]^2 + \text{pen}(m)\right\} \tag{3.4}$$

satisfies

$$\mathbb{E}\left[\|f_A - \tilde{f}\|_n^2 \mathbf{1}_{\Omega_n}\right] \leq C \inf_{m\in\mathcal{M}_n}\left[\|f_A - f_m\|^2 + \text{pen}(m)\right] + C'\frac{s^2\Sigma_n}{n} \tag{3.5}$$

where C and C' are universal constants and

$$\Omega_n = \left\{ \omega / \left| \frac{\|t\|_n^2}{\|t\|^2} - 1 \right| \leq \frac{1}{2},\ \forall t \in \bigcup_{m,m' \in \mathcal{M}_n} (S_m + S_{m'}) \setminus \{0\} \right\}. \tag{3.6}$$

Comments

- For the proof of this result we use an exponential martingale inequality given by Meyer [23] and chaining arguments that can also be found in Barron *et al.* [7] to state exponential bounds on supremum of empirical processes.
- One can also define Ω_n by

$$\left\{ \omega / \left| \frac{\|t\|_n^2}{\|t\|^2} - 1 \right| \leq \rho,\ ,\ \forall t \in \bigcup_{m,m' \in \mathcal{M}_n} (S_m + S_{m'}) \setminus \{0\} \right\}$$

  for some ρ chosen to be less than one, then (3.5) holds for some constant C that now depends on ρ.
- A precise calibration of the penalty term (best choices of ϑ and L_m's) can be determined by carrying out simulation experiments (see the related work for density estimation by Birgé and Rozenholc [9]).
- When the $\vec{X}_i$'s are random variables independent of the ε_i's the indicator set $\mathbf{I}_{\Omega_n}$ can be removed in (3.5) (see Sect. 5). We emphasize that in this case no assumption on the type of dependency between the $\vec{X}_i$'s is required.

Below, we present a useful corollary which makes the performance of $\tilde{f}$ more precise when Ω_n (as defined by (3.6)) is known to occur with high probability. Indeed, assume that:

$$(\mathbf{H}_\mu) \quad \text{There exists} \quad \ell > 1 \quad \text{such that} \quad \mathbb{P}(\Omega_n^c) \leq \frac{C_\ell}{n^\ell},$$

then the following result holds:

Corollary 3.1. *Let us consider Model (1.1) where f is an unknown function belonging to $\mathbb{L}_2(A,\mu) \cap \mathbb{L}_\infty(A,\mu)$. Under the Assumptions of Theorem 3.1 and* $(\mathbf{H}_\mu)$*, the PLSE $\tilde{f}$ defined by (3.3) satisfies*

$$\mathbb{E}\left[\|f_A - \tilde{f}\|_n^2 \right] \leq C \inf_{m \in \mathcal{M}_n} \left[\|f_A - f_m\|^2 + \mathrm{pen}(m) \right] + C' \frac{s^2 \Sigma_n}{n} + C'' \frac{\|f_A\|_\infty^2 + s^2}{n} \tag{3.7}$$

where C and C' are universal constants, and C'' depends on C_ℓ and ℓ only.

The constants C and C' in Corollary 3.1 are the same as those in Theorem 3.1. The proof of Corollary 3.1 is deferred to Section 6. We shall then see that if S_m contains the constant functions then $\|f_A\|_\infty^2$ can be replaced by $\|f_A - \int f_A \mathrm{d}\mu\|_\infty^2$. Comments on Condition $(\mathbf{H}_\mu)$ are to be found in Section 5.

4. Adaptation in the minimax sense

Throughout this section we take $k = 1$ for sake of simplicity and since we aim at estimating f on some compact set, with no loss of generality we can assume that $A = [0,1]$.

4.1. Two examples of collection of models

This section presents two collections of models which are frequently used for estimation: piecewise polynomials and compactly supported wavelets. In the sequel, J_n denotes some positive integer.

(P) Let $\mathcal{M}_n$ be the set of pairs $(d, \{b_0 = 0 < b_1 < \cdots < b_{d-1} < b_d = 1\})$ when d varies among $\{1, \ldots, J_n\}$ and $\{b_0 = 0 < b_1 < \cdots < b_{d-1} < b_d = 1\}$ among the dyadic knots $N_j/2^{J_n}$ with $N_j \in \mathbb{N}$. For each $m = (m_1, m_2) \in \mathcal{M}_n$ we define S_m as the linear span generated by the piecewise polynomials of degree less than r based on the dyadic knots given by m_2. More precisely, if $m_1 = d$ and $m_2 = \{b_0 = 0 < b_1 < \cdots < b_{d-1} < b_d = 1\}$ then S_m consists of all the functions of the form

$$t = \sum_{j=1}^{d} P_j \mathbb{1}_{[b_{j-1}, b_j[},$$

where the P_j's are polynomials of degree less than r. Note that $\dim(S_m) = rm_1$. We denote by $\mathcal{S}_n$ the linear space S_m corresponding to the choice $m_1 = 2^{J_n}$ and $m_2 = \{j/2^{J_n},\ j = 0, \ldots, 2^{J_n}\}$. Since $\dim(\mathcal{S}_n) = r2^{J_n}$, we impose the natural constraint $r2^{J_n} \leq n$.

By choosing for all $m \in \mathcal{M}_n$ $L_m = \ln(n/r)/r$, Σ_n defined by (3.2) remains bounded by a constant that is free from n. Indeed for each $d \in \{1, \ldots, J_n\}$,

$$|\{m \in \mathcal{M}_n /\ m_1 = d\}| = C^{d-1}_{2^{J_n}-1} \leq C^d_{2^{J_n}},$$

where C^d_k denotes the binomial coefficient $\left(\begin{array}{c} k \\ d \end{array} \right)$. Thus,

$$\begin{aligned} \sum_{m \in \mathcal{M}_n} \mathrm{e}^{-L_m D_m} &\leq \sum_{d=1}^{2^{J_n}} C^d_{2^{J_n}} \mathrm{e}^{-\ln(n/r)d} \leq (1 + \exp(-\ln(n/r)))^{2^{J_n}} \\ &\leq \exp(n/r \exp(-\ln(n/r))) = e \end{aligned}$$

using that $2^{J_n} \leq n/r$.

