



HAL
open science

Model selection and randomization for weakly dependent time series forecasting

Pierre Alquier, Olivier Wintenberger

► **To cite this version:**

Pierre Alquier, Olivier Wintenberger. Model selection and randomization for weakly dependent time series forecasting. 2009. hal-00362151v1

HAL Id: hal-00362151

<https://hal.science/hal-00362151v1>

Preprint submitted on 17 Feb 2009 (v1), last revised 3 Jul 2012 (v4)

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

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

MODEL SELECTION AND AGGREGATION FOR WEAKLY DEPENDENT TIME SERIES FORECASTING

PIERRE ALQUIER⁽¹⁾ AND OLIVIER WINTENBERGER⁽²⁾

ABSTRACT. Observing a stationary time series, we propose in this paper new procedures in two steps for the prediction of the next value of the time series. Following machine learning theory paradigm, the first step consists in determining randomized estimators, or "experts", in (possibly numerous) different predictive models. In the second step estimators are obtained by model selection or randomization associated with exponential weights of these experts. We prove Oracle inequalities for both estimators and provide some applications for linear, artificial Neural Networks and additive non-parametric predictors.

1. INTRODUCTION

When observing a time series, one crucial issue is to predict first future value with the observed past values. Since the seminal works of Akaike, see for example [1], different model selection procedures have been studied for inferring how many observed past values are needed for predicting the next value. Efficiency of different penalized empirical risk minimizers such that AIC, BIC, Mallows, APE's predictors have been proved when the observations satisfy a linear auto-regressive model, see for instance Ing [17]. The main issue in this context is to determine the order of an efficient predictive linear autoregressive model and then to estimate its coefficients. There the model fitted by the observations is assumed to belong into the same class than the predictive models.

In the same time, model selection procedure have been hugely improved using learning theory in the independent and identically distributed (iid for short) case, see Vapnik [27] and Massart [20] among others. Results such that Oracle inequalities have been settled in very extended context. Even if the true model does not belong into one of the models proposed by the experts recent procedures ensure that the risk is as small as possible. However, few works have been done for dependent observations, principally in two direction: penalized least square and randomization techniques. Baraud *et al.* [5] proved Oracle inequalities with respect to the quadratic loss and under β -mixing condition. Their penalized empirical risk minimizers select an efficient predictive model when the number of useful past values is known. Recently, the theory of individual sequences leads also to Oracle inequalities for risk of prediction. Randomization with exponential weights of experts advices predicts the observations as if it was a deterministic sequence. We refer the reader to Lugosi and Cesa-Bianchi [19] for more details. Good predictors are then obtained given the expert devices. But the form of the expert devices given the observations is not given

2000 *Mathematics Subject Classification.* Primary 62M20; Secondary 62G05, 68Q32, 68T05, 62M10, 91E40.

Key words and phrases. Time series prediction, autoregression estimation, adaptive inference, statistical learning, randomized estimators, aggregation of estimators, model selection, weak dependence.

and then the form of the predictors is not tractable.

In this paper, we give Oracle inequalities for the \mathbb{L}^1 -risk of prediction of some stationary time series. We introduce two new procedures that find an efficient predictive model associated with an efficient number of past values. To prove this we use the PAC-Bayesian approach introduced by McAllester [21]. This general theoretical framework has proved to efficiently give Oracle inequalities in many iid frameworks, see Catoni [7, 8, 9], Audibert [4] and Alquier [2]. There exist procedures and Oracle inequalities in the dependent cases, see Baraud *et al.* [5] and Modha and Masry [23]. In Modha and Masry [23], their procedure use the α -mixing coefficients of the observations. To our knowledge, there is no efficient estimation of this coefficients and their procedure is not implementable in practice. In Baraud *et al.* [5], the Oracle inequality holds only if the β -mixing coefficients and the prediction procedure satisfy together intricate conditions. Here again, as β -mixing coefficients are not estimable there is no way to check those conditions. In this paper, the prediction procedures are for the first time completely free of the dependence properties of the observations. It represents an important progress for learning theory applications with dependent observations.

Let us assume that we observe (X_1, \dots, X_n) from a stationary time series $X = (X_t)_{t \in \mathbb{Z}}$ distributed as π_0 on $\mathcal{X}^{\mathbb{Z}}$ where \mathcal{X} is an Hilbert space equipped with its usual norm $\|\cdot\|$. For each θ in the set of parameter Θ we associate a $p(\theta)$ -autoregressive function f_θ from $\mathcal{X}^{p(\theta)}$ to \mathcal{X} that represents a predictive model. Then each $\theta \in \Theta$ is associated with a predictor $f_\theta(X_{n-1}, \dots, X_{n-p(\theta)})$. The risk of prediction is the absolute loss $R(\theta)$ defined as:

$$R(\theta) = \pi_0 \left[\left\| f_\theta(X_{p(\theta)}, \dots, X_1) - X_{p(\theta)+1} \right\|^2 \right],$$

where here and all along the paper $\pi[h] = \int h d\pi$ for any measure π and any integrable function h . The choice of this risk instead of the classic quadratic loss is due to its Lipschitzian property, very well suited with the dependence context here. The main objective of this paper is to determine two different procedures that give estimators $\hat{\theta}_n$ with associated risk $R(\hat{\theta}_n)$ satisfying an Oracle inequality - in other words, $R(\hat{\theta}_n)$ is not far from $\inf_{\Theta} R$.

As we have to deal with different models and different delays in the same time, it is convenient to split the set Θ in subsets of the form:

$$\Theta = \bigcup_{p=1}^{\lfloor \frac{n}{2} \rfloor} \Theta_p \text{ with } \Theta_p = \bigcup_{\ell=1}^{m_p} \Theta_{p,\ell},$$

where $m_p > 0$ has to be fixed carefully. The set Θ_p consists in different predictive models that need the same number of past values. To fix the idea, let us give the simple example additive non parametric predictive models when $\mathcal{X} = \mathbb{R}$, see Subsection 4.3 for more details. Let us define

$$\hat{X}_{n+1} = \sum_{i=0}^{\hat{p}} \hat{f}_i(X_{n-i}).$$

Then we fix $\hat{\theta}_n = ((\hat{f}_i)_{0 \leq i \leq \hat{p}})$. We split

$$\Theta = \bigcup_{p=1}^{\lfloor \frac{n}{2} \rfloor} \Theta_p = \bigcup_{p=1}^{\lfloor \frac{n}{2} \rfloor} \{(f_i)_{0 \leq i \leq p} \in \mathcal{A}_p\}$$

where C is a compact subset of \mathbb{R} and \mathcal{A}_p is a compact subset of \mathcal{F}^{p+1} for \mathcal{F} the set of integrable functions from \mathbb{R} to \mathbb{R} . Under suitable conditions on \mathcal{F} , there exists an ordered functional basis $(\varphi_i)_{i \geq 1}$. Then the index ℓ corresponds to the number of the first functionals in the basis that we consider. Then $f_i = \sum_{j=1}^{\ell} a_{i,j} \varphi_j$ for each i and $\Theta_{p,\ell} = \{(a_{i,j})_{0 \leq i \leq p, 1 \leq j \leq \ell}\}$.

The common first step of our two prediction procedures consists on proposing a randomized estimator $\tilde{\theta}_{p,\ell}$ for each subset $\Theta_{p,\ell}$. Then we propose two different estimators $\hat{\theta}$ and $\tilde{\theta}$ of a parameter θ associated with an efficient predictive model. The first procedure is a model selection that provides $(\hat{p}, \hat{\ell})$. It leads to the natural choice $\hat{\theta} = \tilde{\theta}_{\hat{p}, \hat{\ell}}$. Our model selection criterion for each indices (p, ℓ) is close to the following penalized empirical risk criterion

$$r_n(\hat{\theta}_{p,\ell}) + \sqrt{\frac{K_n d_{p,\ell}}{n-p}} \ln(d_{p,\ell} n),$$

where $r_n(\theta)$ is the empirical risk, $d_{p,\ell}$ is a measure of the complexity of $\Theta_{p,\ell}$, highly related to its dimension, and $K_n > 0$ is independent of p, ℓ . The second procedure is a second randomization step on the indexes (p, ℓ) that gives $(\tilde{p}, \tilde{\ell})$ and then leads to the corresponding estimator $\tilde{\theta} = \tilde{\theta}_{\tilde{p}, \tilde{\ell}}$. The exponential weights associated to each indices (p, ℓ) have the same form than the ones used for randomizing expert devices in the theory of individual sequence. They deeply depends on a parameter $K_n > 0$.

The value of K_n has to be fixed arbitrarily and it has lot of consequences on the sharpness of the Oracle inequalities we obtained. For bounded observations, the best is to fix it larger than some constant depending on the (non-estimable) dependence properties of the observations. If we fail, remark that a less good Oracle inequality still holds, see the results in Section 3. For possibly unbounded observations, we can fix it proportional to $\ln(n)$ independently on the observations. Such choice leads to an additional logarithmic term in the rate of convergence. But remark that even for K_n fixed as a constant we over-penalized the expected risk there is always additional logarithmic terms in the rates of the Oracle inequalities, see below. So we can fix as a rule of thumb $K_n = C \ln(n)$ for some known C and our procedure is free of the dependence properties of the observations, see Subsection 3.3 for more details.

Let us resume the main results of this paper for K_n fixed to $\ln(n)$. For bounded observations, we prove a Probably Approximately Correct Oracle inequality: for n large enough, with probability at least $1 - \varepsilon$

$$R(\hat{\theta}_n) \leq \min_{p,\ell} \left\{ \inf_{\Theta_{p,\ell}} R(\theta) + C \sqrt{\frac{d_{p,\ell}}{n-p}} \ln(d_{p,\ell}/\varepsilon) \ln^2(n) \right\},$$

where C is a constant. For possibly unbounded observations, we obtain Oracle inequalities in expectation. More precisely, we obtain that for n sufficiently large

$$\pi_0[R(\hat{\theta}_n)] \leq \min_{p,\ell} \left\{ \inf_{\Theta_{p,\ell}} R(\theta) + C \sqrt{\frac{d_{p,\ell}}{n-p}} \ln(d_{p,\ell}) \ln^2(n) \right\},$$

where C is constant. This result can be compared with those of Baraud, Comte and Viennet [5] and Modha and Masry [23]. They achieve respectively Oracles inequalities of the form

$$\begin{aligned} \pi_0[R'(\hat{\theta}_{p,n})] &\leq \left(1 + \frac{1}{C}\right)^2 \min_{\ell} \left\{ \inf_{\Theta_{p,\ell}} R'(\theta) + C^3 \frac{d_{p,\ell}}{n-p} \right\} \text{ for each } p, \\ \pi_0[R'(\hat{\theta}_n)] &\leq \left(1 + \frac{1}{C}\right) \min_{p,\ell} \left\{ \inf_{\Theta_{p,\ell}} R'(\theta) + C \left(\frac{K_n d_{p,\ell}}{n-p}\right)^c \ln(d_{p,\ell}) \right\} \end{aligned}$$

where R' is the excess quadratic risk, $0 < c < 1$ is a constant depending on the dependence structure of the observations and C is fixed by the statistician. Our Oracle inequalities are sharper than the ones of [23]. Baraud *et al.* [5] achieve the optimal rates and we fail, but with a loss in the constant. Moreover, as already noticed, those authors are not fully adaptive in p .

To obtain such Oracle inequalities, sharp exponential inequalities are used in the dependent setting. For this, weakly dependence properties on the observations are assumed. This dependent setting might be more general than the mixing one, see the monograph of Dedecker *et al.* [10]. Here we use in the bounded cases the θ_∞ -coefficients (also called γ -mixing coefficients) introduced in Rio [24] to derive a sharp Hoeffding inequality in the dependent framework. These coefficients generalize the uniform mixing ones. In the unbounded cases we use generic models called chains with infinite memory introduced by Doukhan and Wintenberger [14] that includes many classical econometric models such that ARMA, GARCH and LARCH. Here we work under restrictions of additive forms that unfortunately exclude unbounded volatility models, see Subsection 2.4 for more details. Our dependent framework is not comparable with the β - or α -mixing one as it deals with some dynamical systems that are not mixing, see Andrews [3] and Dedecker and Prieur [11] or details on these counter-examples.

The paper is organized as follows. First some notation, the framework and the predictors are introduced in Section 2. Then the Oracles inequalities and some comments follow in Section 3. The main results of this Section are applied for Linear predictors, artificial Neural Networks predictors and Non-parametric Auto-Regressive predictors in Section 4. Finally the proofs are collected in Section 5.

2. PRELIMINARIES

Let $X = (X_t)_{t \in \mathbb{Z}}$ be a stationary process taking values in a measurable Hilbert space $(\mathcal{X}, \mathcal{B})$ (with norm $\|\cdot\|$ and scalar product $\langle \cdot, \cdot \rangle$). Assume that X is distributed as π_0 .

2.1. The predictive models. For any $p \in \{1, \dots, n-1\}$ and any $\ell \in \{1, \dots, m_p\}$ with $m_p > 0$, any parameter θ of the set $\Theta_{p,\ell}$, compact subset of \mathbb{R}^q for some $q < \infty$, is identified with a function

$f_\theta : \mathcal{X}^p \rightarrow \mathcal{X}$. Let us assume that there exists a sequence $(a_j(\theta))_{j \in \{1, \dots, p\}}$ satisfying the relation

$$(2.1) \quad \sum_{j=1}^p a_j(\theta) \leq L.$$

such that for any $(x_1, \dots, x_p), (y_1, \dots, y_p) \in \mathcal{X}^p$ we have:

$$(2.2) \quad \left| f_\theta(x_1, \dots, x_p) - f_\theta(y_1, \dots, y_p) \right| \leq \sum_{j=1}^p a_j(\theta) \|x_j - y_j\|.$$

Moreover, we assume that the Θ_p are disjoint sets for all $p \in \{1, \dots, \lfloor n/2 \rfloor\}$ so that any $\theta \in \Theta$ belongs to one and only one Θ_p . We write $p(\theta)$ the corresponding value of p . Let us define the set of indexes:

$$M = \bigcup_{p=0}^{\lfloor \frac{n}{2} \rfloor} \{p\} \times \{1, \dots, m_p\}.$$

Finally, \mathcal{T} denotes a σ -algebra on Θ , and for any $(p, \ell) \in M$, $\mathcal{T}_{p,\ell}$ denote the restriction of \mathcal{T} on $\Theta_{p,\ell}$.

2.2. The risk. For a chosen $\theta \in \Theta$ from the observations, we check the ability of $\widehat{X}_N^\theta = f_\theta(X_{N-1}, \dots, X_{N-p(\theta)})$ to predict X_N for any $N \in \mathbb{Z}$. The error of prediction $R(\theta)$ is the expectation of the absolute loss of X_N by \widehat{X}_N^θ which do not depend on N by stationarity conditional on the value of θ :

$$R(\theta) = \pi_0 \left[\left\| X_1 - \widehat{X}_1^\theta \right\| \right].$$

The objective is to determine $\widehat{\theta}_n$ such that its risk is close to $R(\bar{\theta})$ where $\bar{\theta} \in \arg \min_{\theta \in \Theta} R(\theta)$. We define also the values $\bar{\theta}_{p,\ell}$ for any $(p, \ell) \in M$ by $\bar{\theta}_{p,\ell} \in \arg \min_{\theta \in \Theta_{p,\ell}} R(\theta)$.

