MODEL SELECTION FOR WEAKLY DEPENDENT TIME SERIES FORECASTING

PIERRE ALQUIER[(1)] AND OLIVIER WINTENBERGER[(2)]

Abstract. Observing a stationary time series, we propose a two-step procedure for the prediction of the next value of the time series. The first step follows machine learning theory paradigm and consists in determining a set of possible predictors as randomized estimators in (possibly numerous) different predictive models. The second step follows the model selection paradigm and consists in choosing one predictor with good properties among all the predictors of the first steps. We study our procedure for two different types of observations: causal Bernoulli shifts and bounded weakly dependent processes. In both cases, we give oracle inequalities: the risk of the chosen predictor is close to the best prediction risk in all predictive models that we consider. We apply our procedure for predictive models such as linear predictors, neural networks predictors and non-parametric autoregressive predictors.

Contents

2000 *Mathematics Subject Classification.* Primary 62M20; Secondary 62G05, 68Q32, 68T05, 62M10, 91E40.
Key words and phrases. Time series prediction, autoregression estimation, adaptative inference, statistical learning, randomized estimators, aggregation of estimators, model selection, weak dependence.

1. Introduction

When observing a time series, one crucial issue is to predict the (non-observed) first future value using the observed past values. Since the seventies, different model selection procedures have been studied for inferring how many observed past values are needed for predicting the next value. Procedures such as AIC ([1]), BIC (Schwarz [26]) and APE (Ing [17]) are used by practitioners to select a reasonable linear predictor. When the observations satisfy a linear model, those procedures are proved to be asymptotically efficient (see Ing [17] for more details).

In the same time, the progress of statistical learning theory in the iid setting brought new perspectives in model selection (see Vapnik [29] and Massart [20] among others). Typical machine-learning procedures allow to choose a predictor among a family, with the guarantee that this predictor performs almost as well as the best possible predictor of the family (called the oracle). Such results are called oracle inequalities; they provide guarantees on the quality of the prediction without any parametric assumption on the observations.

Few works have been done in the context of dependent observations. The machine learning theory was used successfully in the time series prediction context by Modha and Masry [22]. However, their procedure relies on the knowledge of the α-mixing coefficients. To our knowledge, there is no efficient estimation of this coefficients and their procedure seems difficult to use in practice. Baraud *et al.* [5] use the model selection point of view to perform regression and auto-regression on dependent observations. They prove powerful oracle inequalities when the observations satisfy an additive auto-regressive model. When the observations are Harris recurrent Markov chains, Lacour [18] gives also oracle inequalities for a procedure completely free of the dependence properties. An alternative point of view is provided by the theory of individual sequences prediction (see Lugosi and Cesa-Bianchi [19] or Stoltz [28]). In these works, no assumption on the observations - not even a stochastic assumption - is done and oracle inequalities are given.

In this paper, our objectives are the following:

(1) to build various predictors of different form and using different number of past observations,
(2) to select one of these predictors *without any assumptions on the distribution of the observations*,
(3) to prove oracle inequalities under weak assumptions on the observed time series.

In the end of this introduction, let us briefly present our two-step procedure.

Let us 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 $\|\cdot\|$. Let us fix a (possibly large) family of predictors $\{f_\theta, \theta\in\Theta\}$: for any θ and any t, f_θ applied to the past values $(X_{t-1}, X_{t-2}, \dots, X_1)$ is a possible prediction of X_t. We discretize the family of predictors by the number p of past values they use for prediction. Thus we assume that

$$\Theta = \bigcup_{p=1}^{\lfloor \frac{n}{2} \rfloor} \Theta_p$$

where the Θ_p are disjoint in order that for any $\theta \in \Theta$, there is only one p such that $\theta \in \Theta_p$. Now, for any $\theta \in \Theta_p$, f_θ is a function $\mathcal{X}^p \to \mathcal{X}$ and at any time t, $f_\theta(X_{t-1}, ..., X_{t-p})$ is a prediction of X_t according to θ and denoted $\hat{X}_t^\theta$. As the predictor f_θ may take different forms (linear, neural networks,...), we write

$$\Theta_p = \bigcup_{\ell=1}^{m_p} \Theta_{p,\ell}$$

for a given $m_p \in \{1, \ldots, n\}$. Finally, the risk $R(\theta)$ of the prediction is defined by

$$R(\theta) = \pi_0 \left[\| f_\theta(X_{t-1}, ..., X_{t-p}) - X_t \| \right] = \pi_0 \left[\left\| \hat{X}_t^\theta - X_t \right\| \right],$$

where here and all along the paper $\pi[h] = \int h d\pi$ for any measure π and any integrable function h. Note that $R(\theta)$ does not depend on t as X is stationary.