(W) For all integer j let $\Lambda(j)$ be the set $\{(j,k),\ k = 1, \ldots, 2^j\}$. Let us consider the $\mathbb{L}_2$-orthonormal system of compactly supported wavelets of regularity r,

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

built by Cohen *et al.* [10]; for a precise description and use, see Donoho and Johnstone [13]. These new functions derive from Daubechies' [11] wavelets at the interior of $[0,1]$ and are boundary corrected at the "edges". For some positive J_n, let $\mathcal{S}_n$ be the linear span of the $\phi_{J_0,k}$'s for $(J_0,k) \in \Lambda(J_0)$ together with the $\varphi_{j,k}$'s for $(j,k) \in \bar{\Lambda}_n = \cup_{J=J_0}^{J_n-1} \Lambda(J)$. We have that $\dim(\mathcal{S}_n) = 2^{J_0} + \sum_{j=J_0}^{J_n-1} |\Lambda(j)| = 2^{J_n} \leq n$ if $J_n \leq \ln_2(n)$. We take $\mathcal{M}_n = \mathcal{P}(\bar{\Lambda}_n)$, ($\mathcal{P}(A)$ denotes the power of the set A) and for each $m \in \mathcal{M}_n$, define S_m as the linear space generated by the $\phi_{J_0,k}$'s for $(J_0,k) \in \Lambda(J_0)$ and the $\varphi_{j,k}$'s for $(j,k) \in m$.

We choose $L_m = \ln(n)$ in order to bound Σ_n by a constant that does not depend on n:

$$\sum_{m \in \mathcal{M}_n} \mathrm{e}^{-L_m D_m} \leq \sum_{D=1}^{2^{J_n}} C^D_{2^{J_n}} \mathrm{e}^{-\ln(n)D} \leq (1 + \exp(-\ln(n)))^{2^{J_n}} \leq \exp(n \exp(-\ln(n))) = e$$

using that $2^{J_n} \leq n$.

4.2. **Two results about adaptation in the minimax sense**

For $p \geq 1$ and $\alpha > 0$, we set

$$|t|_{\alpha,p} = \sup_{y>0} y^{-\alpha} w_d(t,y)_p, \quad d = [\alpha] + 1$$
$$|t|_\infty = \sup_{x,y\in[0,1]} |t(x) - t(y)|$$

where $w_d(t,.)_p$ denotes the modulus of smoothness of t. For a precise definition of those notions, we refer to DeVore and Lorentz [12], Chapter 2, Section 7. We recall that a function t belongs to the Besov space $\mathcal{B}_{\alpha,p,\infty}([0,1])$ if $|t|_{\alpha,p} < \infty$.

In this section we show how an adequate choice of the collection of models leads to an estimator $\tilde{f}$ that is adaptive in the minimax sense (up to a constant) over Besov bodies of the form

$$\mathbb{B}_{\alpha,p,\infty}(R_1,R_2) = \{t \in \mathcal{B}_{\alpha,p,\infty}(A)/ \quad |t|_{\alpha,p} \leq R_1,\ |t|_\infty \leq R_2\}$$

with $p \geq 1$. In a related regression framework, the case $p \geq 2$ was considered in Baraud *et al.* [6] and it is shown there that weak moment conditions on the ε_i's are sufficient to obtain such estimators. We shall take advantage here of the strong integrability assumption on the ε_i's to extend the result to the case where $p \in [1,2[$. The PLSE defined by (3.3) with the collections **(W)** or **(P)** described in Section 4.1 (and the corresponding L_m's) achieves the minimax rates up to a $\ln(n)$ factor. The extra $\ln(n)$ factor is due to the fact that those collections are "too big" for the problem at hand. In the sequel, we exhibit a subcollection of models **(W')** out of **(W)** which has the property to be both "small" enough to avoid the $\ln(n)$ factor in the convergence rate and "big" enough to allow the PLSE to be rate optimal. The choice of this subcollection comes from the compression algorithm field and we refer to Birgé and Massart [8] for more details. It is also proved there how to obtain a suitable collection from piecewise polynomials instead of wavelets.

For $a > 2$ and $x \in (0,1)$, let us set

$$K_j = [\mathcal{L}(2^{J-j})2^J] \ \text{ and } \ \mathcal{L}(x) = \left(1 - \frac{\ln x}{\ln 2}\right)^{-a}, \tag{4.1}$$

where $[x]$ denotes the integer part of x, and

$$L(a) = 1 + \sum_{j=0}^{+\infty} \frac{1 + (a + \ln(2))j}{(1+j)^a}\,. \tag{4.2}$$

Then we define the new collection of models (we take the notations used in the description of collection **(W)**) by:

(W') For $J \in \{J_0, \ldots, J_n - 1\}$, let

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

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

For each $J \in \{J_0, \ldots, J_n - 1\}$ and $m \in \mathcal{M}_n^J$,

$$2^J \leq D_m = 2^J + \sum_{j=J}^{J_n-1} K_j \leq 2^J \left(1 + \sum_{j=1}^{+\infty} j^{-a}\right). \tag{4.3}$$

Hence, for each J, the linear spaces belonging to the collection $\{S_m, m \in \mathcal{M}_n^J\}$ have their dimension of order 2^J. Besides, it will be shown in Section 8 that the space $\cup_{m\in\mathcal{M}_n^J} S_m$ has good (nonlinear) approximation properties with respect to functions belonging to inhomogeneous Besov spaces.

We give a first result under the assumption that μ is absolutely continuous with respect to the Lebesgue measure on $[0,1]$.

Proposition 4.1. *Assume that* $(\mathbf{H}_\mu)$ *and* $(\mathbf{H}_{X,\varepsilon})$ *hold and that* μ *admits a density with respect to the Lebesgue measure on* $[0,1]$ *that is bounded from above by some constant* h_1*. Consider the collection of models* **(W')** *with* J_n *such that* $2^{J_n} \geq \Gamma n/\ln^b(n)$ *for some* $b > 0$ *and* $\Gamma > 0$*. Let* $p \in [1, +\infty]$ *and set*

$$\left(\frac{1}{p} - \frac{1}{2}\right)_+ \leq \alpha_p = \begin{cases} \frac{1}{2}\left(\frac{1}{p} - \frac{1}{2}\right)\left[1 + \sqrt{\frac{2+3p}{2-p}}\right] & \text{if } \ p < 2 \\ 0 & \text{else.} \end{cases}$$

If $\alpha_p < \alpha \leq r$ *then* $\forall (R_1, R_2) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$*, the PLSE defined by (3.3) with* $L_m = L(a)$ *for all* $m \in \mathcal{M}_n$ *satisfies*

$$\sup_{f\in\mathbb{B}_{\alpha,p,\infty}(R_1,R_2)} \mathbb{E}\left[\|f - \tilde{f}\|_n^2\right] \leq C_1 n^{-\frac{2\alpha}{2\alpha+1}} \tag{4.4}$$

where C_1 *depends on* α*,* a*,* s*,* h_1*,* R_1*,* R_2*,* b *and* Γ*.*

We now relax the assumption that μ is absolutely continuous with respect to the Lebesgue measure.