The risk $R(\theta)$ cannot be computed as the distribution π_0 is unknown. So we introduce its empirical counterpart $r_n(\theta)$ as,

$$r_n(\theta) = \frac{1}{n - p(\theta)} \sum_{t=p(\theta)+1}^n \left\| X_t - \widehat{X}_t^\theta \right\|.$$

2.3. The estimators. For any model $(p, \ell) \in M$ let us choose a probability measure $\pi_{p,\ell}$ on the measurable space $(\Theta_{p,\ell}, \mathcal{T}_{p,\ell})$ - an usual choice for $\pi_{p,\ell}$ is the Lebesgue measure on $\Theta_{p,\ell}$ that is often a compact subset of \mathbb{R}^d for some $d > 0$, but note that the choice of the various parameters involved in this subsection is discussed later in the paper and illustrated by a simple example. Let us also choose some prior weights on the models: $w_{p,\ell} \geq 0$ such that $\sum_{(p,\ell) \in M} w_{p,\ell} \leq 1$. This choice will be discussed later.

For any measure π and any measurable function h such that $\pi[\exp(h)] < +\infty$, we define the Gibbs measure $\pi\{h\}$ through the equation:

$$(2.3) \quad \frac{d\pi\{h\}}{d\pi}(\theta) = \frac{\exp(h(\theta))}{\pi[\exp(h)]}.$$

Let us put

$$\mathcal{G} = \left\{ 8, 16, \dots, 2^{\lfloor \frac{\ln(n^2)}{\ln 2} \rfloor} \right\}.$$

Now let us choose some $K_n \geq 1$, see Remark 3.3.3 for a discussion of this choice. For any $(p, \ell) \in M$ and $\lambda > 0$ we define:

$$\hat{R}(p, \ell, \lambda) = -\frac{1}{\lambda} \ln \int_{\Theta_{p,\ell}} \exp(-\lambda r_n(\theta)) d\pi_{p,\ell}(\theta) + \frac{1}{\lambda} \ln \frac{|\mathcal{G}|}{w_{p,\ell}} + \frac{\lambda K_n^2}{n \left(1 - \frac{p}{n}\right)^2}.$$

Now, we propose, for any model $(p, \ell) \in M$ and parameter $\lambda \in \mathcal{G}$ the following estimation procedure: draw

$$\tilde{\theta}_{p,\ell}^\lambda \sim \pi_{p,\ell}\{-\lambda r_n\}.$$

Then, we propose two procedures to select a model $(p, \ell) \in M$ and a parameter $\lambda \in \mathcal{G}$.

The first procedure is a model selection procedure, we choose

$$(\hat{p}, \hat{\ell}, \hat{\lambda}) = \arg \min_{\substack{(p,\ell) \in M \\ \lambda \in \mathcal{G}}} \hat{R}(p, \ell, \lambda)$$

and we have the estimator

$$\hat{\theta} = \tilde{\theta}_{\hat{p}, \hat{\ell}}^{\hat{\lambda}}.$$

The second procedure proceeds by randomization on all the models. For any $\lambda \in \mathcal{G}$, we define the weights

$$w_{p,\ell}^\lambda = \frac{\exp\left(-\hat{R}(p, \ell, \lambda)\right)}{\sum_{(p', \ell') \in M} \exp\left(-\hat{R}(p', \ell', \lambda)\right)},$$

and we draw $(\tilde{p}^\lambda, \tilde{\ell}^\lambda)$ randomly according to the weights $(w_{p,\ell}^\lambda)_{(p,\ell)}$ and finally choose

$$\tilde{\lambda} = \arg \min_{\lambda \in \mathcal{G}} \hat{R}\left(\tilde{p}^\lambda, \tilde{\ell}^\lambda, \lambda\right)$$

and we have the estimator

$$\tilde{\theta} = \tilde{\theta}_{\tilde{p}^{\tilde{\lambda}}, \tilde{\ell}^{\tilde{\lambda}}}^{\tilde{\lambda}}.$$

2.4. Assumptions on the observations. In order to achieve our main results, we need to give some assumptions on the observations. We give below two very different settings of works. One is based on a specific (but wide) unbounded model so-called chain with infinite memory (or complete connection). The other one referred on the bounded case associated under a condition of weakly dependence type. See section 4 for some examples.

2.4.1. Chains with infinite memory. We study chains with Infinite memory introduced in [14] and we refer to it as Assumption **(CIM)**. Let ξ_t for $t \in \mathbb{Z}$ be independent and identically distributed variables distributed as μ on a Banach space \mathcal{X}' called the innovations. We assume that the innovations norm admits a Laplace transform, more formally that for all $c \in \mathbb{R}$, we have $\mu[\exp(c\|\xi_0\|)] < \infty$. We write this Laplace transform $\Psi(c) := \mu[\exp(c\|\xi_0\|)]$. We will say that **(CIM)** is satisfied if $X = (X_t)_{t \in \mathbb{Z}}$ is the solution of the equation

$$(2.4) \quad X_t = F(X_{t-1}, X_{t-2}, \dots; \xi_t) \text{ almost everywhere,}$$

for some function $F : \mathcal{X}^{(\mathbb{N} \setminus \{0\})} \times \mathcal{X}' \rightarrow \mathcal{X}$. Assume also that there exists some u satisfying, for all $x = (x_k)_{k \in \mathbb{N} \setminus \{0\}}$, $x' = (x'_k)_{k \in \mathbb{N} \setminus \{0\}} \in \mathcal{X}^{(\mathbb{N} \setminus \{0\})}$ such that there exists $N > 0$ as $x_k = x'_k = 0$ for all $k > N$, the condition

$$(2.5) \quad \|F(x; y) - F(x'; y')\| \leq \sum_{j=1}^{\infty} a_j(F) \|x_j - x'_j\| + u \|y - y'\|,$$

$$(2.6) \quad \text{with} \quad \sum_{j=1}^{\infty} a_j(F) := a(F) < 1,$$

Using directly Theorem 3.1 of [14] we derive the following proposition:

Proposition 2.1. *There exists a unique stationary causal solution X of equation (2.4) satisfying $\pi[\|X_0\|^r] < \infty$ for any $1 \leq r < \infty$.*

2.4.2. Bounded weakly dependent processes. We refer to this case, described below as Assumption **(WDP)**. In all this subsection we assume that X is bounded, i.e. $\|X\|_{\infty} < \infty$. In our context the appropriate weakly dependence notion is relying on the coefficients $\theta_{\infty, n}(1)$ introduced by Dedecker *et al.* [10]. This is a version of the γ -mixing of Rio [25] adapted to stationary time series. If Z is a bounded random variable on $(\Omega, \mathcal{A}, \mathbb{P})$, for any σ -algebra \mathfrak{G} of \mathcal{A} we put:

$$\theta_{\infty}(\mathfrak{G}, Z) = \sup_{f \in \Lambda_1} \left\| \mathbb{E}(f(Z) | \mathfrak{G}) - \mathbb{E}(f(Z)) \right\|_{\infty},$$

where Λ_1 is the set of real 1-Lipschitz functions and \mathbb{E} is the expectation with respect to the distribution of X . In our context it is convenient to define the σ -algebra $\mathfrak{G}_p = \sigma(X_t, t \leq p)$ for any $p \in \mathbb{Z}$ and

$$\theta_{\infty, k}(r) = \sup \left\{ \theta_{\infty}(\mathfrak{G}_p, (X_{j_1}, \dots, X_{j_\ell})), \quad p + r \leq j_1 < \dots < j_\ell, \quad 1 \leq \ell \leq k \right\},$$

Assumption **(WDP)** refers to the cases where $\theta_{\infty, n}(1)$ is well defined for the process $X = (X_t)_{t \in \mathbb{Z}}$. Let us give examples of time series satisfying **(WDP)**.

Bounded chains with infinite memory are θ_{∞} weakly-dependent. Suppose that X is the solution of equation (2.4) associated with an innovation which is bounded, i.e. $\|\xi_0\|_{\infty} < \infty$, then we have the following result,

Lemma 2.2. *Under condition (2.6) there exists a unique causal stationary process X solution of the equation (2.4). This solution is bounded by $u\|\xi_0\|_{\infty}/(1-a)$ and*

$$\theta_{\infty, n}(1) \leq 2 \frac{u\|\xi_0\|_{\infty}}{1-a(F)} \sum_{r=1}^n \inf_{0 < p < r} \left\{ a(F)^{r/p} + \sum_{j=p}^{\infty} a_j(F) \right\}.$$

The proof of Lemma 2.2 is given in the section dedicated to proofs, actually, in Subsection 5.5.

Uniform φ -mixing sequences are also θ_{∞} weakly-dependent. Let us recall the definition of the φ -mixing coefficients introduced in [16];

$$\varphi(r) = \sup_{(A, B) \in \mathfrak{G}_0 \times \mathfrak{F}_r} |\pi(B/A) - \pi(B)|$$

where $\mathfrak{F}_r = \sigma(Y_t, t \geq r)$. The class of φ -mixing processes is large, it includes in particular uniform ergodic Markov Chains, see [13].

Proposition 2.3. *If $(X_t)_{t \in \mathbb{Z}}$ is a stationary bounded (by $C > 0$) process then*

$$\theta_{\infty, n}(1) \leq 2C \sum_{r=1}^n \varphi(r).$$

The proof of the Proposition 2.3 is given in Subsection 5.5.

3. MAIN RESULTS

We first give a result that holds with a probability that may be as close as one as possible, then we give a result in expectation. Then some remarks on these two oracle inequalities are given. In all the sequel, we work under the assumption that, for every $(p, \ell) \in M$ there exists a constant $1 \leq d_{p, \ell} < \infty$ such that

$$(3.1) \quad \sup_{\gamma > e} \left\{ \frac{-\ln \pi_{p, \ell} [\exp(-\gamma (R - R(\bar{\theta}_{p, \ell})))]}{\ln(\gamma)} \right\} = d_{p, \ell}.$$

Even if this definition of the "dimension" of each sets $\Theta_{p, \ell}$ is non standard and comes artificially in this form from the PAC-Bayesian approach, it is linked with the standard notions of dimensions like the Vapnik or entropy one. More precisely we have the following result.

Proposition 3.1. *Let $\dim \in \mathbb{N}^*$, $x > 0$, and \mathcal{B}_x be the closed ℓ^1 -ball in \mathbb{R}^{\dim} of radius $x > 0$ and centered at 0. If we assume that $\Theta_{p, \ell} = \mathcal{B}_{c_{p, \ell}}$ for $c_{p, \ell} > 0$, that $\pi_{p, \ell}$ is the Lebesgue measure on $\Theta_{p, \ell}$ and that $\theta \rightarrow R(\theta)$ is a C -Lipschitz function then we have:*

$$(3.2) \quad d_{p, \ell} \leq \dim \times \left(1 + \ln \left(c_{p, \ell} \left(\frac{Ce}{\dim} \vee \frac{1}{c_{p, \ell} - \|\bar{\theta}_{p, \ell}\|} \right) \right) \right).$$

The proof of this result is given at the end of subsection 5.4.

3.1. Oracle inequality with large probability. The following inequality is a PAC result. It is very convenient to built confidence intervals of prediction, see [27] for example for more details on such confident intervals.

Theorem 3.2. *Let us assume that relation (3.1) is satisfied. Then for all n such that $n \ln^2 n \geq (8eK_n)^2$, under **(WDP)**, with π_0 -probability at least $1 - \varepsilon$ he have:*

$$R(\hat{\theta}) \text{ and } R(\tilde{\theta}) \leq \left(1 \vee \frac{2k_n^2}{k_n^2 + K_n^2} \right) \inf_{d_{p, \ell} \leq nK_n} \left\{ R(\bar{\theta}_{p, \ell}) + K_n \left[\frac{2 + (k_n/K_n)^2}{1 - p/n} \sqrt{\frac{d_{p, \ell}}{n}} \ln(d_{p, \ell} n) + \frac{4 \ln \frac{3 \ln n}{\varepsilon w_{p, \ell}}}{\sqrt{d_{p, \ell} n \ln(d_{p, \ell} n)}} \right] \right\}$$

where

$$(3.3) \quad k_n = \frac{\|X\|_{\infty} + 2\theta_{\infty, n}(1)}{1 + L}.$$

The proof of this result is given in the section dedicated to proofs, more precisely in Subsection 5.3 page 17.

3.2. Oracle inequality in expectation. The following oracle inequality holds in expectation. It is a weaker result than the previous one of Theorem 3.2 but the setting is more general as it holds under both **(WDP)** and **(CIM)**.

Theorem 3.3. *Let us assume that relation (3.1) is satisfied. Then for all n such that $n \ln^2 n \geq (8eK_n)^2$, we have:*

$$\begin{aligned} \pi_0[R(\hat{\theta})] \text{ and } \pi_0[R(\tilde{\theta})] &\leq \left(1 \vee \frac{2k_n^2}{k_n^2 + K_n^2}\right) \inf_{d_{p,\ell} \leq nK_n} \left\{ R(\bar{\theta}_{p,\ell}) \right. \\ &\quad \left. + K_n \left[\frac{2 + (k_n/K_n)^2}{1 - p/n} \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell}n) + \frac{4 \ln \frac{12 \ln n}{w_{p,\ell}}}{\sqrt{d_{p,\ell}n \ln(d_{p,\ell}n)}} \right] \right\} + \frac{3(1+L)\Psi(c^*) \ln(n)}{\sqrt{n}}. \end{aligned}$$

where as previously

$$k_n = \frac{\|X\|_\infty + 2\theta_{\infty,n}(1)}{1+L} \text{ under } \mathbf{(WDP)},$$

and where under **(CIM)**,

$$(3.4) \quad \theta_{\infty,n}^* = 1 + 2 \sum_{r=1}^n \inf_{0 < k < r} \left\{ a(F)^{r/k} + \sum_{j=k}^{\infty} a_j(F) \right\}, \quad c^* = \frac{u\theta_{\infty,n}^*}{2(1-a(F))} \text{ and } k_n = \frac{\ln n}{1+L}.$$

The proof of this result is given in Subsection 5.3 page 21.

3.3. Comments on the main results.

3.3.1. Comparison with other results. Oracles inequalities in expectation have already been proved in Modha and Masry [23] and Baraud *et al.* [5]. Their approach are based on traditional mixing coefficients and on classical penalized minimizers of the empirical risk estimators. As already said in the introduction, except the fact that they work with the quadratic loss, their results are very comparable with ours. Our rates are always smaller than those of Modha and Masry [23], as in their case it depends on the decrease rates of the mixing coefficients. The results in Baraud *et al.* [5] are very competitive with ours. They achieve the optimal rate of convergence, i.e. the optimal one in the iid case, but they pay it with a multiplicative constant larger than 1 in the oracle inequality. More important, their approach depends on the (unobservable) mixing properties of the observations through intricate conditions on the model dimension and on the penalization. This drawback of their approach is due to the use of the β -mixing coefficients of the time series. The weak dependence coefficients used here lead to a sharper Hoeffding type inequality than the β -mixing coefficients, see Rio [24] and its consequence the Theorem 5.6 of this paper. Then it is for the first time possible to consider here predictors free of the dependence properties of the observed time series. Remark that as the choice of the dependence framework is orthogonal to the one of the estimation procedures, it should be interesting to study classical penalized empirical risk minimizers in the weakly dependence framework used here.