The mathematical counterparts of the points (1), (2) and (3) of our objectives are the following. The point (1) corresponds to build, on the basis of the observations, an estimator $\hat{\theta}_{p,\ell}$ in each model $\Theta_{p,\ell}$, for $1 \leq p \leq \lfloor n/2 \rfloor$ and $1 \leq \ell \leq m_p$. The point (2) consists in defining a procedure to choose a $\hat{\theta}$ among all the possible $\hat{\theta}_{p,\ell}$. Finally, for point (3), we prove that $R(\hat{\theta})$ is close to $\inf_{\theta \in \Theta} R(\theta)$. To attain our objectives we use the PAC-Bayesian paradigm (introduced by Shawe-Taylor and Williamson [27] and McAllester [21]). Using this approach, Catoni [7, 8, 9], Audibert [4], Alquier [2], Tsybakov and Dalalyan [10] solve points (1), (2) and (3) simultaneously for various regression and classification problems in the iid setting. In this paper we build a procedure that gives a predictor $\hat{\theta}$ satisfying, under general conditions on X and with probability at least $1 - \varepsilon$,

$$R(\hat{\theta}) \leq \inf_{d_{p,\ell} \leq n} \left\{ \min_{\theta \in \Theta_{p,\ell}} R(\theta) + \text{cst.} \sqrt{\frac{d_{p,\ell}}{n}} \log^{5/2}(n) \right\} + \text{cst.} \frac{\log \frac{1}{\varepsilon}}{\sqrt{n}}$$

where cst > 0 is an explicit constant and $d_{p,\ell}$ an estimate of the complexity of $\Theta_{p,\ell}$.

To obtain such oracle inequalities, we use sharp estimates (close to the ones in the iid case) on the Laplace transform of partial sums in dependent settings. For bounded observations we use the θ_∞-coefficients (see [11]), introduced in Rio [24] as the γ-mixing coefficients. These coefficients generalize the uniform mixing ones. For unbounded observations we use the causal Bernoulli shifts representation. It includes all classical linear ARMA models and also the more general chains with infinite memory introduced by Doukhan and Wintenberger [14]. These bounded and unbounded dependent frameworks are not comparable with the β or α-mixing ones as they include some dynamical systems that are not mixing, see Andrews [3] and Dedecker and Prieur [12] for details. Finally, it is important to note that our prediction procedure is the same for the two dependence frameworks. It does not depend on any unknown dependence coefficients; we believe that similar oracle inequalities for our predictor can be proved in other weakly dependent settings.

The paper is organized as follows: First, the prediction procedure is detailed in Section 2; Second, the assumptions on the observed time series and the corresponding oracle inequalities are given in Section 3. In Section 4 are given some examples of time series for which these oracle inequalities hold. Our procedure applied on some possible prediction models are given in Section

5. Linear predictors (with simulations), neural networks predictors and non-parametric predictors are considered. Finally the complete proofs are collected in Section 6.

2. The prediction procedure

We observe $(X_1, \ldots, X_n)$ from a stationary time process $X = (X_t)_{t \in \mathbb{Z}}$ distributed as π_0 on $\mathcal{X}^{\mathbb{Z}}$ where $\mathcal{X}$ is an Hilbert space equipped with its usual norm $\|\cdot\|$. We fix a family of predictors $\{f_\theta, \theta \in \Theta\}$ with

$$\Theta = \bigcup_{p=1}^{\lfloor \frac{n}{2} \rfloor} \Theta_p = \bigcup_{p=1}^{\lfloor \frac{n}{2} \rfloor} \left(\bigcup_{\ell=1}^{m_p} \Theta_{p,\ell} \right)$$

such that $m_p \geq n$ and $p(\theta)$ is the only p such that $\theta \in \Theta_p$. For any $\theta \in \Theta$, we denote $\hat{X}_t^\theta = f_\theta(X_{t-1}, ..., X_{t-p})$ and $R(\theta) = \pi_0 \left[\left\| \hat{X}_t^\theta - X_t \right\| \right]$.