Proposition 4.2. *Assume that* $(\mathbf{H}_\mu)$ *and* $(\mathbf{H}_{X,\varepsilon})$ *hold. Consider the collection of models* **(W')** *with* J_n *such that* $2^{J_n} \geq \Gamma n/\ln^b(n)$ *for some* $b > 0$ *and* $\Gamma > 0$*. Let* $p \in [1, +\infty]$ *and set*

$$\alpha'_p = \frac{1 + \sqrt{2p+1}}{2p}\,.$$

If $\alpha'_p < \alpha \leq r$ *then* $\forall (R_1, R_2) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$*, the PLSE defined by (3.3) with* $L_m = L(a)$ *for all* $m \in \mathcal{M}_n$ *satisfies*

$$\sup_{f\in\mathbb{B}_{\alpha,p,\infty}(R_1,R_2)} \mathbb{E}\left[\|f - \tilde{f}\|_n^2\right] \leq C_2 n^{-\frac{2\alpha}{2\alpha+1}} \tag{4.5}$$

where C_2 *depends on* α*,* a*,* s*,* R_1*,* R_2*,* b *and* Γ*.*

Equations (4.4) and (4.5) hold for $R_2 = +\infty$ if the left-hand-side term is replaced by

$$\sup_{f\in\mathbb{B}_{\alpha,p,\infty}(R_1,+\infty)} \mathbb{E}\left[\|f - \tilde{f}\|_n^2 \mathbb{1}_{\Omega_n}\right]$$

i.e. no assumption on $\|f\|_\infty$ is required provided that the indicator function $\mathbb{1}_{\Omega_n}$ is added.

We shall see in Section 5 that Condition $(\mathbf{H}_\mu)$ need not be assumed to hold when the sequences $(\vec{X}_i)_{i=1,\ldots,n}$ and $(\varepsilon_i)_{i=1,\ldots,n}$ are independent. Moreover in this case one can assume R_2 to be infinite. The proofs of Propositions 4.1 and 4.2 are deferred to Section 8.

5. Study of Ω_n and condition $(\mathbf{H}_\mu)$

In this section, we study Ω_n and we give sufficient conditions for $(\mathbf{H}_\mu)$ to hold. For this purpose, we examine various dependency structures for the joint law of the $\vec{X}_i$'s and the ε_i's.

5.1. Case of independent sequences $(\vec{X}_i)_{i=1,\dots,n}$ and $(\varepsilon_i)_{i=1,\dots,n}$

We start with the case of deterministic $\vec{X}_i$'s. In this context it is clear from the definition of Ω_n that $\mathbb{P}(\Omega_n) = 1$. Thus the indicator $\mathbb{1}_{\Omega_n}$ can be removed in (3.5). More precisely under the assumptions of Theorem 3.1 we have that for some universal constants C and C'

$$\mathbb{E}\left[\|f_A - \tilde{f}\|_n^2\right] \leq C \inf_{m\in\mathcal{M}_n}\left[\|f_A - f_m\|_n^2 + \mathrm{pen}(m)\right] + C'\frac{s^2\Sigma_n}{n}\cdot \tag{5.1}$$

If the sequences $(\vec{X}_i)_{i=1,\dots,n}$ and $(\varepsilon_i)_{i=1,\dots,n}$ are independent then by conditionning over the $\vec{X}_i$'s (5.1) holds and it is enough to average over the $\vec{X}_i$'s to recover (3.5) where the indicator of Ω_n is removed. In conclusion in this context, Inequality (3.7) holds for any function $f \in \mathbb{L}_2(A,\mu)$ with $C" = 0$. Let us emphasize again that in this case no assumption on the type of dependency of the $\vec{X}_i$'s is required.

5.2. Case of β-mixing $\vec{X}_i$'s

The next proposition presents some dependency situations where Assumption $(\mathbf{H}_\mu)$ is fulfilled: more precisely, we can check this assumption when the variables are geometrically or arithmetically β-mixing. We refer to Kolmogorov and Rozanov [19] for a precise definition of β-mixing and to Ibragimov [17], Volonskii and Rozanov [31] or Doukhan [14] for examples. A sequence of random vectors is said to be geometrically β-mixing if the decay of their β-mixing coefficients, $(\beta_k)_{k\geq 0}$, is exponential, that is if there exists two positive numbers M and θ such that $\beta_k \leq M\mathrm{e}^{-\theta k}$ for all $k \geq 0$. The sequence is said to be arithmetically β-mixing if the decay is hyperbolic, that is if there exists two positive numbers M and θ such that $\beta_k \leq Mk^{-\theta}$ for all $k > 0$.

Since our results are expressed in terms of μ-norm, we introduce a condition ensuring that there exists a connection between this and the ν-norm. We recall that ν denotes the Lebesgue measure.

(C1): The restriction of μ to the set A admits a density h_X w.r.t. the Lebesgue measure such that: $0 < h_0 \leq h_X \leq h_1$ where h_0 and h_1 are some fixed constants chosen such that $h_0 \leq 1 \leq h_1$.

A typical situation where **(C1)** is satisfied is once again the autoregressive model (2.1): in the particular case where $k = 1$ and where the stationary distribution μ_ε of the ε_i's is equivalent to the Lebesgue measure, it follows from Duflo [15] that the variables X_i's admit a density h_X w.r.t. the Lebesgue measure on $\mathbb{R}$ which satisfies: $h_X(y) = \int h_\varepsilon[y - f(x)]h_X(x)\mathrm{d}x$. Then h_X is a continuous function and since A is a compact, there exist two constants $h_0 > 0$ and $h_1 \geq 1$ such that $h_0 \leq h_X(x) \leq h_1$, $\forall x \in A$.