3.3.2. Choice of the weights. When $k_n \geq K_n$ the order of convergence of $\pi_0[R(\hat{\theta})]$ to $R(\bar{\theta}_{p,\ell})$ is given by the expression

$$\frac{K_n}{k_n} \left[\sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell}n) + \frac{\ln \frac{\ln n}{w_{p,\ell}}}{\sqrt{nd_{p,\ell} \ln(d_{p,\ell}n)}} \right]$$

and so, by

$$\frac{K_n}{k_n} \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell} n)$$

as soon as

$$\frac{\ln \frac{\ln n}{w_{p,\ell}}}{\sqrt{nd_{p,\ell}} \ln(d_{p,\ell} n)} \leq \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell} n).$$

Then if the weights are chosen such that satisfying the condition

$$(3.5) \quad w_{p,\ell} \geq \frac{e^{-d_{p,\ell} \ln^2(nd_{p,\ell})}}{\ln(n)}.$$

As the sums of these weights have to be less than 1, such choice is possible if

$$\sum_{(p,\ell) \in M} e^{-d_{p,\ell}} \leq 1.$$

If this condition is not satisfied, there may be a loss in the bound on the risk of the estimator due to the choice of the weights. This loss is classical in learning theory and has nothing to do with the PAC-bayesian approach used here. In some way it means that we cannot perform an efficient model selection if we have too many models.

3.3.3. Choice of the parameter K_n . The best choice for the parameter K_n is k_n that depends on the known parameter L and on the non-observable dependence structure of the observations. So, we will discuss in practice this choice is delicate. In the sequel of this discussion we work under reasonable weak dependence and complexity conditions, i.e. that $\theta_{\infty,\infty}(1) < \infty$, $\theta_{\infty,\infty}^* < \infty$ if **(CIM)** and (3.1). For bounded φ -mixing processes it is ensure by the summability of the φ -coefficients. If we do not have any reason to assume that the process is bounded, we can fix $K_n = \ln n / (1 + L)$ and $w_{p,\ell}$ as in Subsection 3.3.2 in order to obtain the oracle inequality:

$$\pi_0[R(\hat{\theta})] \text{ and } \pi_0[R(\tilde{\theta})] \leq \inf_{p,\ell} \left\{ R(\bar{\theta}_{p,\ell}) + C \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell} n) \ln(n) \right\}$$

for some constant $C > 0$, as soon as n is sufficiently large. If we assume that the observations are bounded, we can get a refinement choosing K_n as an upper bound for $k_n = (\|X\|_\infty + 2\theta_{\infty,\infty}(1)) / (1 + L)$. If we are lucky and the relation $K_n \geq k_n$ is satisfied, we obtain under **(WDP)** and with probability $1 - \varepsilon$

$$R(\hat{\theta}) \text{ and } R(\tilde{\theta}) \leq \inf_{p,\ell} \left\{ R(\bar{\theta}_{p,\ell}) + C \sqrt{\frac{d_{p,\ell}}{n}} \ln \frac{d_{p,\ell} n}{\varepsilon} \right\}.$$

Remark that it is possible that L goes to infinity with n such that for any fixed $K_n = K$ then $K \geq k_n$ for large n . But then another loss in a power of $\ln(n)$ appears as the "dimension" $d_{p,\ell}$ grows with L , see the application on Neural Networks predictors in Subsection 4.2. If we do a mistake in the upper estimate on k_n , namely $K_n < k_n$, then a multiplicative constant $c \in]1, 2[$ deteriorates the oracle inequality under both **(CIM)** or **(WDP)**:

$$\pi_0[R(\hat{\theta})] \text{ and } \pi_0[R(\tilde{\theta})] \leq c \inf_{p,\ell} \left\{ R(\bar{\theta}_{p,\ell}) + C \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell} n) (k_n)^2 \right\}.$$

Such choice of small K_n no longer ensures the consistency of the estimator. So we recommend to choose in any cases the parameter $K_n = \ln n / (1 + L)$ that is free of dependence properties. As a consequence, we do an over-penalization and the procedure is very conservative, see the discussion based on simulations at the end of Subsection 4.1.

4. APPLICATIONS

In this section we investigate several possible predictors. Note than in all the applications, we work on unions of compact subsets of parameters $\Theta_{p,\ell}$ of \mathbb{R}^{dim} for some dimension $dim \in \mathbb{N}$ associated with the prior measure $\pi_{p,\ell}$ that is the Lebesgue one. The "dimension" $d_{p,\ell}$ is then closely related to dim thanks to Proposition 3.1.

4.1. Linear predictors. Let us first consider the case of linear auto-regressive predictions. More precisely, in the case $\mathcal{X} = \mathbb{R}$ we consider predictive models of the form:

$$f_\theta(X_{N-1}, \dots, X_{N-p}) = \theta_0 + \sum_{i=1}^p \theta_i X_{N-i},$$

where $\theta \in \Theta_p \subset \mathbb{R}^{p+1}$ with by definition, for some $c_p > 0$,

$$\Theta_p = \Theta_{p,1} = \left\{ \theta \in \mathbb{R}^p, \quad \|\theta\|_1 = \sum_{i=0}^p |\theta_i| \leq c_p \right\}.$$

In this simple case $m_p = 1$ for all p such that the index ℓ can be omitted in the sequel. Using Proposition 3.1 it follows that

$$d_{p,\ell} \leq (p+1) \left(1 + \ln \left(c_p \left(\frac{e}{p+1} \vee \frac{1}{c_p - \|\bar{\theta}_p\|} \right) \right) \right),$$

where $\bar{\theta}_p = \arg \min_{\theta \in \Theta_{p,\ell}} R(\theta)$. Let us fix the weights equals to $w_p = 2^{-p-1}$ for all $p \in \{1, \dots, [n/2]\}$. Then the relations $\sum_{(p,\ell) \in M} w_{p,\ell} \leq 1$ and (3.5) is satisfied for large n and we have the following Corollary of Theorem 3.2

Corollary 4.1. *Let us assume that there exists $\xi > 0$ such that for any p , $\|\bar{\theta}_p\|_1 \leq c_p - \xi$. For n large enough, let us assume that that **(CIM)** or **(WDP)** is satisfied, that $K_n \geq k_n^*$, then there exists a $C = C(c_p, \xi, K_n, \|X\|_\infty)$ under **(WDP)** or $C = C(c_p, \xi, K_n, \theta_{\infty,n}^*)$ under **(CIM)**, such that*

$$\pi_0[R(\tilde{\theta})] \text{ and } \pi_0[R(\hat{\theta})] \leq \inf_{1 \leq p < n/2} \left\{ R(\bar{\theta}_p) + CK_n \sqrt{\frac{p}{n}} \ln(n) \right\}.$$

It is a simple consequence of Theorem 3.2 in this context so the proof is omitted.

Linear predictors are expected to be efficient when the observations are solutions of a linear autoregressive model. Let us assume that $(X_T)_{t \in \mathbb{Z}}$ is a stationary solution of an AR(∞) model

$$(4.1) \quad X_t = a_0 + \sum_{i=1}^{\infty} a_i X_{t-i} + \xi_t, \text{ for all } t \in \mathbb{Z}$$

where ξ_t are iid. Here we do not distinguish degenerate cases, i.e. $(a_i)_{i>0}$ may or may not be a sequence of infinitely many non zero numbers. So AR(p) for $p < \infty$ or AR(∞) models are considered in one shot. Assume that

$$\text{(AR)} \quad \sum_{i>0} |a_i| < 1,$$

and that ξ_1 are normally distributed as μ . Then it is easy to check that we are in the case **CIM**. Moreover, the distributions of X_{t+1} conditional on $(X_t, X_{t-1}, \dots, X_{t-p})$ are gaussian then symmetric and the median is also the mean in order that

$$\bar{\theta}_p = (a_0, a_1, a_2, \dots, a_p).$$

In this classical case, we have a corollary of Theorem 3.2

Corollary 4.2. *Let us fix $c_p = 1$ for all p , $w_p = 2^{-p-1}$ and $K_n = \ln(n)/2$. Then there exists a constant $C = C(\mu, (a_i)_{i \in \mathbb{N}})$ linear predictors $\tilde{\theta}$ and $\hat{\theta}$ satisfy, for sufficiently large n :*

$$\begin{aligned} & \pi_0[R(\tilde{\theta})] \text{ and } \pi_0[R(\hat{\theta})] \\ & \leq \mu[|\xi_0|] + \inf_{1 \leq p < n/2} \left\{ \pi_0 \left[\left| \sum_{i>p} a_i X_i \right| \right] + 2 \sqrt{\frac{\sum |a_i| + \mu[|\xi_0|]}{1 - \sum |a_i|}} \sqrt{\frac{p}{n}} \frac{\ln^2 n}{(1-p/n)} \right\} + C \frac{\ln n}{\sqrt{n}}. \end{aligned}$$

The proof is omitted as it is a simple consequence of Corollary 4.1.

Despite its apparent complexity, the procedure used here can be effectively implemented, using Monte Carlo methods, see for example Catoni [8] for an effective implementation of PAC-Bayesian methods. Actually, in this case the performance of the predictions is clearly not optimal on simulations when compared with the estimators in Ing and Wei [18] on the same set of experiments. Our procedure is clearly too conservative due to the minimax-type approach used here that focusses on pessimistic bounds based on the worst cases. The improvement of the practical performances of predictions will be the subject of future works. However, if the reader is interested, the code for the computation of the estimator is available upon request to the authors.

4.2. Neural networks predictors. In this section we consider the bounded case (**WDP**) where $\|X\|_\infty \leq 1$. The neural networks predictors proposed here are close to those in Modha and Masry [23]. The procedure approximates a natural good predictor given by $m_p(X_{n-p}, \dots, X_n)$ where

$$m_p(x) = \text{med}(X_0 | (X_{-p}, \dots, X_{-1}) = x) \text{ for all } x \in \mathbb{R}^p,$$

the median of the distribution of X_0 conditional on p past values (X_{-p}, \dots, X_{-1}) . This non-linear predictor is the optimal one with respect to the \mathbb{L}^1 -risk.

We will now present the predictors which are parametric families of functions based on the abstract neural networks used in Barron [6]. Let us assume that $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is a Lipschitz sigmoidal function such that its tail approach the tails of the unit step at least polynomially fast. More precisely, let us have the assumptions:

(NN): Assume that

- (1) $\phi(u) \rightarrow 1$ as $u \rightarrow \infty$ and $\phi(u) \rightarrow 0$ as $u \rightarrow -\infty$,
- (2) $\phi(u) - \phi(v) \leq D'_1 |u - v|$ for all $u, v \in \mathbb{R}$ and for some $D'_1 > 0$. Set $D_1 = 1 \vee D'_1$.
- (3) $|\phi(u) - \mathbb{1}_{u>0}| \leq D'_2 / |u|^{D_3}$ for $u \in \mathbb{R}$, $u \neq 0$ and for some $D_3 > 0$ and $D'_2 > 0$. Set $D_2 = 1 \vee D'_2$.

(SN): Assume that there exists a complex-valued function \tilde{m}_p on \mathbb{R}^p such that for $x \in \mathbb{R}^p$, we have

$$m_p(x) - m_p(0) = \int_{\mathbb{R}^p} (e^{iwx} - 1) \tilde{m}_p(w) dw$$

and that

$$\int_{\mathbb{R}^p} \|w\|_1 |\tilde{m}_p(w)| dw \leq C'_p < \infty$$

for some $C'_p > 0$. Set $C_p = 1 \vee C'_p$.

Then the predictors express as parametric families of functions. More precisely, they are neural networks with dimension, or "hidden units", ℓ and memory, or "time delay" or "lags", p . Let $a_i \in \mathbb{R}^p$, $b_i \in \mathbb{R}$ and $c_i \in \mathbb{R}$ for $1 \leq i \leq \ell$. Setting $\theta = (a_i, b_i, c_i, c_0)$ for some $c_0 \in \mathbb{R}$, we remark that the dimension of one predictive model is $\ell(p+2) + 1$. The predictors are defined as

$$f_\theta = \text{clip} \left(c_0 + \sum_{i=1}^{\ell} c_i \phi(a_i \cdot x + b_i) \right), \text{ for all } x \in \mathbb{R}^p,$$

where $\text{clip}(y) = y \wedge 1 \vee (-1)$. Now we restrict the parameters to be in the following ball. Define

$$\tau_\ell = 2^{(2D_3+1)/D_3} D_2^{1/D_3} \ell^{(D_3+1)/(2D_3)}$$

where D_1, D_2 and D_3 are as in Assumption (NN). Define also a compact subset

$$\mathcal{B}_{p,\ell} = \left\{ \theta; \sum_{i=1}^{\ell} |c_i| \leq C_p + \frac{1}{3}; \max_{1 \leq i \leq \ell} \|a_i\|_1 \leq \tau_\ell + \frac{1}{3\ell}; \max_{1 \leq i \leq \ell} |b_i| \leq \tau_\ell + \frac{1}{3\ell} \right\}$$

where $\|\cdot\|_1$ denotes the ℓ^1 -norm. Remark that here the constants are added in each direction to have a secure zone of width 1 in the ℓ^1 -norm around the classical parameter set:

$$\mathcal{B}'_{p,\ell} = \left\{ \theta; \sum_{i=1}^{\ell} |c_i| \leq C_p; \max_{1 \leq i \leq \ell} \|a_i\|_1 \leq \tau_\ell; \max_{1 \leq i \leq \ell} |b_i| \leq \tau_\ell \right\}$$

With the help of this secure zone, we have an Oracle inequality where the infimum is taken on the classical sets $\mathcal{B}_{p,\ell}$ for whom the optimal $\bar{\theta}_{p,\ell}$ has got the good approximation properties, see [6]. Moreover as $c_{p,\ell} - \|\bar{\theta}_{p,\ell}\|_1$ is bounded by 1 it implies that for large values of p, ℓ it holds $d_{p,\ell} \leq (\ell(p+2) + 1)(1 + \ln(C_p \vee \ell\tau_\ell + 1))$ applying Proposition 3.1. Finally, let us fix the largest possible value for ℓ as $m_p = \lfloor \sqrt{n/p} \rfloor$. It is enough for having a good approximation thanks to Theorem 3 of [6]. Then we have the following result for neural networks predictors construct on $\mathcal{B}_{p,\ell}(\xi)$:

Corollary 4.3. *Let us assume (WDP) with $\|X\|_\infty \leq 1$, $\theta_{\infty,\infty}(1) < \infty$, (NN) and (SN) with $C_p \leq C'p^c$ for some $C', c > 0$ and all $p \geq 1$. Then if we take $w_p = 1/n$ and K_n is fixed to some K , for all $\epsilon > 0$ there exists a constant $C = C(C', c, D_1, D_2, D_3, K, \epsilon)$ such that for n sufficiently large, with probability at least $1 - \epsilon$,*

$$R(\tilde{\theta}) \text{ and } R(\hat{\theta}) \leq \inf_{1 \leq p \leq \sqrt{n}/\ln(n)} \left\{ \pi_0 [|X_0 - \text{med}(X_0|X_{-1}, \dots, X_{-p})|] + C \frac{p^{1/4} \ln^{3/4} n}{n^{1/4}} \right\}.$$

The proof of this corollary is given in Subsection 5.6. Following the approach of Modha and Masry [23], our estimator is said to be a memory universal predictor with rate $\ln^{3/4}(n)/n^{1/4}$. The rate here is better than the one obtained for the \mathbb{L}^2 -risk in [23], $(\ln(n)/n)^{\frac{5}{2}}$ where $0 < c < 1$ depends on the mixing properties of the process. Remark that the choice of w_p is not optimal as it does not satisfy the relation (3.5). This implies a loss, due to that we do not manage to estimate $\sum \exp(-d_{p,\ell})$ here. However, this loss due to the weights is less than the one due to the fact that here L , the Lipschitz constant of the predictors, goes to ∞ with n . It implies the loss of a square root of the Logarithm through the "dimension" $d_{p,\ell}$, see the proof for more details. Finally, remark that the result is not easily implementable as artificial networks predictors depend on the constants C_p which are not observable.