2.1. The Lipschitz predictors.

Let M denotes the set of all possible pairs (p, ℓ):

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

Let $\mathcal{T}$ be a σ-algebra on Θ and $\mathcal{T}_{p,\ell}$ be its restriction to $\Theta_{p,\ell}$ for any $(p,\ell) \in M$. For any $(p,\ell) \in M$, we assume that $\Theta_{p,\ell}$ is a compact subset of $\mathbb{R}^q$ for some $q < \infty$ (q depends on (p,ℓ)) and that there exists $(a_j(\theta))_{j \in \{1,...,p\}}$ satisfying, for any $(x_1, ..., x_p), (y_1, ..., y_p) \in \mathcal{X}^p$, the relation

$$\|f_\theta(x_1, ..., x_p) - f_\theta(y_1, ..., y_p)\| \leq \sum_{j=1}^{p} a_j(\theta) \|x_j - y_j\| . \tag{2.1}$$

Moreover, we assume that

$$L := \sup_{(p,\ell) \in M} \sup_{\theta \in \Theta_{p,\ell}} \sum_{j=1}^{p} a_j(\theta) \qquad \text{satisfies} \qquad L \leq \log(n) - 1 \tag{2.2}$$

in order to bound the volatility of the predictors uniformly on M.

2.2. The complexity of $\Theta_{p,\ell}$.

To control the complexity of each $\Theta_{p,\ell}$ we assume that, for all $(p,\ell) \in M$, there exist a probability measure $\pi_{p,\ell}$ on the measurable space $(\Theta_{p,\ell}, \mathcal{T}_{p,\ell})$ and a constant $1 \leq d_{p,\ell} < \infty$ satisfying

$$\sup_{\gamma > e} \left\{ \frac{-\log \int_{\Theta_{p,\ell}} \left[\exp\left(-\gamma \left(R(\theta) - R\left(\overline{\theta}_{p,\ell}\right) \right) \right) \right] d\pi_{p,\ell}(\theta)}{\log(\gamma)} \right\} \leq d_{p,\ell}. \tag{2.3}$$

Here $\overline{\theta}_{p,\ell} = \arg\min_{\Theta_{p,\ell}} R$ for any $(p,\ell) \in M$. The parameter $d_{p,\ell}$ is linked with classical dimensions as the Vapnik dimension and entropy measures. In this paper we only investigate the cases where $\pi_{p,\ell}$ is the Lebesgue measure on $\Theta_{p,\ell}$. We have the following result:

Proposition 2.1. *Let $q \in \mathbb{N}^*$, $x > 0$ and $\mathcal{B}_x^q$ be the closed ℓ^1-ball in $\mathbb{R}^q$ of radius $x > 0$ and centered at 0. If $\Theta_{p,\ell} = \mathcal{B}_{c_{p,\ell}}^q$ for $c_{p,\ell} > 0$ and $\theta \to R(\theta)$ is a C-Lipschitz function then we have:*

$$d_{p,\ell} \leq q \times \left(1 + \log \left(c_{p,\ell} \left(\frac{Ce}{q} \vee \frac{1}{c_{p,\ell} - \|\overline{\theta}_{p,\ell}\|} \right) \right) \right). \tag{2.4}$$

The proof of this result is given at the end of Subsection 6.4. Predictive models whom complexity $d_{p,\ell}$ is estimated are given in Section 5.

2.3. **The empirical risk.** As the risk $R(\theta)$ cannot be computed, we use its empirical counterpart $r_n(\theta)$:

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

2.4. **The randomized estimators.** For any $(p,\ell) \in M$, our randomized estimators $\tilde{\theta}^{\lambda}_{p,\ell}$ is drawn randomly through a Gibbs measure

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

We recall that for any measure π and any measurable function h such that $\pi[\exp(h)] < +\infty$, the Gibbs measure denoted $\pi\{h\}$ is defined by the relation:

$$\frac{d\pi\{h\}}{d\pi}(\theta) = \frac{\exp(h(\theta))}{\pi[\exp(h)]}. \tag{2.5}$$

Here the parameter λ is called the temperature (this terminology comes from the statistical thermodynamics). For $n \geq 8e(1+L)$, λ takes values in a finite grid $\mathcal{G}_{p,\ell}$ defined as

$$\mathcal{G}_{p,\ell} = \left\{ g_1 \frac{\sqrt{d_{p,\ell} n}\log(d_{p,\ell} n)}{(1+L)\log^{3/2}(n)}, \ldots, g_{n_0} \frac{\sqrt{d_{p,\ell} n}\log(d_{p,\ell} n)}{(1+L)\log^{3/2}(n)} \right\} \cap \left[2e, \frac{n}{4(1+L)} \right].$$

where $\check{c} \leq g_1 < \ldots < g_{n_0} \leq \hat{c}$ with $2 \leq n_0 \leq n$ and $0 < \check{c} < 2/(1+L) < 2e(1+L) < \hat{c} < \infty$. Remark that when λ grows, $\pi_{p,\ell}\{-\lambda r_n\}$ tends to concentrate around the minimizer of the empirical risk $r_n(.)$.