Proposition 5.1. *Assume that* **(C1)** *holds.*

(i) *If the process $(\vec{X}_i)$ is geometrically β-mixing with constants M and θ and if $\dim(\mathcal{S}_n) \leq n/\ln^3(n)$ then $(\mathbf{H}_\mu)$ is satisfied for the collections* **(P)** *and* **(W)** *with $\ell = 2$ and $C_\ell = C(M,\theta,h_0,h_1)$.*

(ii) *If the process $(\vec{X}_i)$ is arithmetically β-mixing with constants M and $\theta > 12$ and if $\dim(\mathcal{S}_n) \leq n^{1-3/\theta}/\ln(n)$ then $(\mathbf{H}_\mu)$ is satisfied for the collections* **(P)** *and* **(W)** *with $\ell = 2$ and $C_\ell = C(M,\theta,h_0,h_1)$.*

Proof. The result derives from Claim 5 in Baraud *et al.* [6] with $\rho = 1/2$: (4.23) is fulfilled with $\Psi(n) = \ln^2(n)$ in case (i) and $\Psi(n) = n^{3/\theta}$ in case (ii). □

Comments

- Under suitable conditions on the function f the process $(X_i)_{i\geq 1-k}$ generated by the autoregressive model (2.1) is stationary and geometrically (M,θ)-mixing. More precisely, the classical condition is (see Doukhan [14], Th. 7, p. 102):

(H*) (i) The ε_i's are independent and independent of the initial variables $X_0, \ldots, X_{-k+1}$.
(ii) There exists non negative constants $a_1, \ldots, a_k$ and positive constants c_0 and c_1 such that $|f(x)| \leq \sum_{i=1}^k a_i|x_i| - c_1$ if $\max_{i=1,\ldots,k} |x_i| > c_0$ and the unique nonnegative real zero x_0 of the polynomial $P(z) = z^k - \sum_{i=1}^k a_i z^{k-i}$ satisfies $x_0 < 1$. Moreover, the Markov chain $(\vec{X}_i)$ is irreducible with respect to the Lebesgue measure on $\mathbb{R}^k$.
In particular, the irreducibility condition for the Markov chain $(\vec{X}_i)$ is satisfied as soon as μ_ε is equivalent to the Lebesgue measure.

- Examples of arithmetically mixing processes corresponding to the autoregressive model (2.1) can be found in Ango Nze [3].

6. Proof of Theorem 3.1 and Corollary 3.1

In order to detail the steps of the proofs, we demonstrate consecutive claims. From now on we fix some $m \in \mathcal{M}_n$ to be chosen at the end of the proof.

Claim 1: We have

$$\|f_A - \tilde{f}\|_n^2 \leq \|f_A - f_m\|_n^2 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i(\tilde{f} - f_m)(\vec{X}_i) + \text{ pen}(m) - \text{pen}(\hat{m}). \tag{6.1}$$

Proof. Starting from (2.3) we know that $\gamma_n(\tilde{f}) - \gamma_n(f_m) \leq \text{pen}(m) - \text{pen}(\hat{m})$ and since $\gamma_n(\tilde{f}) - \gamma_n(f_m) = \|f - \tilde{f}\|_n^2 - \|f - f_m\|_n^2 - 2n^{-1}\sum_{i=1}^n \varepsilon_i(\tilde{f} - f_m)(\vec{X}_i)$, the claim is proved for f_A replaced by f namely

$$\|f - \tilde{f}\|_n^2 \leq \|f - f_m\|_n^2 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i(\tilde{f} - f_m)(\vec{X}_i) + \text{ pen}(m) - \text{pen}(\hat{m}). \tag{6.2}$$

Noticing that if t is a A-supported function then $\|f - t\|_n^2 = \|f\mathbb{I}_{A^c}\|_n^2 + \|f_A - t\|_n^2$ and applying this identity to $t = \tilde{f}$ and $t = f_m$, we obtain the claim from (6.2) after simplification by $\|f\mathbb{I}_{A^c}\|_n^2$. □

Recall that Ω_n is defined by equation (3.6), and for each $m' \in \mathcal{M}_n$, let

$$G_1(m') = \sup_{t \in B_{m'}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i t(\vec{X}_i),$$

where $B_{m'} = \{t \in S_m + S_{m'} / \ \|t\| \leq 1\}$.

The key of Theorem 3.1 relies on the following proposition which is proved in Section 7.

Proposition 6.1. *Under* **(H$_{(X,\varepsilon)}$)** *for all* $m' \in \mathcal{M}_n$

$$\mathbb{E}\left[\left(G_1^2(m') - (p_1(m') + p_2(m))\right)_+ \mathbb{I}_{\Omega_n}\right] \leq 1.6\kappa s^2 \frac{\mathrm{e}^{-L_{m'}D_{m'}}}{n}$$

where $p_1(m') = \kappa s^2 D_{m'}(1 + L_{m'})/n$, $p_2(m) = \kappa s^2 D_m/n$ *and* κ *is a universal constant (that can be taken to be 38).*

Next, we show

Claim 2: There exists a universal constant C such that

$$C^{-1}\mathbb{E}\left[\|f_A - \tilde{f}\|_n^2 \mathbb{I}_{\Omega_n}\right] \leq \|f_A - f_m\|^2 + \text{pen}(m) + s^2\frac{\Sigma_n}{n}\cdot$$

Proof. From Claim 1 we deduce

$$\|f_A - \tilde{f}\|_n^2 \leq \|f_A - f_m\|_n^2 + 2\|\tilde{f} - f_m\| G_1(\hat{m}) + \text{ pen}(m) - \text{pen}(\hat{m}). \tag{6.3}$$

On Ω_n, we can ensure that $\|\tilde{f} - f_m\| \leq \sqrt{2}\|\tilde{f} - f_m\|_n$, therefore the following inequalities hold

$$\begin{aligned} 2\|\tilde{f} - f_m\| G_1(\hat{m}) &\leq 2\|\tilde{f} - f_m\|_n \sqrt{2} G_1(\hat{m}) \leq \frac{1}{4}\|\tilde{f} - f_m\|_n^2 + 8G_1^2(\hat{m}) \\ &\leq \frac{1}{4}\left(\|\tilde{f} - f_A\|_n + \|f_A - f_m\|_n\right)^2 + 8G_1^2(\hat{m}) \\ &\leq \frac{1}{2}\left(\|\tilde{f} - f_A\|_n^2 + \|f_A - f_m\|_n^2\right) + 8G_1^2(\hat{m}). \end{aligned} \tag{6.4}$$