4.3. Non-parametric auto-regressive predictors. In this section, we propose the following setting coming from economic modelization and studied in [5]. Let us assume that the process $(X_t)_{t \in \mathbb{Z}}$ is a solution of the equation:

$$X_t = f_1(X_{t-1}) + \dots + f_{p_0}(X_{t-p_0}) + \xi_t, \text{ for all } t \in \mathbb{Z}$$

where $\xi_t \sim \mathcal{N}(0, \sigma^2) =: \mu$, f_i are functions $\mathbb{R} \rightarrow \mathbb{R}$ supported by a compact set and p_0 is some unknown finite integer. Remark that, up to scale changing of X , functions f_i are supported by $[-1, 1]$. In order to be in a particular case of **(CIM)**, we assume that for any $i \in \{1, \dots, p_0\}$,

$$(4.2) \quad \exists a_i \in [0; 1[, \forall (x, x') \in [-1; 1]^2, \quad |f_i(x) - f_i(x')| \leq a_i |x - x'|$$

with $a_1 + \dots + a_{p_0} < 1$.

Actually, we assume more regularity on every f_i : they belong to the Hölder class $H(s_i, L_i)$ for $s_i \geq 1$. This means that f_i is derivable $\lfloor s_i \rfloor$ times and that

$$(4.3) \quad \exists L_i > 0, \forall (x, x') \in [-1; 1]^2, \quad |f_i^{(\lfloor s_i \rfloor)}(x) - f_i^{(\lfloor s_i \rfloor)}(x')| \leq L_i |x - x'|^{s_i - \lfloor s_i \rfloor}.$$

Remark that if (4.3) is satisfied with the relation

$$(4.4) \quad \sum_{i=1}^{p_0} \left| f_i^{(\lfloor s_i \rfloor)}(0) \right| + \dots + \left| f_i^{(\lfloor s_i \rfloor)}(0) \right| + L_i < 1$$

then (4.2) follows. Is it well known, see for example Tsybakov [26], that if $(\varphi_j(\cdot))_{j \geq 1}$ is the Fourier basis on $[-1, 1]$, namely $\phi_{2k}(x) = \sqrt{2} \cos(2\pi kx)$ and $\phi_{2k+1}(x) = \sqrt{2} \sin(2\pi kx)$, Assumption 4.3 implies that f_i belongs to a Sobolev class with regularity s_i and so that there is a constant $\gamma_i = \gamma(L_i, s_i)$ such that for any $m \in \mathbb{N} \setminus \{0\}$,

$$\min_{(\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m} \left\{ \int_{-1}^1 \left[f_i(t) - \sum_{j=1}^m \alpha_{i,j} \varphi_j(t) \right]^2 ds \right\}^{\frac{1}{2}} \leq \gamma_i m^{-s_i}.$$

Then natural predictors arise as of the form

$$\widehat{X}_{n+1} = \sum_{i=1}^p \sum_{j=1}^{\ell} \theta_{i,j} \varphi_j(X_{n-i}) =: f_{\theta}(X_n, \dots, X_{n-p})$$

for any $p \in \{1, \dots, \lfloor n/2 \rfloor\}$, any $\ell \in \{1, \dots, m_p = n\}$ and any $\theta_{p,\ell} \in \mathbb{R}$ satisfying the relation

$$\Theta_{p,\ell} = \left\{ \theta \in \mathbb{R}^{p\ell}, \sum_{i=1}^p \sum_{j=1}^{\ell} \theta_{i,j}^2 (2\lfloor j/2 \rfloor)^2 \leq L^2 \right\}.$$

This ensures that any f_{θ} is an L -Lipschitz function. Finally, let us define for any $\ell \in \{1, \dots, n\}$, $i \in \{1, \dots, \lfloor n/2 \rfloor\}$ the coefficients $\bar{\theta}_{p,\ell} \in \mathbb{R}^{p\ell}$ that satisfy the relation

$$\arg \min_{\theta \in \Theta_{p,\ell}} \pi_0 \left[\left| X_n - \sum_{i=1}^p \sum_{j=1}^{\ell} \theta_{i,j} \varphi_j(X_{n-i}) \right| \right]$$

and we obtain as a consequence of Theorem 3.2:

Corollary 4.4. *Let $w_{p,\ell} = n^{-2}$, $K_n = \ln(n)/(1+L)$ and $s = \inf\{s_1, \dots, s_{p_0}\}$. Let us assume that Equations (4.3) and (4.4) are satisfied and that there exists $c > 0$ such that for any $\ell \in \{1, \dots, n\}$ we have*

$$L - \left(\sum_{i=1}^{p_0} \sum_{j=1}^{\ell} (\bar{\theta}_{p_0,\ell})_{i,j}^2 (2\lfloor j/2 \rfloor)^2 \right)^{\frac{1}{2}} \geq c.$$

Then there is a constant

$$C = C(p_0, a, L_1, s_1, \dots, L_{p_0}, s_{p_0}, c)$$

such that

$$\pi_0[R(\hat{\theta})] \text{ and } \pi_0[R(\tilde{\theta})] \leq \mu(|\xi_0|) + Cn^{-\frac{s}{2s+1}} \ln^2(n).$$

The minimax rate of convergence with respect to s_1, \dots, s_{p_0} is achieved up to a loss in $\ln^2(n)$. Remark that the choice of the weights $w_{p,\ell}$ is not optimal as it does not satisfied condition (3.5). But it has no effect in the rate in the Oracle inequality as the loss coming from the weights is smaller than the one coming from the over-penalization. Remark also that the result in [5] achieves the minimax rate of convergence with no extra logarithmic factor. This minimax rate is achieved for the excess \mathbb{L}^2 -risk, not of prediction, but empirically on the distribution of the observed values. We argue that our risk is more natural in the time series forecasting context. However, note that in [5] it is assumed that $p_0 < p_{\max}$ for some known p_{\max} satisfying some relation with the β -mixing coefficients of the observed process. It is restrictive as that model selection procedure depends on p_{\max} and on β -mixing coefficients that are not observable.

5. PROOFS

To present the proofs in a unified version wether we work under **(CIM)** or **(WDP)**, we truncate the observations if we are under **(CIM)**. This method entirely stands in view of the result of Lemma 2.2. More precisely, we truncate the innovations ξ_t and replace them with $\bar{\xi}_t = (\xi_t \wedge C) \vee (-C)$. Now we denote $\bar{X} = (\bar{X}_t)_{t \in \mathbb{Z}}$ the solution of the equation

$$\bar{X}_t = F(\bar{X}_{t-1}, \bar{X}_{t-2}, \bar{X}_{t-3}, \dots; \bar{\xi}_t), \text{ a.e. for all } t \in \mathbb{Z}.$$

This solution exists and satisfies weak dependence conditions, see Lemma 2.2 for more details. To treat both cases in the same way, we denote in the sequel $\bar{X} := X$ and $\|X\|_{\infty} = C$ under **(WDP)**. Moreover, we will use the notation \bar{r} , \bar{R} the risks associated with \bar{X} .

We will now present some useful Lemmas. Their proofs are postponed at the end of the Section.

5.1. Useful Lemmas. The first Lemma gives a bound on the deviations of the risk of \bar{X} . The result derives simply from the Rio's "Hoeffding's type" inequality stated in [24].

Lemma 5.1. *For any $\lambda > 0$ and $\theta \in \Theta$ we have:*

$$\pi_0[\exp(\lambda(\bar{R}(\theta) - \bar{r}_n(\theta)))] \leq \exp\left(\frac{\lambda^2 k_n^2}{n(1-p(\theta)/n)^2}\right),$$

where k_n depends on the nature of the observations, more precisely is given by the relations

$$\left\{ \begin{array}{l} \text{(CIM)} \quad k_n = \frac{uC \left(1 + 2 \sum_{r=1}^n \inf_{0 < k < r} \left\{ a(F)^{r/k} + \sum_{j=k}^{\infty} a_j(F) \right\}\right)}{(1+L)(1-a(F))}, \\ \text{(WDP)} \quad k_n = \frac{\|X\|_{\infty} + 2\theta_{\infty,n}(1)}{1+L}. \end{array} \right.$$

The proof of this Lemma is given in Section 5.4.

We now give a result particularly useful for the so-called "PAC-Bayesian" randomization technique proposed by Catoni [7, 8]. Given a measurable space (E, \mathcal{E}) we let $\mathcal{M}_+^1(E)$ denote the set of all probability measures on (E, \mathcal{E}) . The Kullback divergence is a pseudo-distance on $\mathcal{M}_+^1(E)$ defined, for any $(\pi, \pi') \in [\mathcal{M}_+^1(E)]^2$ by the equation

$$\mathcal{K}(\pi, \pi') = \begin{cases} \pi[\ln(d\pi/d\pi')] & \text{if } \pi \ll \pi', \\ +\infty & \text{otherwise.} \end{cases}$$

Lemma 5.2 (Legendre transform of the Kullback divergence function). *For any $\pi \in \mathcal{M}_+^1(E)$, for any measurable function $h : E \rightarrow \mathbb{R}$ such that $\pi[\exp(h)] < +\infty$ we have:*

$$(5.1) \quad \pi[\exp(h)] = \exp \left(\sup_{\rho \in \mathcal{M}_+^1(E)} \left(\rho[h] - \mathcal{K}(\rho, \pi) \right) \right),$$

with convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of π , the supremum with respect to ρ in the right-hand side is reached for the Gibbs measure $\pi\{h\}$ defined in (2.3).

The proof of this Lemma is omitted here as it can be found in [7] or [8].

With the help of Lemma 5.2, we then can prove a general PAC-Bayesian bound from Lemma 5.1. It consists in an upper-bound for the mean risk of Gibbs estimators in all sub-models.

Lemma 5.3. *Under the assumptions of Theorem 3.2 we have for any $\lambda > 0$ and $(p, \ell) \in M$:*

$$(5.2) \quad \pi_0 \left[\exp \left(\sup_{\rho \in \mathcal{M}_+^1(\Theta_{p,\ell})} \left\{ \lambda \rho[\bar{R} - \bar{r}_n] - \mathcal{K}(\rho, \pi_{p,\ell}) \right\} - \frac{\lambda^2 k_n^2}{n(1-p/n)^2} \right) \right] \leq 1,$$

where k_n has the same expression than in Lemma 5.1.

The proof of this Lemma is given in Section 5.4.

From this result, we derive another PAC-Bayesian bound on the mean risk of any aggregation estimators of all Gibbs estimators. The techniques were developed by Catoni [8, 7] in the iid or exchangeable setting for classification on the basis of the seminal paper of McAllester [21] and extended by Audibert [4] to regression with quadratic loss and Alquier [2] to a general loss function. The scheme use here follows [7].

Lemma 5.4. *For any measurable function $\rho_{p,\ell} : \mathcal{X}^n \rightarrow \mathcal{M}_+^1(\Theta_{p,\ell})$ for $(p, \ell) \in M$ and for any measurable family of weights $\hat{w}_{p,\ell}^\lambda : \mathcal{X}^n \rightarrow [0, 1]$ with*

$$\sum_{(p,\ell) \in M} \hat{w}_{p,\ell}^\lambda \leq 1,$$

under the assumptions of Theorem 3.2 we have:

$$\pi_0 \left[\sup_{\substack{\lambda \in \mathcal{G} \\ \sum_{(p,\ell) \in M} \hat{w}_{p,\ell}^\lambda = 1 \\ \rho_{p,\ell} \in \mathcal{M}_+^1(\Theta_{p,\ell})}} \left\{ \sum_{(p,\ell) \in M} \hat{w}_{p,\ell}^\lambda \rho_{p,\ell} \left[\exp \left(\lambda(\bar{R} - \bar{r}_n) - \ln \frac{d\rho_{p,\ell}}{d\pi_{p,\ell}} - \frac{\lambda^2 k_n^2}{n(1-p/n)^2} + \ln \frac{w_{p,\ell}}{|\mathcal{G}| \hat{w}_{p,\ell}^\lambda} \right) \right] \right\} \right]$$

$$\forall \pi_0 \left[\sup_{\substack{\lambda \in \mathcal{G} \\ (p, \ell) \in M \\ \rho_{p, \ell} \in \mathcal{M}_+^1(\Theta_{p, \ell})}} \left\{ \exp \left(\lambda \rho_{p, \ell} [\bar{R} - \bar{r}_n] - \mathcal{K}(\rho_{p, \ell}, \pi_{p, \ell}) - \frac{\lambda^2 k_n^2}{n(1-p/n)^2} + \ln \frac{w_{p, \ell}}{|\mathcal{G}|} \right) \right\} \right] \leq 1,$$

where k_n has the same expression than in Lemma 5.1.

The proof of this Lemma is given in Section 5.4.

Finally, we present a Lemma that gives a usefull inequality under **(CIM)**. We recall that Ψ denotes here the Laplace transform of the norm of $\|\xi_0\|$, that is assumed to be finite.

Lemma 5.5. *Let us define the following random variable:*

$$g(C) = \sup_{\theta \in \Theta} |r_n(\theta) - \bar{r}_n(\theta)|.$$

We have under **(CIM)** the following inequality, for any $c > 0$:

$$\pi_0[g(C)] \leq \frac{1+L}{1-a(F)} u \Psi(c) C \exp(-cC).$$

The proof of this Lemma is given in Section 5.4.