2.5. **The model selection.** Classically one choose the minimizer of the penalized empirical risk $\arg\min_{p,\ell}[r_n(\tilde{\theta}^{\lambda}_{p,\ell}) + \mathrm{pen}(p,\ell,\lambda)]$, for some well chosen penalization $\mathrm{pen}(p,\ell,\lambda)$, see Massart [20]. Here we consider $\hat{\theta} = \tilde{\theta}^{\hat{\lambda}}_{\hat{p},\hat{\ell}}$ where

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

The model criterion $\hat{R}\,(p,\ell,\lambda)$ is given by the PAC-Bayesian approach:

$$\hat{R}\,(p,\ell,\lambda) = -\frac{1}{\lambda}\log \int_{\Theta_{p,\ell}} \exp\left(-\lambda r_n(\theta)\right) d\pi_{p,\ell}(\theta) + \frac{1}{\lambda}\log\left(n \left\lfloor \frac{n}{2} \right\rfloor m_p\right) + \frac{\lambda(1+L)^2\log^3(n)}{n(1-p/n)^2}.$$

3. Main results

In order to prove that $R(\hat{\theta})$ is close to $\inf_{\theta\in\Theta} R(\theta)$ with hogh probability, we restrict our study to two different contexts. Note that $\hat{\theta}$ is defined independently of these contexts and that a practitioner may compute our predictor on any observed time series.

3.1. **Bounded weakly dependent processes (WDP).** In this case X is bounded, i.e. $\|X\|_\infty := \sup_t \|X_t\| < \infty$. We use the $\theta_{\infty,n}(1)$-coefficients in Dedecker *et al.* [11], a version of the γ-mixing of Rio [25]) adapted to stationary time series. If Z is a bounded variable in $\mathcal{X}^q$ ($q \geq 1$) defined on $(\Omega, \mathcal{A}, \mathbb{P})$, for any σ-algebra $\mathfrak{S}$ of $\mathcal{A}$ we have:

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

where Λ_1 is the set of real 1-Lipschitz functions on $\mathcal{X}^q$ equipped with the norm $\|z\| = \sum_{i=1}^q \|z_i\|$. Let us define the σ-algebra $\mathfrak{S}_p = \sigma(X_t, t \leq p)$ for any $p \in \mathbb{Z}$ and the coefficients

$$\theta_{\infty,k}(1) = \sup \left\{ \theta_\infty\big(\mathfrak{S}_p, (X_{j_1}, \ldots, X_{j_\ell})\big), \quad p+1 \leq j_1 < \cdots < j_\ell, \quad 1 \leq \ell \leq k \right\}.$$

Moreover assume that there is a constant $\mathcal{C} > 0$ such that for any n, $\theta_{\infty,n}(1) < \mathcal{C}$ (the short memory condition). Causal Bernoulli shifts with bounded innovations, uniform φ-mixing sequences and dynamical systems are classical θ_∞ weakly-dependent examples, see Section 4 for more details. In this context we prove the following oracle inequality

Theorem 3.1. *Under* **(WDP)** *and condition* (2.3)*, there are explicit constants*

$$(\mathrm{cst}_1, \mathrm{cst}_2) = \mathrm{cst}(\check{c}, \hat{c}, L, \mathcal{C}, \|X_0\|_\infty)$$

such that for all $n \geq 8e(1+L)$ *with probability at least* $1-\varepsilon$

$$R(\hat{\theta}) \leq \inf_{d_{p,\ell} \leq n} \left\{ \min_{\theta \in \Theta_{p,\ell}} R(\theta) + \mathrm{cst}_1 . \sqrt{\frac{d_{p,\ell}}{n}} \log^{5/2}(n) \right\} + \mathrm{cst}_2 . \frac{\log \frac{1}{\varepsilon}}{\sqrt{n}} + 4(1+L) \left(\frac{(\|X_0\|_\infty + \mathcal{C})^2}{2} - \log^3(n) \right)_+ .$$

The proof of this result is given in Subsection 6.2 page 15.