Combining (6.3) and (6.4) leads on Ω_n to,

$$\begin{aligned} \|f_A - \tilde{f}\|_n^2 &\leq \|f_A - f_m\|_n^2 + \frac{1}{2}\|f_A - \tilde{f}\|_n^2 + \frac{1}{2}\|f_A - f_m\|_n^2 + \text{pen}(m) + \ 8G_1^2(\hat{m}) - \text{pen}(\hat{m}) \\ &\leq \|f_A - f_m\|_n^2 + \frac{1}{2}\|f_A - \tilde{f}\|_n^2 + \frac{1}{2}\|f_A - f_m\|_n^2 + \text{pen}(m) \\ &\quad + \ 8p_2(m) + 8(G_1^2(\hat{m}) - (p_1(\hat{m}) + p_2(m))_+ \\ &\quad + \ 8p_1(\hat{m}) - \text{pen}(\hat{m}). \end{aligned} \tag{6.5}$$

By taking $\vartheta \geq 8\kappa$, we have

$$\text{pen}(m') \geq 8p_1(m'),$$

for all $m' \in \mathcal{M}_n$ and $8p_2(m) \leq \text{pen}(m)$. Thus we derive from (6.5)

$$\frac{1}{2}\|f_A - \tilde{f}\|_n^2 \mathbb{I}_{\Omega_n} \leq \frac{3}{2}\|f_A - f_m\|_n^2 + 2\text{pen}(m) + 8(G_1^2(\hat{m}) - (p_1(\hat{m}) + p_2(m))_+ \mathbb{I}_{\Omega_n},$$

and by taking the expectation on both sides of this inequality we get

$$\begin{aligned} \frac{1}{2}\mathbb{E}\left[\|f_A - \tilde{f}\|_n^2 \mathbb{I}_{\Omega_n}\right] &\leq \frac{3}{2}\|f_A - f_m\|^2 + 2\text{pen}(m) \\ &\quad + \ 8 \sum_{m' \in \mathcal{M}_n} \mathbb{E}\left[\left(G_1^2(m') - (p_1(m') + p_2(m))\right)_+ \mathbb{I}_{\Omega_n}\right]. \end{aligned}$$

We conclude by using Proposition 6.1 and (3.2), and by choosing m among $\mathcal{M}_n$ to minimize $m' \mapsto \|f_A - f_{m'}\|^2 + \text{pen}(m')$. This ends the proof of Theorem 3.1 with $C = 4$ and $C' = 16 \times 1.6\kappa$. □

For the proof of Corollary 3.1, we introduce the notation $\Pi_{\hat{m}}$ for the orthogonal projector (with respect to the usual inner product of $\mathbb{R}^n$) onto the $\mathbb{R}^n$-subspace $\{(t(\vec{X}_1), \ldots, t(\vec{X}_n))'/t \in S_{\hat{m}}\}$. It follows from the definition of the least-squares estimator that $(\tilde{f}(\vec{X}_1), \ldots, \tilde{f}(\vec{X}_n))' = \Pi_{\hat{m}} Y$. Denoting in the same way the function t and the vector $(t(\vec{X}_1), \ldots, t(\vec{X}_n))'$, we see that $\|f_A - \tilde{f}\|_n^2 = \|f_A - \Pi_{\hat{m}} f_A\|_n^2 + \|\Pi_{\hat{m}}\varepsilon\|_n^2 \leq \|f_A\|_n^2 + n^{-1}\sum_{i=1}^n \varepsilon_i^2$. Thus,

$$\mathbb{E}\left[\|f_A - \tilde{f}\|_n^2 \mathbb{I}_{\Omega_n^c}\right] \leq \|f_A\|_\infty^2 \mathbb{P}(\Omega_n^c) + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\varepsilon_i^2 \mathbb{I}_{\Omega_n^c}\right].$$

Let now x and y be positive constants to be chosen later, by a truncation argument we have

$$\mathbb{E}\left[\varepsilon_i^2 \mathbb{I}_{\Omega_n^c}\right] \leq x^2 \mathbb{P}(\Omega_n^c) + \mathbb{E}\left[\varepsilon_i^2 \mathbb{I}_{|\varepsilon_i|>x} \mathbb{I}_{\Omega_n^c}\right] \leq x^2 \mathbb{P}(\Omega_n^c) + \mathbb{E}\left[\varepsilon_i^2 \mathrm{e}^{y|\varepsilon_i| - yx} \mathbb{I}_{|\varepsilon_i|>x} \mathbb{I}_{\Omega_n^c}\right]$$
$$\leq x^2 \mathbb{P}(\Omega_n^c) + 2y^{-2}\mathrm{e}^{-yx}\mathbb{E}\left[\mathrm{e}^{2y|\varepsilon_i|}\mathbb{I}_{\Omega_n^c}\right]$$

by using in the last inequality that for all $u > 0$, $u^2\mathrm{e}^u/2 \leq \mathrm{e}^{2u}$. Now by $(\mathbf{H}_{X,\varepsilon})$ together with Hölder's inequality (we set $\bar{\ell}^{-1} = 1 - \ell^{-1}$) we have

$$\mathbb{E}\left[\mathrm{e}^{2y|\varepsilon_i|}\mathbb{I}_{\Omega_n^c}\right] \leq \mathbb{E}^{1/\bar{\ell}}\left[\mathrm{e}^{2y\bar{\ell}|\varepsilon_i|}\right]\mathbb{P}^{1/\ell}(\Omega_n^c) \leq 2^{1/\bar{\ell}}\mathrm{e}^{2y^2\bar{\ell}s^2}\mathbb{P}^{1/\ell}(\Omega_n^c).$$

Thus we deduce that

$$\mathbb{E}\left[\|f_A - \tilde{f}\|_n^2 \mathbb{I}_{\Omega_n^c}\right] \leq (\|f_A\|_\infty^2 + x^2)\mathbb{P}(\Omega_n^c) + 2^{1+1/\bar{\ell}}y^{-2}\mathrm{e}^{2y^2\bar{\ell}s^2 - yx}\mathbb{P}^{1/\ell}(\Omega_n^c).$$

We now choose $x = 2\sqrt{\bar{\ell}}s$ and $y = 1/x$ and under $(\mathbf{H}_\mu)$ we get

$$\mathbb{E}\left[\|f_A - \tilde{f}\|_n^2 \mathbb{I}_{\Omega_n^c}\right] \leq \left[(\|f_A\|_\infty^2 + 4\bar{\ell}s^2)C_\ell + 2^{3+1/\bar{\ell}}\mathrm{e}^{-1/2}C_\ell^{1/\ell}\bar{\ell}s^2\right]\frac{1}{n}\,.$$

The proof of Corollary 3.1 is completed by combining this inequality with the result of Claim 2.