5.2. Proof of Theorem 3.2. We are now able to give the proof of our main theorem. Let us apply Lemma 5.4. As it holds for any probability measure $\rho_{p, \ell}$ it holds for $\hat{\rho}_{p, \ell} = \pi_{p, \ell}\{-\lambda r_n\}$ associated to any $\Lambda = \mathcal{G}$. We use the inequality $\forall x \in \mathbb{R}, \exp(x) \geq \mathbf{1}_{\mathbb{R}_+^*}(x)$ and the associated Markov inequality:

$$\pi_0 \hat{\rho}_{p, \ell}(A + \ln(\varepsilon) > 0) \leq \pi_0 \hat{\rho}_{p, \ell}[\exp(A)] \varepsilon \leq \varepsilon$$

where $A = \lambda(\bar{R} - \bar{r}_n) - \ln \frac{d\rho_{p, \ell}}{d\pi_{p, \ell}} - \frac{\lambda^2 k_n^2}{n(1-p/n)^2} + \ln \frac{w_{p, \ell}}{|\mathcal{G}| \hat{w}_{p, \ell}^\lambda}$ for any $\lambda \in \mathcal{G}$. Here we used the fact that $\hat{\rho}_{p, \ell}$ is a probability conditional on (X_1, \dots, X_n) in order that $\pi_0 \hat{\rho}_{p, \ell}$ is a well defined probability measure. Moreover, we have used elementary convex inequality to get rid off with the sum of the weights $\hat{w}_{p, \ell}^\lambda$ as they are fixed. With probability $1 - \varepsilon$ on the drawing of the data with respect to π_0 and on the drawing of all the estimators $\hat{\theta}_{p, \ell}^\lambda$ with respect to $\hat{\rho}_{p, \ell}^\lambda$ and on the drawing of p and ℓ according to $\hat{w}_{p, \ell}^\lambda$, we have, for any $(p, \ell) \in M$:

$$(5.3) \quad \bar{R}(\tilde{\theta}_{p, \ell}^\lambda) \leq \bar{r}_n(\tilde{\theta}_{p, \ell}^\lambda) + \frac{\lambda k_n^2}{n(1-p/n)^2} + \frac{1}{\lambda} \ln \left[\frac{d\rho_{p, \ell}}{d\pi_{p, \ell}}(\tilde{\theta}_{p, \ell}^\lambda) \right] + \frac{1}{\lambda} \ln \frac{|\mathcal{G}| \hat{w}_{p, \ell}^\lambda}{w_{p, \ell}} + \frac{1}{\lambda} \ln \frac{1}{\varepsilon}.$$

Using the same technique but with the second part of the result of Lemma 5.4 we obtain for any $(p, \ell) \in M$, $\lambda \in \mathcal{G}$ and $\rho \in \mathcal{M}_+^1(\Theta_{p, \ell})$,

$$(5.4) \quad \int_{\Theta_{p, \ell}} \bar{r}_n(\theta) \rho(d\theta) \leq \int_{\Theta_{p, \ell}} \bar{R}(\theta) \rho(d\theta) + \frac{\lambda k_n^2}{n(1-p/n)^2} + \frac{1}{\lambda} \mathcal{K}(\rho, \pi_{p, \ell}) + \frac{1}{\lambda} \ln \frac{|\mathcal{G}|}{w_{p, \ell}} + \frac{1}{\lambda} \ln \frac{1}{\varepsilon}.$$

Note that (5.3) is equivalent to

$$(5.5) \quad \begin{aligned} \bar{R}(\tilde{\theta}_{p, \ell}^\lambda) &\leq -\frac{1}{\lambda} \ln \int_{\Theta_{p, \ell}} \exp(-\lambda r_n(\theta)) \pi_{p, \ell}(d\theta) + \frac{\lambda k_n^2}{n(1-p/n)^2} + \frac{1}{\lambda} \ln \frac{|\mathcal{G}| \hat{w}_{p, \ell}^\lambda}{w_{p, \ell}} + \frac{1}{\lambda} \ln \frac{1}{\varepsilon} \\ &\quad + \bar{r}_n(\tilde{\theta}_{p, \ell}^\lambda) - r_n(\tilde{\theta}_{p, \ell}^\lambda) \\ &\leq \hat{R}(p, \ell, \lambda) + \frac{1}{\lambda} \ln \frac{\hat{w}_{p, \ell}^\lambda}{\varepsilon} + \frac{\lambda(k_n^2 - K_n^2)}{n(1-p/n)^2} + \left| \bar{r}_n(\tilde{\theta}_{p, \ell}^\lambda) - r_n(\tilde{\theta}_{p, \ell}^\lambda) \right| \end{aligned}$$

so we obtain

$$(5.6) \quad \bar{R}(\tilde{\theta}_{p,\ell}^\lambda) \leq \hat{R}(p, \ell, \lambda) + \frac{1}{\lambda} \ln \frac{\hat{w}_{p,\ell}^\lambda}{\varepsilon} + \frac{\lambda(k_n^2 - K_n^2)}{n(1-p/n)^2} + g(C) \mathbb{1}_{(\mathbf{CIM})}.$$

First let us study the estimator $\hat{\theta}$. For any $\lambda > 0$, let us choose $\hat{w}_{p,\ell}^\lambda = 1$ when (p, ℓ) minimizes $\hat{R}(p, \ell, \lambda)$ and 0 otherwise. Remembering that $(\hat{p}, \hat{\ell}, \hat{\lambda}) = \arg \min \hat{R}(p, \ell, \lambda)$, we obtain:

$$(5.7) \quad \bar{R}(\hat{\theta}) \leq \inf_{p,\ell,\lambda} \hat{R}(p, \ell, \lambda) + \frac{\hat{\lambda}(k_n^2 - K_n^2)}{n(1-\hat{p}/n)^2} - \frac{1}{\hat{\lambda}} \ln \varepsilon + g(C) \mathbb{1}_{(\mathbf{CIM})}.$$

Now, we are going to upper bound the term $\hat{R}(p, \ell, \lambda)$. From inequality (5.4) we derive that

$$\begin{aligned} -\frac{1}{\lambda} \ln \int_{\Theta_{p,\ell}} \exp(-\lambda r_n(\theta)) \pi_{p,\ell}(d\theta) &= \inf_{\rho \in \mathcal{M}_+^1(\Theta_{p,\ell})} \left\{ \int_{\Theta_{p,\ell}} r_n(\theta) \rho(d\theta) + \frac{1}{\lambda} \mathcal{K}(\rho, \pi_{p,\ell}) \right\} \\ &\leq \inf_{\rho \in \mathcal{M}_+^1(\Theta_{p,\ell})} \left\{ \int_{\Theta_{p,\ell}} \bar{r}_n(\theta) \rho(d\theta) + \frac{1}{\lambda} \mathcal{K}(\rho, \pi_{p,\ell}) \right\} + g(C) \mathbb{1}_{(\mathbf{CIM})} \\ &\leq \inf_{\rho \in \mathcal{M}_+^1(\Theta_{p,\ell})} \left\{ \int_{\Theta_{p,\ell}} \bar{R}(\theta) \rho(d\theta) + \frac{2}{\lambda} \mathcal{K}(\rho, \pi_{p,\ell}) \right\} + \frac{\lambda k_n^2}{n(1-p/n)^2} + \frac{1}{\lambda} \ln \frac{|\mathcal{G}|}{\varepsilon w_{p,\ell}} + g(C) \mathbb{1}_{(\mathbf{CIM})} \\ &= -\frac{2}{\lambda} \ln \int_{\Theta_{p,\ell}} \exp\left(-\frac{\lambda}{2} \bar{R}(\theta)\right) \pi_{p,\ell}(d\theta) + \frac{\lambda k_n^2}{n(1-p/n)^2} + \frac{1}{\lambda} \ln \frac{|\mathcal{G}|}{\varepsilon w_{p,\ell}} + g(C) \mathbb{1}_{(\mathbf{CIM})}. \end{aligned}$$

So we obtain:

$$(5.8) \quad \hat{R}(p, \ell, \lambda) \leq -\frac{2}{\lambda} \ln \int_{\Theta_{p,\ell}} \exp\left(-\frac{\lambda}{2} \bar{R}(\theta)\right) \pi_{p,\ell}(d\theta) + \frac{\lambda(k_n^2 + K_n^2)}{n(1-p/n)^2} + \frac{1}{\lambda} \ln \frac{|\mathcal{G}|}{\varepsilon w_{p,\ell}} + g(C) \mathbb{1}_{(\mathbf{CIM})}.$$

Now, let us remark that, as soon as $\lambda > 2e$, we have that

$$-\ln \pi_{p,\ell} \left[\exp\left(-\frac{\lambda}{2} (R - R(\bar{\theta}_{p,\ell}))\right) \right] \leq d_{p,\ell} \ln \frac{\lambda}{2}$$

as we work under Assumption (3.1) and it easily follows that

$$\begin{aligned} -\ln \pi_{p,\ell} \left[\exp\left(-\frac{\lambda}{2} \bar{R}\right) \right] &\leq -\ln \pi_{p,\ell} \left[\exp\left(-\frac{\lambda}{2} R\right) \right] + \frac{\lambda}{2} \pi_0[g(C)] \mathbb{1}_{(\mathbf{CIM})} \\ &= -\ln \pi_{p,\ell} \left[\exp\left(-\frac{\lambda}{2} (R - R(\bar{\theta}_{p,\ell}))\right) \right] + \frac{\lambda}{2} R(\bar{\theta}_{p,\ell}) + \frac{\lambda}{2} \pi_0[g(C)] \mathbb{1}_{(\mathbf{CIM})} \\ &\leq d_{p,\ell} \ln \frac{\lambda}{2} + \frac{\lambda}{2} R(\bar{\theta}_{p,\ell}) + \frac{\lambda}{2} \pi_0[g(C)] \mathbb{1}_{(\mathbf{CIM})}. \end{aligned}$$

We plug this result into the inequality (5.8) to obtain:

$$(5.9) \quad \hat{R}(p, \ell, \lambda) \leq R(\bar{\theta}_{p,\ell}) + \frac{1}{\lambda} \left(d_{p,\ell} \ln^2 \frac{\lambda}{2} + \ln \frac{|\mathcal{G}|}{\varepsilon w_{p,\ell}} \right) + \frac{\lambda(k_n^2 + K_n^2)}{n(1-p/n)^2} + (g(C) + \pi_0[g(C)]) \mathbb{1}_{(\mathbf{CIM})}.$$

Now we can conclude to the result of Theorem 3.2. We work under **(WDP)** so that $(g(C) + \pi_0[g(C)])\mathbb{1}_{(\mathbf{CIM})} = 0$ and that $\bar{R} = R$. It remains to collect the informations of the inequalities (5.7) and (5.9). We obtain:

$$(5.10) \quad R(\hat{\theta}) \leq \inf_{p,\ell,\lambda} \left\{ \hat{R}(p, \ell, \lambda) \right\} + \frac{\hat{\lambda}(k_n^2 - K_n^2)}{n(1 - \hat{p}/n)} - \frac{1}{\hat{\lambda}} \ln \varepsilon$$

$$\leq \inf_{p,\ell,\lambda} \left\{ R(\bar{\theta}_{p,\ell}) + \frac{1}{\lambda} \left(d_{p,\ell} \ln^2 \frac{\lambda}{2} + \ln \frac{|\mathcal{G}|}{\varepsilon w_{p,\ell}} \right) + \frac{\lambda(k_n^2 + K_n^2)}{n(1 - p/n)^2} \right\}$$

$$+ \frac{\hat{\lambda}(k_n^2 - K_n^2)}{n(1 - \hat{p}/n)^2} - \frac{1}{\hat{\lambda}} \ln \varepsilon$$

So it remains to get rid of the two last terms. First of all, let us control $(1/\hat{\lambda}) \ln(1/\varepsilon)$. Remember that $\hat{\lambda}$ is a the minimizer of

$$\hat{R}(\hat{p}, \hat{\ell}, \lambda) = -\frac{1}{\lambda} \ln \int_{\Theta_{\hat{p}, \hat{\ell}}} \exp(-\lambda r_n(\theta)) d\pi_{\hat{p}, \hat{\ell}}(\theta) + \frac{1}{\lambda} \ln \frac{|\mathcal{G}|}{w_{\hat{p}, \hat{\ell}}} + \frac{\lambda K_n^2}{n \left(1 - \frac{\hat{p}}{n}\right)^2}$$

$$= G(\lambda) + \frac{1}{\lambda} \ln |\mathcal{G}| + \frac{\lambda K_n^2}{n \left(1 - \frac{\hat{p}}{n}\right)^2}$$

where G is a decreasing function as

$$G'(\lambda) = \frac{1}{\lambda^2} \ln \int_{\Theta_{\hat{p}, \hat{\ell}}} \exp(-\lambda r_n(\theta)) d\pi_{\hat{p}, \hat{\ell}}(\theta) - \frac{1}{\lambda} \int_{\Theta_{\hat{p}, \hat{\ell}}} r_n(\theta) d\pi_{\hat{p}, \hat{\ell}}\{-\lambda r_n\}(\theta) - \frac{1}{\lambda} \ln \frac{1}{w_{p,\ell}}$$

and we can check that each of these three term is negative. So this means that $\hat{\lambda} > \check{\lambda}$ where $\check{\lambda}$ is the minimizer of

$$\frac{1}{\lambda} \ln |\mathcal{G}| + \frac{\lambda K_n^2}{n \left(1 - \frac{\hat{p}}{n}\right)^2};$$

it appears that $\check{\lambda}$ is known in explicit form and so we obtain

$$\frac{1}{\check{\lambda}} \ln \frac{1}{\varepsilon} \leq \frac{1}{\check{\lambda}} \ln \frac{1}{\varepsilon} = \frac{K_n}{\left(1 - \frac{\hat{p}}{n}\right) \sqrt{n \ln |\mathcal{G}|}} \ln \frac{1}{\varepsilon} \leq \frac{2K_n}{\sqrt{n}} \ln \frac{1}{\varepsilon}.$$

So, Inequality 5.10 becomes

$$(5.11) \quad R(\hat{\theta}) \leq \inf_{p,\ell,\lambda} \left\{ R(\bar{\theta}_{p,\ell}) + \frac{1}{\lambda} \left(d_{p,\ell} \ln^2 \frac{\lambda}{2} + \ln \frac{|\mathcal{G}|}{\varepsilon w_{p,\ell}} \right) + \frac{\lambda(k_n^2 + K_n^2)}{n(1 - p/n)^2} \right\}$$

$$+ \frac{\hat{\lambda}(k_n^2 - K_n^2)}{n(1 - \hat{p}/n)^2} + \frac{2K_n}{\sqrt{n}} \ln \frac{1}{\varepsilon}.$$

Let us now consider two cases: $k_n \leq K_n$ and $k_n > K_n$. If $k_n \leq K_n$, Inequality 5.11 becomes

$$(5.12) \quad R(\hat{\theta}) \leq \inf_{p,\ell,\lambda} \left\{ R(\bar{\theta}_{p,\ell}) + \frac{1}{\lambda} \left(d_{p,\ell} \ln^2 \frac{\lambda}{2} + \ln \frac{|\mathcal{G}|}{\varepsilon w_{p,\ell}} \right) + \frac{\lambda(k_n^2 + K_n^2)}{n(1 - p/n)^2} \right\} + \frac{2K_n}{\sqrt{n}} \ln \frac{1}{\varepsilon}.$$

and let us replace the infimum with respect to λ by the specific value

$$\lambda^*(p, \ell) = \frac{1 - p/n}{K_n} \sqrt{d_{p,\ell} n} \ln(d_{p,\ell} n)$$

that well balances the second and the third term of the sum in (5.9). Now if $n^2 \geq \lambda^*(p, \ell) \geq 4e$ then we can find a $\lambda' \in \mathcal{G}$ such that $\lambda' > 2e$ and $\lambda' \leq \lambda^*(p, \ell) \leq 2\lambda'$. This holds if, for example, $p \leq n/2$, $n \ln^2 n \geq (8eK_n)^2$ and $d_{p,\ell} \leq nK_n$. This leads to the following inequality

$$(5.13) \quad R(\hat{\theta}) \leq \inf_{d_{p,\ell} \leq nK_n} \left\{ R(\bar{\theta}_{p,\ell}) + \frac{k_n^2/K_n + 2K_n}{1 - p/n} \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell} n) + \frac{K_n \ln \frac{|\mathcal{G}|}{\varepsilon w_{p,\ell}}}{(1 - p/n) \sqrt{d_{p,\ell} n} \ln(d_{p,\ell} n)} \right\} + \frac{2K_n}{\sqrt{n}} \ln \frac{1}{\varepsilon}.$$