3.2. **Causal Bernoulli shifts (CBS).** Let $\mathcal{X}'$ be some Banach space equipped with a norm also denoted $\|\cdot\|$. Let $H : \mathcal{X}'^{\mathbb{N}} \mapsto \mathcal{X}$ be a Lipschitzian function, i.e. such that there exists $(a_j(H))_{j \in \mathbb{N}}$ satisfying, for any $v = (v_j)_{j \in \mathbb{N}}$, $v' = (v'_j)_{j \in \mathbb{N}} \in \mathcal{X}'^{\mathbb{N}}$, the relations:

$$\big\|H(v) - H(v')\big\| \leq \sum_{j=0}^{\infty} a_j(H) \|v_j - v'_j\|, \tag{3.1}$$

$$\text{with } \sum_{j=0}^{\infty} j a_j(H) < +\infty. \tag{3.2}$$

We denote $\sum_{j=0}^{\infty} a_j(H) := a(H)$, $\sum_{j=0}^{\infty} j a_j(H) = \tilde{a}(H)$. The causal Bernoulli shifts are defined by the relation

$$X_t = H(\xi_t, \xi_{t-1}, \xi_{t-2}, \ldots) \qquad \forall t \in \mathbb{Z}$$

where ξ_t for $t \in \mathbb{Z}$ are iid variables called the innovations and distributed as μ. We assume that the innovation's norm admits a finite Laplace transform $\mu[\exp(c^* \|\xi_0\|)] := \Psi(c^*) < +\infty$ (the Cramer condition) for $c^* \geq a(H)$. Classical examples of such processes are causal linear ARMA models and chains with infinite memory with low-tail innovations, see Section 4 for more details. In this context we prove the following oracle inequality

Theorem 3.2. *Under* **(CBS)** *and condition* (2.3), *there are explicit constants*

$$(\mathrm{cst}'_1, \mathrm{cst}'_2) = \mathrm{cst}'(\check{c}, \hat{c}, L, a(H), \tilde{a}(H), \Psi(1))$$

such that for all $n \geq 8e(1+L)$ with probability at least $1-\varepsilon$

$$R(\hat{\theta}) \leq \inf_{d_{p,\ell}\leq n} \left\{ \min_{\theta\in\Theta_{p,\ell}} R(\theta) + \mathrm{cst}'_1.\sqrt{\frac{d_{p,\ell}}{n}}\log^{5/2}(n) \right\} + \mathrm{cst}'_2.\frac{\log\frac{1}{\varepsilon}}{\sqrt{n}}$$
$$+ \sqrt{\frac{d_{\hat{p},\hat{\ell}}}{n}}\log(d_{\hat{p},\hat{\ell}}n)4(1+L)\hat{c}\left(4a(H)\Psi(a(H)) + 2\log^2(n)\Big(1+\frac{\tilde{a}(H)}{a(H)}\Big)^2 - \log^3(n)\right)_+ .$$

The proof of this result is given in Subsection 6.3 page 17.

3.3. **Comments on the results.** The constants are roughly (but explicitly) estimated in the proofs, see Subsections 6.2 and 6.3. For example, we obtain

$$\mathrm{cst}_1 \leq (1+L)\left(\frac{6}{\check{c}} + 8\hat{c}\Big(1+\|X_0\|_\infty + \mathcal{C}\Big)^2\right) \quad \text{and} \quad \mathrm{cst}_2 \leq \frac{7(1+L)}{\check{c}}.$$

For n sufficiently large, the last terms in the oracle inequalities vanish. Then it exists a constant $C>0$ such that under **(WDP)** or **(CBS)** for all $n \geq 8e(1+L)$ with probability at least $1-\varepsilon$:

$$R(\hat{\theta}) \leq \inf_{d_{p,\ell}\leq n} \left\{ \min_{\theta\in\Theta_{p,\ell}} R(\theta) + C\sqrt{\frac{d_{p,\ell}}{n}}\log^{5/2}(n) \right\} + C\frac{\log\frac{1}{\varepsilon}}{\sqrt{n}}.$$

Similar oracles inequalities have already proved by Modha and Masry [22] and Baraud *et al.* [5]. These inequalities are given in expectation while ours are true with high probability. Remark that integrating our oracle inequalities with respect to ε leads to a result in expectation: there exists a constant $C>0$ independent of n such that in both **(WDP)** and **(CBS)** cases

$$\pi_0\left[R(\hat{\theta})\right] \leq \inf_{d_{p,\ell}\leq n} \left\{ \min_{\theta\in\Theta_{p,\ell}} R(\theta) + C\sqrt{\frac{d_{p,\ell}}{n}}\log^{5/2}(n) \right\}.$$

The converse is not true: results in expectation do not lead to results that hold with high probability.