Moreover, if for all $m \in \mathcal{M}_n$, $\mathbb{1} \in S_m$ then we notice that all along the proof, f can be replaced by $f + c = g$ where c is a given constant. Indeed, in this case, $g_m = f_m + c$, $\hat{g}_m = \hat{f}_m + c$, so that $f - f_m = g - g_m$ and $f - \hat{f}_m = g - \hat{g}_m$. If we choose $c = -\int f_A \mathrm{d}\mu$, we find the same result with $\|f_A\|_\infty$ replaced by $\|f_A - \int f_A \mathrm{d}\mu\|_\infty$ in the last inequality. □

7. Proof of Proposition 6.1

7.1. A key lemma

To prove the proposition we use the following lemma which is inspired by a work on exponential inequalities for martingales due to Meyer [23] (Prop. 4, p. 168).

Lemma 7.1. *Assume that Condition* $(\mathbf{H}_{X,\varepsilon})$ *holds, then for any positive numbers* ϵ, v *we have:*

$$\mathbb{P}\left[\sum_{i=1}^n \varepsilon_i t(\vec{X}_i) \geq n\epsilon, \quad \|t\|_n^2 \leq v^2\right] \leq \exp\left(-\frac{n\epsilon^2}{2s^2v^2}\right). \tag{7.1}$$

Proof. Let $M_n = \sum_{i=1}^n \varepsilon_i t(\vec{X}_i)$, $M_0 = 1$ and $\mathcal{G}_n$ the σ-field generated by the ε_i's, for $i < n$ and the $\vec{X}_i$'s for $i \leq n$. Note that $\mathbb{E}(M_n) = 0$. For each $\lambda > 0$ we have

$$\mathbb{P}\left[M_n \geq n\epsilon,\ \|t\|_n^2 \leq v^2\right] \leq \exp\left(-\lambda n\epsilon + nv^2s^2\lambda^2/2\right)\mathbb{E}\left[\exp(\lambda M_n - \lambda^2 n\|t\|_n^2 s^2/2)\right].$$

Let

$$Q_n = \exp\left(\lambda M_n - \frac{1}{2}\lambda^2 s^2 n\|t\|_n^2\right) = \exp\left(\lambda M_n - \frac{1}{2}\lambda^2 s^2 \sum_{i=1}^n t^2(\vec{X}_i)\right)$$

we find that:

$$\begin{aligned}\mathbb{E}(Q_n|\mathcal{G}_n) &= Q_{n-1}\mathbb{E}\left[\exp\left(\lambda(M_n - M_{n-1}) - \frac{1}{2}\lambda^2 s^2 t^2(\vec{X}_n)\right)|\mathcal{G}_n\right] \\ &= Q_{n-1}\exp\left(-\frac{1}{2}\lambda^2 s^2 t^2(\vec{X}_n)\right)\mathbb{E}\left(\exp(\lambda\varepsilon_n t(\vec{X}_n))|\mathcal{G}_n\right) \\ &\leq Q_{n-1}\exp\left(-\frac{1}{2}\lambda^2 s^2 t^2(\vec{X}_n)\right)\exp\left(\frac{1}{2}\lambda^2 s^2 t^2(\vec{X}_n)\right) = Q_{n-1},\end{aligned}$$

using the independence between ε_n and $\vec{X}_n$ together with Assumption $(\mathbf{H}_{X,\varepsilon})$. Then $\mathbb{E}Q_n \leq \mathbb{E}Q_{n-1}$ which leads to $\mathbb{E}Q_n \leq \mathbb{E}Q_0 = 1$. Thus

$$\mathbb{P}\left[M_n \geq n\epsilon,\ \|t\|_n^2 \leq v^2\right] \leq \exp(-n\sup_{\lambda>0}(\lambda\epsilon - \lambda^2 s^2 v^2/2)) = \exp\left(-n\frac{\epsilon^2}{2s^2v^2}\right).$$

This proves (7.1). □

7.2. **Proof of Proposition 6.1**

Throughout this section we set

$$Z_n(t) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i t(\vec{X}_i).$$

The proof of Proposition 6.1 is based on a chaining argument which has also been used by van de Geer [30] for an analogous purpose. Indeed it is well known (see Lorentz *et al.* [21], Chap. 15, Prop. 1.3, p. 487) that, in a linear subspace $S \subset \mathbb{L}_2(A,\mu)$ of dimension D, we can find a finite δ-net, $T_\delta \subset \mathbb{B}$, where $\mathbb{B}$ denotes the unit ball of S, such that

- for each $0 < \delta < 1$, $|T_\delta| \leq \left(\frac{3}{\delta}\right)^D$;
- for each $t \in \mathbb{B}$, there exists $t_\delta \in T_\delta$ such that $\|t - t_\delta\| \leq \delta$.

We apply this result to the linear space $S_m + S_{m'}$ of dimension $D(m') \leq D_m + D_{m'}$. We consider δ_k-nets, $T_k = T_{\delta_k}$, with $\delta_k = \delta_0 2^{-k}$ ($\delta_0 < 1$ that is to be chosen later) and we set $H_k = \ln(|T_k|)$. Given some point $t \in B_{m'} = \{t \in S_m + S_{m'} / \ \|t\| \leq 1\}$, we can find a sequence $\{t_k\}_{k\geq 0}$ with $t_k \in T_k$ such that $\|t - t_k\|^2 \leq \delta_k^2$. Thus we have the following decomposition that holds for any $t \in B_{m'}$