Now, let us consider the case where $k_n > K_n$ (this is the difficult case). We have

$$\begin{aligned} \frac{\hat{\lambda}(k_n^2 - K_n^2)}{n(1 - \hat{p}/n)^2} &= \frac{k_n^2 - K_n^2}{k_n^2 + K_n^2} \frac{\hat{\lambda}(k_n^2 + K_n^2)}{n(1 - \hat{p}/n)^2} \\ &\leq \frac{k_n^2 - K_n^2}{k_n^2 + K_n^2} \left[-\frac{1}{\lambda} \ln \int_{\Theta_{\hat{p}, \hat{\ell}}} \exp(-\lambda r_n(\theta)) d\pi_{\hat{p}, \hat{\ell}}(\theta) + \frac{1}{\lambda} \ln \frac{|\mathcal{G}|}{w_{\hat{p}, \hat{\ell}}} + \frac{\lambda(k_n^2 + K_n^2)}{n(1 - \frac{\hat{p}}{n})^2} \right] \\ &= \frac{k_n^2 - K_n^2}{k_n^2 + K_n^2} \inf_{p, \ell, \lambda} \hat{R}(p, \ell, \lambda) \end{aligned}$$

by definition of $(\hat{p}, \hat{\ell}, \hat{\lambda})$ and so, using Inequality 5.10, we obtain

$$(5.14) \quad R(\hat{\theta}) \leq \left(1 + \frac{k_n^2 - K_n^2}{k_n^2 + K_n^2} \right) \inf_{p, \ell, \lambda} \left\{ R(\bar{\theta}_{p,\ell}) + \frac{1}{\lambda} \left(d_{p,\ell} \ln^2 \frac{\lambda}{2} + \ln \frac{|\mathcal{G}|}{\varepsilon w_{p,\ell}} \right) + \frac{\lambda(k_n^2 + K_n^2)}{n(1 - p/n)^2} \right\} + \frac{2K_n}{\sqrt{n}} \ln \frac{1}{\varepsilon}.$$

The same particular value of λ leads to

$$(5.15) \quad R(\hat{\theta}) \leq \left(\frac{k_n^2 + K_n^2}{k_n^2 + K_n^2} + \frac{k_n^2 - K_n^2}{k_n^2 + K_n^2} \right) \inf_{d_{p,\ell} \leq nK_n} \left\{ R(\bar{\theta}_{p,\ell}) + \frac{k_n^2/K_n + 2K_n}{1 - p/n} \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell} n) + \frac{K_n \ln \frac{|\mathcal{G}|}{\varepsilon w_{p,\ell}}}{(1 - p/n) \sqrt{d_{p,\ell} n} \ln(d_{p,\ell} n)} \right\} + \frac{2K_n}{\sqrt{n}} \ln \frac{1}{\varepsilon}.$$

If we combine 5.13 and 5.15 on both cases, and if we remark that $|\mathcal{G}| \leq \log_2(n^2) \leq 3 \ln(n)$, we obtain

$$R(\hat{\theta}) \leq \left(1 \vee \frac{2k_n^2}{k_n^2 + K_n^2} \right) \inf_{d_{p,\ell} \leq nK_n} \left\{ R(\bar{\theta}_{p,\ell}) + K_n \left[\frac{2 + (k_n/K_n)^2}{1 - p/n} \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell} n) \right. \right.$$

$$\left. + \frac{4 \ln \frac{3 \ln n}{\varepsilon w_{p,\ell}}}{\sqrt{d_{p,\ell} n \ln(d_{p,\ell} n)}} \right\}$$

that ends the proof.

5.3. Proof of Theorem 3.3. Here we deal with both cases **(WDP)** and **(CIM)** at the same time. We also use the results given in the Proof of the Theorem 3.2. More precisely, it has been shown that for any (p, ℓ) such that $d_{p,\ell} \leq nK_n$ the following inequality can not hold with probability larger than ε :

$$\begin{aligned} \bar{R}(\hat{\theta}) \text{ and } \bar{R}(\tilde{\theta}) \geq & \left(1 \vee \frac{2k_n^2}{k_n^2 + K_n^2} \right) \left\{ R(\bar{\theta}_{p,\ell}) + K_n \left[\frac{2 + (k_n/K_n)^2}{1 - p/n} \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell} n) \right. \right. \\ & \left. \left. + \frac{4 \ln \frac{3 \ln n}{\varepsilon w_{p,\ell}}}{\sqrt{d_{p,\ell} n \ln(d_{p,\ell} n)}} \right] \right\} + (g(C) + \pi_0[g(C)]) \mathbb{1}_{(\mathbf{CIM})} \end{aligned}$$

and so for

$$\begin{aligned} R(\hat{\theta}) \text{ and } R(\tilde{\theta}) \geq & \left(1 \vee \frac{2k_n^2}{k_n^2 + K_n^2} \right) \left\{ R(\bar{\theta}_{p,\ell}) + K_n \left[\frac{2 + (k_n/K_n)^2}{1 - p/n} \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell} n) \right. \right. \\ & \left. \left. + \frac{4 \ln \frac{3 \ln n}{\varepsilon w_{p,\ell}}}{\sqrt{d_{p,\ell} n \ln(d_{p,\ell} n)}} \right] \right\} + (g(C) + 2\pi_0[g(C)]) \mathbb{1}_{(\mathbf{CIM})} = a_n(p, \ell) + b_n(p, \ell) \log \frac{1}{\varepsilon}. \end{aligned}$$

Let us now deal with $\hat{\theta}$ (the proof for $\tilde{\theta}$ is similar). For any (p, ℓ) with $d_{p,\ell} \leq nK_n$,

$$\pi_0 \left[e^{\frac{R(\hat{\theta}) - a_n(p,\ell)}{2b_n(p,\ell)}} \geq \varepsilon^{-\frac{1}{2}} \right] \leq \varepsilon,$$

this leads to

$$\pi_0 \left[e^{\frac{R(\hat{\theta}) - a_n(p,\ell)}{2b_n(p,\ell)}} \right] = \int_0^\infty \pi_0 \left[e^{\frac{R(\hat{\theta}) - a_n(p,\ell)}{2b_n(p,\ell)}} \geq t \right] dt \leq \int_0^\infty \left(1 \wedge \frac{1}{t^2} \right) dt = 2$$

and so

$$\pi_0 \left[R(\hat{\theta}) - a_n(p, \ell) \right] \leq 2b_n(p, \ell) \ln 2.$$

Replacing $a_n(p, \ell)$ and $b_n(p, \ell)$ by their definitions we obtain

$$\begin{aligned} \pi_0 \left[R(\hat{\theta}) \right] \leq & \left(1 \vee \frac{2k_n^2}{k_n^2 + K_n^2} \right) \left\{ R(\bar{\theta}_{p,\ell}) + K_n \left[\frac{2 + (k_n/K_n)^2}{1 - p/n} \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell} n) \right. \right. \\ & \left. \left. + \frac{4 \ln \frac{12 \ln n}{w_{p,\ell}}}{\sqrt{d_{p,\ell} n \ln(d_{p,\ell} n)}} \right] \right\} + 3\pi_0[g(C)] \mathbb{1}_{(\mathbf{CIM})} \end{aligned}$$

Under **(WDP)** we then get the desired result. Under **(CIM)**, we use the result of Lemma 5.5 to choose C in order to well balance $k_n(C)$ given in Lemma 5.1 and $g(C)$. We fix it equal to

$$C^* = \frac{\ln n}{2c^*} \text{ and } c^* = \frac{u \left(1 + 2 \sum_{r=1}^n \inf_{0 < k < r} \left\{ a(F)^{r/k} + \sum_{j=k}^\infty a_j(F) \right\} \right)}{2(1 - a(F))}.$$

Remark that this choice is independent of p, ℓ . This ends the proof for $\hat{\theta}$ and the same results hold also for $\tilde{\theta}$.

5.4. Proofs of Lemmas 5.1, 5.3, 5.4, 5.5 and of Proposition 3.1.

Proof of Lemma 5.1. The proof of this Lemma is based on the application of a useful inequality from Rio [24] on \overline{X} . Let us first recall this result:

Theorem 5.6. *Let $Y = (Y_t)_{t \in \mathbb{Z}}$ be a stationary time series bounded by C distributed as π_0 on $\mathcal{X}^{\mathbb{Z}}$. Let h be a 1-Lipschitz function of $\mathcal{X}^n \rightarrow \mathbb{R}$, i.e. such that:*

$$(5.16) \quad \forall (x_1, y_1, \dots, x_n, y_n) \in \mathcal{X}^{2n}, \quad \left| h(x_1, \dots, x_n) - h(y_1, \dots, y_n) \right| \leq \sum_{i=1}^n \|x_i - y_i\|.$$

Then for every $t \in \mathbb{R}$ we have:

$$\pi_0 [\exp(t(\pi_0[h(X_1, \dots, X_n)] - h(X_1, \dots, X_n)))] \leq \exp\left(\frac{t^2}{8}n(C + 2, \theta_{\infty, n}(1))^2\right).$$

Proof of Theorem 5.6. We achieve this version of Theorem 1 of [24] remarking that we can rewrite the inequality (3) in [24] as, for any 1-Lipschitz function g :

$$\Gamma(g) = \|\mathbb{E}(g(X_{\ell+1}, \dots, X_n) | \mathcal{F}_\ell) - \mathbb{E}(g(X_{\ell+1}, \dots, X_n))\|_\infty \leq \theta_{\infty, n-\ell}(1).$$

It leads to the result of Lemma 5.6 when bounding $\sum_{r=1}^n (C + \theta_{\infty, r}(1))^2$ with $n(C + \theta_{\infty, n}(1))^2$ as $\sup_{1 \leq r \leq n} \theta_{\infty, r}(1) \leq \theta_{\infty, n}(1)$. \square

We now apply the result of Theorem 5.6 on $Y = \overline{X}$ to obtain the result of Lemma 5.1. Let us fix $\lambda > 0$, $(p, \ell) \in M$, $\theta \in \Theta_{p, \ell}$ and $t = (1 + L)\lambda / [n - p(\theta)]$ and the function h defined by:

$$h(x_1, \dots, x_n) = \frac{1}{1 + L} \sum_{i=p(\theta)+1}^n \|x_i - f_\theta(x_{i-1}, \dots, x_{i-p(\theta)})\|.$$

We easily check that h satisfies condition 5.16 in order to apply Rio's inequality. Note that:

$$\begin{aligned} & \left| h(x_1, \dots, x_n) - h(y_1, \dots, y_n) \right| \\ & \leq \frac{1}{1 + L} \sum_{i=p(\theta)+1}^n \left| \|x_i - f_\theta(x_{i-1}, \dots, x_{i-p(\theta)})\| - \|y_i - f_\theta(y_{i-1}, \dots, y_{i-p(\theta)})\| \right| \\ & \leq \frac{1}{1 + L} \sum_{i=p(\theta)+1}^n \|x_i - y_i - f_\theta(x_{i-1}, \dots, x_{i-p(\theta)}) + f_\theta(y_{i-1}, \dots, y_{i-p(\theta)})\| \\ & \leq \frac{1}{1 + L} \sum_{i=p(\theta)+1}^n \|x_i - y_i\| + \frac{1}{1 + L} \sum_{i=p(\theta)+1}^n \|f_\theta(x_{i-1}, \dots, x_{i-p(\theta)}) - f_\theta(y_{i-1}, \dots, y_{i-p(\theta)})\| \\ & \leq \frac{1}{1 + L} \sum_{i=p(\theta)+1}^n \|x_i - y_i\| + \frac{L}{1 + L} \sum_{i=p(\theta)+1}^n \sum_{j=1}^{p(\theta)} a_j(\theta) \|x_{i-j} - y_{i-j}\| \\ & \leq \frac{1}{1 + L} \sum_{i=p(\theta)+1}^n \|x_i - y_i\| + \frac{L}{1 + L} \sum_{i=1}^n \|x_i - y_i\| \end{aligned}$$

$$\leq \sum_{i=1}^n \|x_i - y_i\|.$$

The direct application of Theorem 5.6 ends the proof under **(WDP)**. Under **(CIM)** k_n is computed in view of the estimate of $\theta_{\infty,n}(1)$ obtained in Lemma 2.2. \square

Proof of Lemma 5.3. Integrate the inequality in Lemma 5.1 with respect $\pi_{p,\ell}$ on $\Theta_{p,\ell}$ (then $p(\theta) = p$) for any $(p, \ell) \in M$ in order to obtain:

$$\pi_{p,\ell}[\pi_0[\exp(\lambda(\bar{R} - \bar{r}_n))]] \leq \exp\left(\frac{\lambda^2 k_n^2}{n(1-p/n)^2}\right).$$

Fubini's Theorem implies that

$$\pi_0 \left[\pi_{p,\ell} \left[\exp\left(\lambda(\bar{R} - \bar{r}_n) - \frac{\lambda^2 k_n^2}{n(1-p/n)^2}\right) \right] \right] \leq 1.$$

Applying Lemma 5.2 for $\pi = \pi_{p,\ell}$ and $h = \lambda(\bar{R} - \bar{r}_n) - \lambda^2 k_n^2 / (n(1-p/n)^2)$ on $\mathcal{M}_+^1(\Theta_{p,\ell})$ leads to the inequality:

$$\pi_0 \left[\exp\left(\sup_{\rho \in \mathcal{M}_+^1(\Theta_{p,\ell})} \{\lambda\rho[\bar{R} - \bar{r}_n] - \mathcal{K}(\rho, \pi_{p,\ell})\} - \frac{\lambda^2 k_n^2}{n(1-p/n)^2}\right) \right] \leq 1.$$

This ends the proof. \square

Proof of Lemma 5.4. First, let us choose $\lambda \in \Lambda$. Let $h_{p,\ell}^\lambda$ denotes, for any $(p, \ell) \in M$:

$$h_{p,\ell}^\lambda = \sup_{\rho_{p,\ell} \in \mathcal{M}_+^1(\Theta_{p,\ell})} \left\{ \lambda\rho_{p,\ell}[\bar{R} - \bar{r}_n] - \mathcal{K}(\rho_{p,\ell}, \pi_{p,\ell}) \right\} - \frac{\lambda^2 k_n^2}{n(1-p/n)^2}.$$

From Lemma 5.3 applied on the different $\mathcal{M}_+^1(\Theta_{p,\ell})$ we have, for any $(p, \ell) \in M$:

$$\pi_0 \left[\sum_{(p,\ell) \in M} w_{p,\ell} \exp(h_{p,\ell}^\lambda) \right] \leq 1.$$

Now we apply Inequality (5.1) in Lemma 5.2 for $\pi = \sum_{(p,\ell) \in M} w_{p,\ell} \delta_{(p,\ell)}$ and $h = \sum_{(p,\ell) \in M} h_{p,\ell}^\lambda \mathbb{1}_{\Theta_{p,\ell}}$ and we obtain

$$\pi_0 \left[\exp\left(\sup_{\sum_{(p,\ell) \in M} w'_{p,\ell} = 1} \left\{ \sum_{(p,\ell) \in M} w'_{p,\ell} h_{p,\ell}^\lambda - \sum_{(p,\ell) \in M} w'_{p,\ell} \ln(w'_{p,\ell}/w_{p,\ell}) \right\}\right) \right] \leq 1$$

and, by Jensen's inequality, and replacing $h_{p,\ell}^\lambda$ by its definition,

$$(5.17) \quad \pi_0 \left[\sup_{\sum_{(p,\ell) \in M} w'_{p,\ell} = 1} \left\{ \sum_{(p,\ell) \in M} w'_{p,\ell} \sup_{\rho_{p,\ell} \in \mathcal{M}_+^1(\Theta_{p,\ell})} \exp\left(\lambda\rho_{p,\ell} \left[\lambda(\bar{R} - \bar{r}_n) - \ln \frac{d\rho_{p,\ell}}{d\pi_{p,\ell}} \right] - \frac{\lambda^2 k_n^2}{n(1-p/n)^2} + \ln \frac{w_{p,\ell}}{w'_{p,\ell}} \right) \right\} \right] \leq 1.$$

By Jensen again, we obtain a bound for the first term in the sum bounded in Lemma 5.4:

$$\pi_0 \left[\sup_{\sum_{(p,\ell) \in M} w'_{p,\ell} = 1} \left\{ \sum_{(p,\ell) \in M} w'_{p,\ell} \sup_{\rho_{p,\ell} \in \mathcal{M}_+^1(\Theta_{p,\ell})} \rho_{p,\ell} \left[\exp \left(\lambda(\bar{R} - \bar{r}_n) - \ln \frac{d\rho_{p,\ell}}{d\pi_{p,\ell}} - \frac{\lambda^2 k_n^2}{n(1-p/n)^2} + \ln \frac{w_{p,\ell}}{w'_{p,\ell}} \right) \right] \right\} \right] \leq 1.$$

Finally, we sum this inequality over all $\lambda \in \mathcal{G}$ to bound the first expectation.