It is difficult to compare our oracle inequalities with the ones in [22] and [5]: contrary with our paper, they deal with the quadratic risk and ($\beta-$ or $\alpha-$) mixing time series. However, let us remark that we obtain the same additional term $\sqrt{d_{p,\ell}/n}$ (called the rate) than in the iid case, up to a multiplicative $\log^{5/2}(n)$ term (called the loss). Baraud *et al.* [5] obtain the same rate than in the iid case associated with the quadratic risk, while Modha and Masry [22] suffer a loss $(n/d_{p,\ell})^c$ for some $c>0$.

4. Examples of time series satisfying **(WDP)** or **(CBS)**

We present several examples of time series satisfying **(WDP)** or **(CBS)**.

4.1. **Causal Bernoulli shifts.** Causal Bernoulli shifts are stationary time series that admit the representation

$$X_t = H(\xi_t, \xi_{t-1}, \xi_{t-2}, \dots) \qquad \forall t \in \mathbb{Z} \tag{4.1}$$

where the ξ_t are iid variables called innovations. Almost all known stationary and ergodic processes have this form. However we work here under the restrictive assumption (4.3). Remark that under this Lipschitz condition the existence of the stationary time series (X_t) follows from (4.1) and it satisfies the Cramer condition as soon as the innovations do. Some examples of causal Bernoulli shifts are presented below.

4.1.1. *Linear models.* Let (X_t) be a real time series admitting the MA(∞) representation

$$X_t = \sum_{j=0}^{\infty} a_j \xi_{t-j} \quad \text{with} \quad \sum_{j=0}^{\infty} j a_j < +\infty.$$

Then it satisfies **(CBS)** if the iid innovations ξ_t satisfy the Cramer condition. As an example there is any causal AR(∞) model $X_t = \phi_0 + \sum_{j=1}^{\infty} \phi_j X_{t-j} + \xi_t$ with $\phi(z) = 1 - \sum_{j=1}^{\infty} \phi_j z^j$ that have no root for $|z| \le 1$ (such that causal ARMA(p,q) models). Indeed, it is a real analytic function on the unit disc, $1/\phi(z)$ is a well defined real analytic function with the expression $\sum_{j=1}^{\infty} \psi_j z^j$ on the unit disc with the coefficients ψ_j that decrease exponentially fast ((3.2) is automatically satisfied).

4.1.2. *Chains with infinite memory.* Chains with infinite memory is a class of time series (X_t) introduced by Doukhan and Wintenberger [14] as the solution of the equation

$$X_t = F(X_{t-1}, X_{t-2}, \dots; \xi_t) \text{ almost everywhere,} \tag{4.2}$$

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

$$\big\|F(x;y) - F(x';y')\big\| \le \sum_{j=1}^{\infty} a_j(F)\|x_j - x'_j\| + u\|y - y'\|, \tag{4.3}$$

$$\text{with} \qquad \sum_{j=1}^{\infty} a_j(F) := a(F) < 1, \tag{4.4}$$

Many non linear econometrics time series are chains with infinite memory. The following Proposition gives sufficient assumptions such that chains with infinite memory satisfy **(CBS)**:

Proposition 4.1. *Under* (4.3) *and* (4.4) *there exists a unique solution* (X_t) *of equation* (4.2) *satisfying* **(CBS)** *if* ξ_0 *satisfies the Cramer condition.*

The proof of Proposition 4.1 is given in Subsection 6.5.

4.2. **Weakly dependent processes.**

4.2.1. *Bounded causal Bernoulli shifts.* Bounded causal Bernoulli shifts are examples of time series satisfying **(WDP)**:

Proposition 4.2. *Under condition* (4.3) *and* (3.2)*, any solution of the equation* (4.1) *is bounded by* $2a(H)\|\xi_0\|_\infty$ *and is weakly dependent* **(WDP)** *with* $\mathcal{C} = 2\|\xi_0\|_\infty \tilde{a}(H)$*.*

The proof of this already known result is given in Subsection 6.5 for completeness. Below are presented two examples of time series satisfying **(WDP)** that are not bounded causal Bernoulli shifts.