$$t = t_0 + \sum_{k=1}^{\infty}(t_k - t_{k-1}).$$

Clearly $\|t_0\| \leq 1$ and for all $k \geq 1$, $\|t_k - t_{k-1}\|^2 \leq 2(\delta_k^2 + \delta_{k-1}^2) = 5\delta_{k-1}^2/2$. In the sequel we denote by $\mathbb{P}_n(.)$ the measure $\mathbb{P}(. \cap \Omega_n)$ (actually only the inequality $\|t\|_n^2 \leq \frac{3}{2}\|t\|^2$ holding for any $t \in S_m + S_{m'}$ is required). Let $(x_k)_{k\geq 0}$ be a sequence of positive numbers that will be chosen later on. Let us set

$$\Delta = \sqrt{3s^2}\left(\sqrt{x_0} + \sum_{k\geq 1}\delta_{k-1}\sqrt{5x_k/2}\right),$$

we have that

$$\mathbb{P}_n\left[\sup_{t\in B_{m'}} Z_n(t) > \Delta\right] = \mathbb{P}_n\left[\exists (t_k)_{k\in\mathbb{N}} \in \prod_{k\in\mathbb{N}} T_k \ / \ Z_n(t_0) + \sum_{k=1}^{+\infty} Z_n(t_k - t_{k-1}) > \Delta\right]$$
$$\leq P_1 + P_2$$

where

$$P_1 = \sum_{t_0\in T_0} \mathbb{P}_n\left[Z_n(t_0) > \sqrt{3s^2x_0}\right],$$
$$P_2 = \sum_{k=1}^{\infty} \sum_{\substack{t_{k-1}\in T_{k-1}\\ t_k\in T_k}} \mathbb{P}_n\left[Z_n(t_k - t_{k-1}) > \delta_{k-1}\sqrt{15s^2x_k/2}\right].$$

Since on Ω_n, $\|t\|_n^2 \leq (3/2)\|t\|^2$ for each $t \in S_m + S_{m'}$, we deduce from Lemma 7.1 that for all $x > 0$

$$\mathbb{P}\left[\left\{Z_n(t) \geq \sqrt{3}s\|t\|\sqrt{x}\right\} \cap \Omega_n\right] \leq \exp(-nx). \tag{7.2}$$

Applying repeatedly this inequality with $t = t_0 \in T_0$ ($\|t_0\| \leq 1$) and with $t = t_k - t_{k-1}$ ($\|t_k - t_{k-1}\|^2 \leq 5\delta_{k-1}^2/2$), we get $P_1 \leq \exp(H_0 - nx_0)$ and $P_2 \leq \sum_{k\geq 1}\exp(H_{k-1} + H_k - nx_k)$. We now choose x_0 such that

$$nx_0 = H_0 + L_{m'}D_{m'} + \tau$$

and for $k \geq 1$, x_k is chosen to satisfy

$$nx_k = H_{k-1} + H_k + kD(m') + L_{m'}D_{m'} + \tau.$$

If $D(m') \geq 1$ then $kD(m') \geq k$ and $\sup_{t\in B_{m'}} Z_n(t)$ being nonnegative we derive

$$\mathbb{P}_n\left(\sup_{t\in B_{m'}} Z_n^2(t) > 3s^2\left[\sqrt{x_0} + \sum_{k\geq 1}\delta_{k-1}\sqrt{5x_k/2}\right]^2\right) \leq \mathrm{e}^{-\tau}\mathrm{e}^{-L_{m'}D_{m'}}\left(1 + \sum_{k=1}^{\infty}\mathrm{e}^{-k}\right)$$
$$\leq 1.6\mathrm{e}^{-\tau}\mathrm{e}^{-L_{m'}D_{m'}}. \tag{7.3}$$

Else, $S_m + S_{m'} = \{0\}$ and obviously (7.3) holds.

Now, it remains to show

$$3ns^2\left(\sqrt{x_0} + \sum_{k\geq 1}\delta_{k-1}\sqrt{5x_k/2}\right)^2 \leq \kappa s^2(D_{m'}(1 + L_{m'}) + D_m + \tau).$$

Indeed by integrating (7.3) with respect to τ we obtain the expected result

$$\mathbb{E}\left[\left(G_1^2(m') - \kappa s^2\frac{D_{m'}(1 + L_{m'}) + D_m}{n}\right)_+ \mathbb{I}_{\Omega_n}\right] \leq 1.6\kappa s^2\frac{\mathrm{e}^{-L_{m'}D_{m'}}}{n}$$

reminding that $G_1(m') = \sup_{t\in B_{m'}} Z_n(t)$.

By Schwarz inequality, we know

$$\left(\sqrt{x_0}+\sum_{k\geq 1}\delta_{k-1}\sqrt{5x_k/2}\right)^2 \leq \left(1+\sum_{k\geq 1}\delta_{k-1}\right)\left(x_0+\frac{5}{2}\sum_{k\geq 1}\delta_{k-1}x_k\right)$$
$$= (1+2\delta_0)\left(x_0+\frac{5}{2}\sum_{k\geq 1}\delta_{k-1}x_k\right).$$

We set $c=c(\delta_0)=\max\{2\ln(2)+1,\ln(9/2\delta_0^2)\}\geq 1$. Since for all $k\geq 0$ $H_k\leq \ln(3/\delta_k)D(m')$, we have for all k

$$\begin{aligned} nx_k &\leq (\ln(9/2\delta_0^2)+k(1+2\ln(2)))D(m')+L_{m'}D_{m'}+\tau \\ &\leq c(k+1)D(m')+L_{m'}D_{m'}+\tau \\ &\leq c(k+1)\left(D_m+D'_m(1+L_{m'})+\tau\right). \end{aligned}$$

Thus,

$$\begin{aligned} n\left(x_0+\frac{5}{2}\sum_{k\geq 1}\delta_{k-1}x_k\right) &\leq c\left(1+5\delta_0\sum_{k=1}^{\infty}(k+1)2^{-k}\right)(D_{m'}(1+L_{m'})+D_m+\tau) \\ &\leq c(1+15\delta_0)(D_{m'}(1+L_{m'})+D_m+\tau), \end{aligned}$$

and the result follows since $3c(1+2\delta_0)(1+15\delta_0)\leq 38=\kappa$ for $\delta_0=0.0138$. □