The second expectation is bounded by choosing specific weights $w'_{p,\ell}$ in the supremum in inequality (5.17) such that $w'_{p,\ell} = 1$ for $(p, \ell) = \arg \max_M \{h_{p,\ell}\}$:

$$\pi_0 \left[\sup_{\substack{(p,\ell) \in M \\ \rho_{p,\ell} \in \mathcal{M}_+^1(\Theta_{p,\ell})}} \left\{ \exp \left(\lambda \rho_{p,\ell} [\bar{R} - \bar{r}_n] - \mathcal{K}(\rho_{p,\ell}, \pi_{p,\ell}) - \frac{\lambda^2 k_n^2}{n(1-p/n)^2} + \ln w_{p,\ell} \right) \right\} \right] \leq 1.$$

Again a summation over all $\lambda \in \mathcal{G}$ leads to the result. This ends the proof. \square

Proof of Lemma 5.5. From the proof of the Lemma 5.1, we already know that $|\bar{r}_n(\theta) - r_n(\theta)| \leq (1+L) \sum_{i=1}^n \|X_i - \bar{X}_i\|$. This bound holds uniformly on Θ . Now we are reduced to estimate $\pi_0[\|X_0 - \bar{X}_0\|]$. For this, we use the assumption (2.5) and the stationarity of X and \bar{X} . More precisely:

$$\begin{aligned} \pi_0[\|X_0 - \bar{X}_0\|] &\leq u\mu[\|\xi_0 - \bar{\xi}_0\|] + \sum_{j \geq 1} a_j(F) \pi_0[\|X_{-j} - \bar{X}_{-j}\|] \\ &\leq u\mu[\|\xi_0\| \mathbf{1}_{\|\xi_0\| > C}] + a(F) \pi_0[\|X_{-j} - \bar{X}_{-j}\|]. \end{aligned}$$

The result follows from the estimate $\mu[\|\xi_0\| \mathbf{1}_{\|\xi_0\| > C}] \leq \mu[\exp(c\|\xi_0\|)] C \exp(-cC)$ for any $c > 0$. \square

Now give the proof of the useful Proposition 3.1.

Proof of Proposition 3.1. Let us introduce a parameter $\zeta > 0$ then we have

$$\begin{aligned} -\frac{1}{\gamma} \ln \pi_{p,\ell} [\exp(-\gamma(R - R(\bar{\theta}_{p,\ell})))] - \zeta &= -\frac{1}{\gamma} \ln \pi_{p,\ell} [\exp(-\gamma(R - R(\bar{\theta}_{p,\ell}) - \zeta))] \\ &\leq -\frac{1}{\gamma} \ln \pi_{p,\ell} (R(\theta) - R(\bar{\theta}_{p,\ell}) \leq \zeta) \end{aligned}$$

Then we directly derive from the definition of $d_{p,\ell}$ that

$$d_{p,\ell} \leq \sup_{\gamma > e} \frac{\inf_{\zeta > 0} \{\zeta \gamma - \ln \pi_{p,\ell} (R(\theta) - R(\bar{\theta}_{p,\ell}) \leq \zeta)\}}{\ln \gamma}.$$

So

$$\zeta \gamma - \dim \ln \frac{\zeta}{C c_{p,\ell}} \leq \dim \wedge \gamma C (c_{p,\ell} - \|\bar{\theta}_{p,\ell}\|) + \dim \ln \left(\frac{C c_{p,\ell} \gamma}{\dim} \vee \frac{c_{p,\ell}}{c_{p,\ell} - \|\bar{\theta}_{p,\ell}\|} \right).$$

Now if $\dim \leq \gamma C (c_{p,\ell} - \|\bar{\theta}_{p,\ell}\|)$ then we get the estimate $\dim(1 + \ln(C c_{p,\ell} \gamma / \dim)) / \ln \gamma$ which decreases with γ . We then get the desired bound when the supremum is established for $\gamma = e \vee \dim / (C (c_{p,\ell} - \|\bar{\theta}_{p,\ell}\|))$. If $\dim \geq \gamma C (c_{p,\ell} - \|\bar{\theta}_{p,\ell}\|)$ then we get the estimate $(\gamma C (c_{p,\ell} - \|\bar{\theta}_{p,\ell}\|) + \dim \ln(c_{p,\ell} / (c_{p,\ell} - \|\bar{\theta}_{p,\ell}\|))) / \ln \gamma$ which increases with γ . Then we have to consider the case γ as

large as possible, that is when $\dim = \gamma C(c_{p,\ell} - \|\bar{\theta}_{p,\ell}\|)$ but then we are going back to an already treated case. \square

5.5. Proofs of results stated in Subsection 2.4. We first prove the existence of a solution of the chains with infinite memory (2.4). Then we prove the θ_∞ -weak dependence properties of this solution when innovations are bounded and of the bounded φ -mixing processes.

Proof of Proposition 2.1. Let us fix some $r \geq 1$. Then $\mu[\|\xi_0\|^r] < \infty$ from the Jensen's inequality as ξ_0 admits finite moments of exponential orders. Now we want to apply Theorem 3.1 of [14] to F that satisfies

$$\|F(0; \xi_0)\|_r \leq u \|\xi_0\|_r < \infty,$$

denoting $\|\cdot\|_r$ for the \mathbb{L}^r -norm $(\mu[\cdot^r])^{1/r}$. We also have for any $x = (x_j)_{j \in \mathbb{N}}$ and $x' = (x'_j)_{j \in \mathbb{N}}$ the following relation

$$\|F(x; \xi_0) - F(x'; \xi_0)\|_r \leq \sum_{j=1}^{\infty} a_j(F) \|x_j - x'_j\|.$$

Using assumption (2.6) we obtain the existence of a unique causal stationary solution to equation (2.4) such that $\pi[\|X_0\|^r] < \infty$. Finally this result holds for all $1 \leq r < \infty$ and we have proved the proposition. \square

In the sequel, we prove results of Lemmas 2.2 and 2.3. These two estimates of the θ_∞ -coefficients are obtained via a common classical technique that we present shortly below, see [11] for more details. The so-called coupling techniques consist in constructing a version $(X_t^*)_{t \in \mathbb{Z}}$ distributed as $(X_t)_{t \in \mathbb{Z}}$ and such that $(X_t^*)_{t > 0}$ is independent of $\mathfrak{S}_0 = \sigma(X_t, t \leq 0)$. If this process $(X_t^*)_{t > 0}$ is well defined, then it gives sharp estimates of the quantity $\theta_{\infty,n}(1)$ as we have the following version of the Kantorovitch-Rubinstein duality, see [11] for more details:

Lemma 5.7. *For any version $(X_t^*)_{t \in \mathbb{Z}}$ we have*

$$(5.18) \quad \theta_{\infty,n}(1) \leq \sum_{i=1}^n \|\mathbb{E}(\|X_i - X_i^*\|/\mathfrak{S}_0)\|_\infty.$$

For the sake of completeness, we recall the proof of this lemma.

Proof of Lemma 5.7. To compute a bound on the coefficients $\theta_\infty(\mathfrak{S}, Z)$ for this solution we first need to introduce coupling arguments coming from Dedecker *et al.* [12] associated with the τ_∞ coefficients defined as

$$\tau_\infty(\mathfrak{S}, Z) = \left\| \sup_{f \in \Lambda_1} \left| \mathbb{E}(f(Z)|\mathfrak{S}) - \mathbb{E}(f(Z)) \right| \right\|_\infty.$$

First note that this coefficient is in fact the same than θ_∞ . But we prefer in this section this formulation as it lets appear the supremum on the class of Lipschitz function, the Weisserstien metrics. If the space is enough rich, we have, as in [12], the Kantorovitch-Rubinstein equation

$$(5.19) \quad \tau_\infty(\mathfrak{S}, Z) = \inf_{Z^* \in V} \|\mathbb{E}(\|Z - Z^*\|/\mathfrak{S})\|_\infty$$

where V is the set of the random variables Z^* distributed as Z but independent of \mathfrak{S} .

To bound $\tau_{\infty,n}(1)$ for all n we have to consider a version of the whole process $(X_t)_{t \in \mathbb{Z}}$ denoted $(X_t^*)_{t \in \mathbb{Z}}$ such that $(X_t^*)_{t > 0}$ is independent of $\mathfrak{G}_0 = \sigma(X_t, t \leq 0)$. As we equip \mathcal{X}^n with the norm $\|(x_1, \dots, x_n)\| = \sum_{i=1}^n \|x_i\|$ we immediately get the inequality

$$\tau_{\infty,n}(1) \leq \|\mathbb{E}(\|(X_1, \dots, X_n) - (X_1^*, \dots, X_n^*)\| \mid \mathfrak{G}_0)\|_{\infty} \leq \sum_{i=1}^n \|\mathbb{E}(\|X_i - X_i^*\| \mid \mathfrak{G}_0)\|_{\infty}.$$

□

We conclude the proofs of Lemmas 2.2 and 2.3 by choosing carefully the version in order to get efficient bounds on $\|\mathbb{E}(\|X_i - X_i^*\| \mid \mathcal{M}_0)\|_{\infty}$ for all i . As there exists plenty of different coupling schemes in the literature, we now have to choose one that gives $(X_t^*)_{t \in \mathbb{Z}}$ that give efficient bounds for each Lemmas. In the Lemma 2.2 we use the forward coupling, see [22] for more details on this techniques. The maximal coupling of [15] is used for the proof of Lemma 2.3.

Proof of Lemma 2.2. We apply theorem 3.1 of [14] checking the relations (here $F(0; 0)$ is fixed to 0 for convenience)

$$\begin{aligned} \|F(0; \xi_0)\|_r \|_{\infty} &\leq u \|\xi_0\|_{\infty} < \infty, \\ \|F(x; \xi_0) - F(x'; \xi_0)\|_{\infty} &\leq \sum_{j=1}^{\infty} a_j(F) \|x_j - x'_j\|, \end{aligned}$$

where $\|\cdot\|_{\infty}$ denotes the $\mathbb{L}^{\infty}(\mu)$ -norm. As (2.6) holds, we can conclude of the existence of a unique causal stationary solution to (2.4) such that $\|X_0\|_{\infty} < \infty$. Moreover, it follows easily from the construction that $\|X_0\|_{\infty} < u \|\xi_0\|_{\infty} / (1 - a)$.

Now let us use the coupling Lemma 5.7 on X_t^* that we construct as follows. Let $(\xi^*)_{t \in \mathbb{Z}}$ be a stationary sequence distributed as $(\xi_t)_{t \in \mathbb{Z}}$, independent of $(\xi_t)_{t \leq 0}$ and such that $\xi_t = \xi_t^*$ for $t > 0$. Let $(X_t^*)_{t \in \mathbb{Z}}$ be the solution of the equation

$$X_t^* = F(X_{t-1}^*, X_{t-2}^*, \dots; \xi_t^*), \quad \text{a.e..}$$

Let $p \neq 0$ be an integer and $(X_t^{(p)})_{t \in \mathbb{Z}}$ be the solution, bounded by $u \|\xi_0\|_{\infty} / (1 - a)$, of the equation

$$(5.20) \quad X_t^{(p)} = F^{(p)}(X_{t-1}^{(p)}, \dots, X_{t-p}^{(p)}; \xi_t),$$

with $F^{(p)}(x_1, \dots, x_p; \xi) = F(x_1, \dots, x_p, 0, \dots; \xi)$ for all $(x_1, \dots, x_p) \in \mathcal{X}^p$. Let $(X_t^{(p)*})_t$ be the solution of Equation (5.20) with the innovation $(\xi^*)_{t \in \mathbb{Z}}$. This coupling scheme is the forward coupling one for the $(X_t^{(p)})_{t \in \mathbb{Z}}$ for all p . We have

$$\begin{aligned} &\|\mathbb{E}(\|X_r - X_r^*\| \mid \mathcal{M}_0)\|_{\infty} \\ &\leq \|\mathbb{E}(\|X_r - X_r^{(p)}\| \mid \mathcal{M}_0)\|_{\infty} + \|\mathbb{E}(\|X_r^{(p)} - X_r^{(p)*}\| \mid \mathcal{M}_0)\|_{\infty} + \|\mathbb{E}(\|X_r^* - X_r^{(p)*}\| \mid \mathcal{M}_0)\|_{\infty}. \end{aligned}$$

Remark that

$$\begin{aligned} \|\mathbb{E}(\|X_r^{(p)} - X_r^{(p)*}\| \mid \mathcal{M}_0)\|_{\infty} &\leq \|\mathbb{E}(\|F^{(p)}(X_{r-1}^{(p)}, \dots, X_{r-p}^{(p)}; \xi_r) - F^{(p)}(X_{r-1}^{(p)*}, \dots, X_{r-p}^{(p)*}; \xi_r)\| \mid \mathcal{M}_0)\|_{\infty} \\ &\leq \sum_{j=1}^p a_j(F) \|\mathbb{E}(\|X_{r-j}^{(p)} - X_{r-j}^{(p)*}\| \mid \mathcal{M}_0)\|_{\infty}. \end{aligned}$$