4.2.2. *Uniform φ-mixing processes.* Let us remind the definition of the φ-mixing coefficients introduced in Ibragimov [16];

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

where $\mathfrak{F}_r = \sigma(Y_t, t \geq r)$. The class of φ-mixing processes gives examples of time series that satisfied **(WDP)**:

Proposition 4.3. *If (X_t) is a stationary bounded process then it satisfies* **(WDP)** *with*

$$\theta_{\infty,n}(1) \leq 2\|X_0\|_\infty \sum_{r=1}^{n} \varphi(r).$$

The proof of this already known result is given in Subsection 6.5 for completeness. Remark that (X_t) satisfies the short memory condition as soon as $(\varphi(r))$ is summable. All uniform ergodic Markov chains are examples of φ-mixing processes with short memory, see Doukhan [13].

4.2.3. *Dynamical systems on* $[0,1]$. The AR(1) process $X_t = 2^{-1}(X_{t-1} + \xi_t)$ with ξ_t Bernoulli distributed is not mixing, see [3] for more details. Through a reversion of the time, it can be viewed as a dynamical system $X_t = T(X_{t+1})$ where $T(x) = 2x$ if $0 \leq x < 1/2$, $T(x) = 2x - 1$ if $1/2 \leq x \leq 1$. Dedecker and Prieur [12] extended this counter-example to processes (X_t) such that $X_t = T(X_{t+1})$ where T is an expanding map on $[0,1]$, see Section 4.4 of [12] for a proper definition. Then (X_t) satisfies **(WDP)** with $\mathcal{C} = K\sigma/(1-\sigma)$ where $K > 0$, $0 \leq \sigma < 1$, see Section 7.2 of [12].

5. Examples of predictors

We give some examples of Lipschitz predictors where we can estimate the complexity of the $\Theta_{p,\ell}$ and then apply our main results. In this section $C > 0$ is a constant independent of ε and n that may be different from one inequality to another.

5.1. **Linear predictors.** Let $\mathcal{X} = \mathbb{R}$ and we consider predictors of the form:

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

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

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

($m_p = 1$ for all p and we omit the index ℓ). Using Proposition 2.1 it follows that

$$d_p \leq (p+1)\log\left(eB\left(\frac{e}{p+1} \vee \frac{1}{B - \|\overline{\theta}_p\|}\right)\right),$$

where $\overline{\theta}_p = \arg\min_{\Theta_p} R(\theta)$. As a consequence of Theorems 3.1 and 3.2 we obtain:

Corollary 5.1. *If $\|\overline{\theta}_p\|_1 \leq B - e/(p+1)$ for all $p \geq 0$, then, under* **(WDP)** *or* **(CBS)**, *for all $n \geq 8e(1+L)$ with probability at least $1-\varepsilon$:*

$$R(\hat{\theta}) \leq \inf_{p+1 \leq n/2}\left\{\min_{\theta\in\Theta_p} R(\theta) + C\sqrt{\frac{p}{n}}\log^{5/2}(n)\right\} + C\frac{\log\frac{1}{\varepsilon}}{\sqrt{n}}.$$

As an application, consider the innovations ξ_t iid satisfying the Cramer condition and $\mathrm{med}(\xi_0) = 0$. If (X_t) is a causal AR(p_0) process ($0 \leq p_0 < \infty$) of the form

$$X_t = a_0 + \sum_{j=1}^{p_0} a_j X_{t-j} + \xi_j \text{ for all } t \in \mathbb{Z}.$$

If $B \geq \sum_{j=0}^{p} |a_j| + e/(p+1)$ for all $0 \leq p \leq p_0$, the error of the best linear predictor is $\mu[|\varepsilon_j|]$. Corollary 5.1 then implies that for any $0 < \varepsilon < 1$ and any $n \geq 2(p_0+1)$, it holds:

$$R(\hat{\theta}) - \mu[|\varepsilon_0|] \leq C\left(\sqrt{\frac{p_0}{n}}\log^{5/2}(n) + \frac{\log\frac{1}{\varepsilon}}{\sqrt{n}}\right) \qquad \text{with probability at least } 1-\varepsilon.$$

For $\varepsilon > 0$ fixed independently of n, the rate of convergence of the excess risk is estimated by $\sqrt{p_0/n}\log^{5/2}(n)$. Note that $\hat{\theta}$ achieves this rate even if p_0 is unknown. One says that our procedure is adaptive in p_0 and, using the terminology of [22], memory-universal.