8. Proof of Propositions 4.1 and 4.2

First we check that equation (3.2) leads to a finite Σ_n. Using the classical inequality on the binomial coefficients

$$\ln(C_{2^j}^{K_j})\leq K_j\left(1+\ln(2^j/K_j)\right),$$

we get

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

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

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

Thus by applying Corollary 3.1 with

$$\mathrm{pen}(m)=\vartheta s^2\frac{D_m}{n}(1+L(a)),$$

we obtain by using (4.3)

$$\mathbb{E}\left[\|f_A-\tilde{f}\|_n^2\right] \leq C \inf_{J\in\{0,\ldots,J_n\}}\left[\|f_A-\tilde{f}_J\|^2+\vartheta s^2\frac{C_a 2^J}{n}(1+L(a))\right] + C'\frac{s^2\Sigma_n}{n} \;+\; C''\frac{R_2+s^2}{n}, \tag{8.1}$$

where $C_a = 1+\sum_{j\geq 1} j^{-a}$. We know from Birgé and Massart [8] that $\forall f \in \mathbb{B}_{\alpha,p,\infty}(R_1,R_2)$, $\forall J \in \{0,\ldots,J_n\}$ there exists some $\tilde{f}_J \in \bigcup_{m\in\mathcal{M}_n^J} S_m$ such that

- if $r \geq \alpha > (1/p-1/2)_+$

$$\|f-\tilde{f}_J\| \leq \sqrt{h_1}\|f-\tilde{f}_J\|_\nu \leq C(h_1,R_1,\Gamma)\left[2^{-\alpha J}+\left(\frac{n}{\ln^b(n)}\right)^{-\alpha+(1/p-1/2)_+}\right] \tag{8.2}$$

- if $r \geq \alpha > 1/p$

$$\|f-\tilde{f}_J\| \leq \|f-\tilde{f}_J\|_\infty \leq C(R_1,\Gamma)\left[2^{-\alpha J}+\left(\frac{n}{\ln^b(n)}\right)^{-\alpha+1/p}\right]. \tag{8.3}$$

By minimizing (8.1) with respect to J and using (8.2) (respectively (8.3)) we obtain (4.4) (respectively (4.5)) noting that for $\alpha > \alpha_p$ (respectively $\alpha > \alpha'_p$)

$$\left(\frac{n}{\ln^b(n)}\right)^{-\alpha+(1/p-1/2)_+} \leq n^{-2\alpha/(2\alpha+1)}$$

(respectively $(n/\ln^b(n))^{-\alpha+1/p} \leq n^{-2\alpha/(2\alpha+1)}$) at least for n large enough.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, in *Proc. 2nd International Symposium on Information Theory*, edited by P.N. Petrov and F. Csaki. Akademia Kiado, Budapest (1973) 267-281.

[2] H. Akaike, A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19** (1984) 716-723.

[3] P. Ango Nze, Geometric and subgeometric rates for markovian processes in the neighbourhood of linearity. *C. R. Acad. Sci. Paris* **326** (1998) 371-376.

[4] Y. Baraud, Model selection for regression on a fixed design. *Probab. Theory Related Fields* **117** (2000) 467-493.

[5] Y. Baraud, *Model selection for regression on a random design*, Preprint 01-10. DMA, École Normale Supérieure (2001).

[6] Y. Baraud, F. Comte and G. Viennet, Adaptive estimation in autoregression or β-mixing regression *via* model selection. *Ann. Statist.* (to appear).

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

[8] L. Birgé and P. Massart, An adaptive compression algorithm in Besov spaces. *Constr. Approx.* **16** (2000) 1-36.

[9] L. Birgé and Y. Rozenholc, *How many bins must be put in a regular histogram*. Working paper (2001).

[10] A. Cohen, I. Daubechies and P. Vial, Wavelet and fast wavelet transform on an interval. *Appl. Comput. Harmon. Anal.* **1** (1993) 54-81.

[11] I. Daubechies, *Ten lectures on wavelets*. SIAM: Philadelphia (1992).

[12] R.A. Devore and C.G. Lorentz, *Constructive Approximation*. Springer-Verlag (1993).

[13] D.L. Donoho and I.M. Johnstone, Minimax estimation *via* wavelet shrinkage. *Ann. Statist.* **26** (1998) 879-921.

[14] P. Doukhan, *Mixing properties and examples*. Springer-Verlag (1994).

[15] M. Duflo, *Random Iterative Models*. Springer, Berlin, New-York (1997).

[16] M. Hoffmann, On nonparametric estimation in nonlinear AR(1)-models. *Statist. Probab. Lett.* **44** (1999) 29-45.

[17] I.A. Ibragimov, On the spectrum of stationary Gaussian sequences satisfying the strong mixing condition I: Necessary conditions. *Theory Probab. Appl.* **10** (1965) 85-106.

[18] M. Kohler, *On optimal rates of convergence for nonparametric regression with random design*, Working Paper. Stuttgart University (1997).
[19] A.R. Kolmogorov and Y.A. Rozanov, On the strong mixing conditions for stationary Gaussian sequences. *Theory Probab. Appl.* **5** (1960) 204-207.
[20] K.C. Li, Asymptotic optimality for C_p, C_l cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15** (1987) 958-975.
[21] G.G. Lorentz, M. von Golitschek and Y. Makokov, *Constructive Approximation, Advanced Problems*. Springer, Berlin (1996).
[22] C.L. Mallows, Some comments on C_p. *Technometrics* **15** (1973) 661-675.
[23] A. Meyer, *Quelques inégalités sur les martingales d'après Dubins et Freedman*, Séminaire de Probabilités de l'Université de Strasbourg. Vols. 68/69 (1969) 162-169.
[24] D.S. Modha and E. Masry, Minimum complexity regression estimation with weakly dependent observations. *IEEE Trans. Inform. Theory* **42** (1996) 2133-2145.
[25] D.S. Modha and E. Masry, Memory-universal prediction of stationary random processes. *IEEE Trans. Inform. Theory* **44** (1998) 117-133.
[26] M. Neumann and J.-P. Kreiss, Regression-type inference in nonparametric autoregression. *Ann. Statist.* **26** (1998) 1570-1613.
[27] B.T. Polyak and A. Tsybakov, A family of asymptotically optimal methods for choosing the order of a projective regression estimate. *Theory Probab. Appl.* **37** (1992) 471-481.
[28] R. Shibata, Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** (1976) 117-126.
[29] R. Shibata, An optimal selection of regression variables. *Biometrika* **68** (1981) 45-54.
[30] S. Van de Geer, Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.* **23** (1995) 1779-1801.
[31] V.A. Volonskii and Y.A. Rozanov, Some limit theorems for random functions. I. *Theory Probab. Appl.* **4** (1959) 179-197.