Denote $u_t = \sup_{j \geq t} \|\mathbb{E}(\|X_j^{(p)} - X_j^{(p)*}\| | \mathcal{M}_0)\|_\infty$ for all $t \in \mathbb{Z}$ that is bounded with $2u\|\xi_0\|_\infty/(1-a)$ by construction. Moreover for all $n > 0$, $u_n \leq a(F)u_{n-p}$ and then $u_n \leq a(F)^{\lfloor n/p \rfloor + 1}u_0$. Using that $\tilde{u}_0 \leq 2\|X_0^{(p)}\|_\infty \leq 2u\|\xi_0\|_\infty/(1-a)$ and that $\lfloor n/p \rfloor + 1 \geq n/p$ we obtain

$$\|\mathbb{E}(\|X_r^{(p)} - X_r^{(p)*}\| | \mathcal{M}_0)\|_\infty \leq 2 \frac{u\|\xi_0\|_\infty}{1-a} a(F)^{r/p}.$$

For the other terms we use the Lipschitz condition on F

$$\|\mathbb{E}(\|X_r - X_r^{(p)}\| | \mathcal{M}_0)\|_\infty \leq \frac{u\|\xi_0\|_\infty}{1-a} \cdot \sum_{j=p}^{\infty} a_j(F) \quad \text{and} \quad \|\mathbb{E}(\|X_r^* - X_r^{(p)*}\| | \mathcal{M}_0)\|_\infty \leq \frac{u\|\xi_0\|_\infty}{1-a} \cdot \sum_{j=p}^{\infty} a_j(F).$$

Finally, merging those two bounds we get

$$\|\mathbb{E}(\|X_r - X_r^*\| | \mathcal{M}_0)\|_\infty \leq 2 \frac{u\|\xi_0\|_\infty}{1-a} \inf_{0 < p < r} \left\{ a(F)^{r/p} + \sum_{j=p}^{\infty} a_j(F) \right\}.$$

We conclude by using (5.18) in Lemma 5.7. \square

Proof of Lemma 2.3. Here we will consider the maximal coupling scheme of [15]. There exists a version $(X_t^*)_{t \in \mathbb{Z}}$ such that

$$\|\mathbb{P}(X_t \neq X_t^* \text{ for some } t \geq r | \mathfrak{G}_0)\|_\infty = \sup_{(A,B) \in \mathfrak{G}_0 \times \mathfrak{F}_r} |\mathbb{P}(A/B) - P(B)| = \varphi(r).$$

Now let us denote \mathcal{X} the state space of $(X_t)_{t \in \mathbb{Z}}$. As $\|X_t\|_\infty \leq C$ we can always fix \mathcal{X} such that $\|x - y\| \leq 2C\mathbb{1}_{x \neq y}$ for every x, y in \mathcal{X} . Thus we have:

$$\begin{aligned} \|\mathbb{E}(\|X_i - X_i^*\| | \mathfrak{G}_0)\|_\infty &\leq \|\mathbb{E}(\mathbb{1}_{X_i \neq X_i^*} | \mathfrak{G}_0)\|_\infty \\ &\leq 2C\|\mathbb{E}(\mathbb{1}_{X_i \neq X_i^*} | \mathfrak{G}_0)\|_\infty \\ &\leq 2C\|\mathbb{P}(X_i \neq X_i^* | \mathfrak{G}_0)\|_\infty \\ &\leq 2C\varphi(i). \end{aligned}$$

The last inequality follows from the rough bound

$$\mathbb{P}(X_i \neq X_i^* | \mathfrak{G}_0) \leq \mathbb{P}\left(\bigcup_{t \geq i} X_t \neq X_t^* | \mathfrak{G}_0\right).$$

We conclude by using (5.18) in Lemma 5.7. \square

5.6. Proofs of results in Section 4. We proof the Corollaries 4.3 and 4.4 of Theorem 3.2 applied in the context of Neural Networks and projection in the Fourier basis predictors.

Proof of Proposition 4.3. Firstly we check that all the predictors are L -Lipschitz functions of the observations. For any $x, y \in \mathbb{R}^p$, as the function called clip is 1-Lipschitz, we have

$$\begin{aligned} |f_\theta(x) - f_\theta(y)| &\leq \left| \sum_{k=1}^{\ell} c_k (\phi(a_k \cdot x + b_k) - \phi(a_k \cdot y + b_k)) \right| \\ &\leq D_1 \sum_{k=1}^{\ell} |c_k| |a_k \cdot (x - y)| \leq D_1 \sum_{k=1}^{\ell} |c_k| \|a_k\|_1 \|x - y\|_\infty \leq D_1 \sum_{k=1}^{\ell} \|a_k\|_1 \sum_{i=1}^p |c_k| |x_i - y_i|. \end{aligned}$$

Then when $\theta \in \mathcal{B}_{p,\ell}$ we are sure that $L = D_1(\tau_\ell + (3\ell)^{-1})(C_p + 1/3)$ is a convenient Lipschitz constant.

Secondly we use the approximation estimates given in [6]. More precisely, using Jensen to estimate \mathbb{L}_1 -risk by \mathbb{L}_2 -risk, we know that

$$\pi_0 \left[\left| \text{med}(X_0 | X_{-1}, \dots, X_{-p}) - f_{\bar{\theta}_{p,\ell}}(X_{-1}, \dots, X_{-p}) \right| \right] \leq \frac{2C_p}{\sqrt{\ell}}.$$

Then using Theorem 3.2 there exists some constant $C > 0$ such that for sufficiently large n and as soon as

$$(5.21) \quad K \geq \frac{\|X\|_\infty + 2\theta_{\infty,\infty}(1)}{1 + L},$$

we have

$$\pi_0[R(\hat{\theta})] \text{ and } \pi_0[R(\tilde{\theta})] \leq \inf_{p,\ell} \left\{ R(\bar{\theta}_{p,\ell}) + C \left(\sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell}n) + \frac{\ln n}{\sqrt{d_{p,\ell}n} \ln(d_{p,\ell}n)} \right) \right\}.$$

Then let us remark that (5.21) is always satisfied for sufficiently large n as in fact L goes to ∞ with n through τ_ℓ and ℓ . On the opposite, using the estimate of $d_{p,\ell}$ and the assumption on C_p we know that for n large and for some constant C it holds $d_{p,\ell} \leq Cp\ell \ln(n)$. Combining with the approximation bound, it holds

$$\pi_0[R(\hat{\theta})] \text{ and } \pi_0[R(\tilde{\theta})] \leq \inf_{p,\ell} \left\{ \pi_0 [|X_0 - \text{med}(X_0 | X_{-1}, \dots, X_{-p})|] + \frac{2C_p}{\sqrt{\ell}} + C \sqrt{\frac{p\ell}{n}} \ln^{3/2}(n) \right\}.$$

When fixing $\ell = \sqrt{n/p} \ln^{-3/2}(n)$ the result follows. \square

Proof of Proposition 4.4. Let us apply Theorem 3.3 and we obtain for some constant C the relation

$$\begin{aligned} \pi_0[R(\hat{\theta})] \text{ and } \pi_0[R(\tilde{\theta})] &\leq \inf_{p,\ell} \left\{ R(\bar{\theta}_{p,\ell}) + C \sqrt{\frac{d_{p,\ell}}{n}} \ln(d_{p,\ell}n) \ln(n) \right\} \\ &\leq \inf_{\ell} \left\{ R(\bar{\theta}_{p_0,\ell}) + C \sqrt{\frac{d_{p_0,\ell}}{n}} \ln(d_{p_0,\ell}n) \ln(n) \right\}. \end{aligned}$$

Now, we have

$$\begin{aligned} R(\bar{\theta}_{p_0,\ell}) &= \inf_{\theta \in \Theta} \pi_0 \left[\left| X_{p+1} - f_{\bar{\theta}_{p_0,\ell}}(X_p, \dots, X_1) \right| \right] \\ &\leq \pi_0 \left[\left| X_{p+1} - \sum_{i=1}^{p_0} f_i(X_{p-i}) \right| \right] + \inf_{\theta \in \Theta} \mathbb{E}_{\pi_0} \left(\left| \sum_{i=1}^{p_0} f_i(X_{p-i}) - \sum_{i=1}^{p_0} \sum_{j=1}^n \theta_{i,j} \varphi_j(X_{p-i}) \right| \right) \\ &\leq \mu(|\xi_0|) + \inf_{\theta \in \Theta} \sum_{i=1}^{p_0} \pi_0 \left[\left| f_i(X_1) - \sum_{j=1}^n \theta_{i,j} \varphi_j(X_1) \right| \right]. \end{aligned}$$

Now, note that the hypothesis on the process implies that X_1 has a density upper bounded by $1/\sqrt{2\pi\sigma^2}$ and so we obtain

$$\begin{aligned}
 R(\bar{\theta}_{p_0, \ell}) &\leq \mu(|\xi_0|) + \frac{1}{\sqrt{2\pi\sigma^2}} \inf_{\theta \in \Theta} \sum_{i=1}^{p_0} \int \left| f_i(x) - \sum_{j=1}^n \theta_{i,j} \varphi_j(x) \right| dx \\
 &\leq \mu(|\xi_0|) + \frac{1}{\sqrt{2\pi\sigma^2}} \inf_{\theta \in \Theta} \sum_{i=1}^{p_0} \left(\int \left[f_i(x) - \sum_{j=1}^n \theta_{i,j} \varphi_j(x) \right]^2 dx \right)^{\frac{1}{2}} \\
 &\leq \mu(|\xi_0|) + \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{i=1}^{p_0} \gamma_i \ell^{-s_i} \leq \mu(|\xi_0|) + \frac{\sum_{i=1}^{p_0} \gamma_i}{\sqrt{2\pi\sigma^2}} \ell^{-s}.
 \end{aligned}$$

So now we have

$$(5.22) \quad \pi_0[R(\hat{\theta})] \text{ and } \pi_0[R(\tilde{\theta})] \leq \mu(|\xi_0|) + \inf_{\ell} \left\{ \ell^{-s} \frac{\sum_{i=1}^{p_0} \gamma_i}{\sqrt{2\pi\sigma^2}} + C \sqrt{\frac{d_{p_0, \ell}}{n}} \ln(d_{p_0, \ell} n) \ln(n) \right\}$$

Now, we estimate $d_{p_0, \ell}$ using Proposition 3.1 and we obtain

$$d_{p, \ell} = p\ell \left(1 + \ln \left(L \left(\frac{e}{p\ell} \vee \frac{1}{a} \right) \right) \right).$$

We plug it into Equation (5.22) to obtain for some $C > 0$ and sufficiently large n

$$\pi_0[R(\hat{\theta})] \text{ and } \pi_0[R(\tilde{\theta})] \leq \mu(|\xi_0|) + \inf_{\ell} \left\{ \ell^{-s} \frac{\sum_{i=1}^{p_0} \gamma_i}{\sqrt{2\pi\sigma^2}} + C \sqrt{\frac{p_0 \ell}{n}} \ln(p_0 \ell n) \ln(n) \right\}.$$

In particular fixing ℓ proportional to $n^{\frac{1}{2s+1}}$ leads to the result. \square

REFERENCES

- [1] AKAIKE, H. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory (Tsahkadzor, 1971)* (1973), 267–281.
- [2] ALQUIER, P. PAC-bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods in Statistics* 17, 4 (2008), 279–304.
- [3] ANDREWS, D. W. K. Nonstrong mixing autoregressive processes. *J. Appl. Probab.* 21, 4 (1984), 930–934.
- [4] AUDIBERT, J.-Y. Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré: Probability and Statistics* 40, 6 (2004), 685–736.
- [5] BARAUD, Y., COMTE, F., AND VIENNET, G. Adaptative estimation in autoregression or β -mixing regression via model selection. *The Annals of Statistics* 29, 3 (2001), 839–875.
- [6] BARRON, A. R. Approximation and estimation bounds for artificial neural networks. *Machine Learning* 14 (1994), 115 – 133.
- [7] CATONI, O. A PAC-bayesian approach to adaptative classification. *Preprint Laboratoire de Probabilités et Modèles Aléatoires* (2003).
- [8] CATONI, O. *Statistical Learning Theory and Stochastic Optimization, Lecture Notes in Mathematics (Saint-Flour Summer School on Probability Theory 2001)*. Springer, 2004.
- [9] CATONI, O. *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, vol. 56 of *Lecture Notes-Monograph Series*. IMS, 2007.
- [10] DEDECKER, J., DOUKHAN, P., LANG, G., LEÓN, J. R., LOUHICHI, S., AND PRIEUR, C. *Weak Dependence, Examples and Applications*, vol. 190 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 2007.
- [11] DEDECKER, J., AND PRIEUR, C. New dependence coefficients: Examples and applications to statistics. *Probability Theory and Related Fields* 132 (2005), 203–235.
- [12] DEDECKER, J., PRIEUR, C., AND RAYNAUD DE FITTE, P. Parametrized kantorovich-rubinstein theorem and application to the coupling of random variables. In *Dependence in Probability and Statistics* (New York, 2006), P. Bertail, P. Doukhan, and P. Soulier, Eds., vol. 187 of *Lecture Notes in Statistics*, Springer, pp. 105–121.
- [13] DOUKHAN, P. *Mixing*, vol. 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994.

- [14] DOUKHAN, P., AND WINTENBERGER, O. Weakly dependent chains with infinite memory. *Stochastic Processes and their Applications* 118 (2008), 1997–2013.
- [15] GOLDSTEIN, S. Maximal coupling. *Z. Wahrsch. verw. Gebiete* 46 (1979), 193–204.
- [16] IBRAGIMOV, I. Some limit theorems for stationary processes. *Theory of Probab. Appl.* 7 (1962), 349–382.
- [17] ING, C.-K. Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Ann. Statist.* 35 (2007), 1238–1277.
- [18] ING, C.-K., AND WEI, C.-Z. Order selection for same-realization predictions in autoregressive processes. *Ann. Statist.* 33 (2005), 2423–2474.
- [19] LUGOSI, G., AND CESA-BIANCHI, N. *Prediction, Learning and Games*. Cambridge University Press, 2006.
- [20] MASSART, P. *Concentration Inequalities and Model Selection, Lecture Notes in Mathematics (Saint-Flour Summer School on Probability Theory 2003)*. Springer, 2006.
- [21] MCALLESTER, D. A. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT, Madison, WI, 1998)* (1998), ACM, pp. 230–234.
- [22] MEYN, S., AND TWEEDIE, R. *Markov chains and stochastic stability*. Springer-Verlag, London, 1993.
- [23] MODHA, D. S., AND MASRY, E. Memory-universal prediction of stationary random processes. *IEEE transactions on information theory* 44, 1 (1998), 117–133.
- [24] RIO, E. Ingalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Académie des Sciences de Paris, Serie I* 330 (2000), 905–908.
- [25] RIO, E. *Théorie asymptotique des processus aléatoires faiblement dépendants*, vol. 31 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2000.
- [26] TSYBAKOV, A. *Introduction à l'estimation non paramétrique*. Mathématiques et Applications, Springer, 2004.
- [27] VAPNIK, V. N. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1998.

LABORATOIRE DE PROBABILITÉS ET MODÈLES ALÉATOIRES, UNIVERSITÉ PARIS 7 (DIDEROT), 175, RUE DU CHEVALERET, 75252 PARIS CEDEX 05, FRANCE, ET CREST-LS.

E-mail address: alquier@math.jussieu.fr

CEREMADE PLACE DU MARCHAL DE LATTRE DE TASSIGNY 75775 PARIS CEDEX 16 FRANCE

E-mail address: wintenberger@ceremade.dauphine.fr