Consider now that (X_t) comes from an AR(∞) model of the form

$$X_t = a_0 + \sum_{i=1}^{\infty} a_i X_{t-i} + \xi_t, \text{ for all } t \in \mathbb{Z}. \tag{5.1}$$

If $B \geq \sum_{j=0}^{p} |a_j| + e/(p+1)$ for all $p \geq 0$, we have $\overline{\theta}_p = (a_0, ..., a_p)$. Then we roughly bound $R(\overline{\theta}_p) = \pi_0[|\sum_{i>p} a_i X_{-i} + \xi_0|] \leq \mu[|\xi_0|] + \pi_0[|X_{-i}|]\sum_{i>p}|a_i|$ and with probability at least $1-\varepsilon$:

$$R(\hat{\theta}) - \mu[|\xi_0|] \leq \inf_{p+1\leq n/2}\left[\pi_0[|X_0|]\sum_{i>p}|a_i| + C\sqrt{\frac{p}{n}}\log^{5/2}(n)\right] + C\frac{\log\frac{1}{\varepsilon}}{\sqrt{n}}.$$

In this non-parametric setting, to obtain a rate of convergence for the excess risk we have to specify the decay rate of the $|a_i|$. For example, if

$$\exists\gamma > 0, \exists\beta > 0, \forall p : \quad \sum_{i>p}|a_i| \leq \frac{\gamma}{p^\beta}$$

then the convergence rate is $(\log^5(n)/n)^{\frac{\beta}{2\beta+1}}$ (consider the optimal $p = n^{1/(2\beta+1)}\log^{5/(2\beta+1)}(n)$).

Simulations. We implement our linear prediction procedure using the R software [23]. We compare the results to the one obtained using the standard ARIMA procedure of R with the AIC criterion for model selection. Our theoretical penalization terms, driven by «the worst-case type» bounds, are necessarily pessimistic: our procedure systematically over-penalizes large models. Thus, for having an efficient procedure in practice, adjustments have been done. However we aim with these simulations to show that

(1) our linear prediction procedure is easily implementable;
(2) its performances are reasonable when the implemented penalization term is smaller than the theoretical one.

We only consider observations from simulations of AR(p_0) models of the form

$$X_t = \sum_{i=1}^{p_0} a_i X_{t-i} + \xi_t$$

where the ξ_t are iid, either $\mathcal{N}(0,\sigma^2)$-distributed, either $(\delta_0 + \mathcal{E}(\lambda))/2$ distributed, where δ_0 is the Dirac mass on 0 and $\mathcal{E}(\lambda)$ the exponential distribution with parameter $\lambda > 0$. In both cases the Cramer condition is satisfied and $\mathrm{med}(\xi_0) = 0$. In the first case, mean and median are equal whereas this is no longer true in the second case. Thus the minimizers of the ℓ_1- and quadratic risk differ in the second case.

We use $p_0 = 3$, $a_1 = 0.2$, $a_2 = 0.3$, $a_3 = 0.2$, $\sigma^2 \in \{1, 3\}$, $\lambda \in \{1, 1/\sqrt{12}\}$, and $n = 500$,

$$\Theta = \bigcup_{p=1}^{8} \Theta_p = \bigcup_{p=1}^{8} \{\theta \in \mathbb{R}^p : \|\theta\|_1 \leq 1\}$$

and

$$\lambda \in \mathcal{G} = \{2, 4, 8, ..., 1024\}.$$

In view of our procedure, we compute the simplified penalized criterion

$$(\hat{\lambda}, \hat{p}) = \arg \min_{\substack{1 \leq p \leq 8 \\ \lambda \in \mathcal{G}}} -\frac{1}{\lambda} \log \int_{\Theta_{p,\ell}} \exp\left(-\lambda r_n(\theta)\right) d\pi_{p,\ell}(\theta) + \lambda \frac{K^2}{n}.$$

The theoretical value $K = 2(\log n)^{3/2} \approx 9$ systematically over-penalizes the large models and always selects the simplest one ($p = 1$). Thus, we fix in practice $K = 0,1$. To compute the criteria, the integrand term is approximated using an acceptation-reject algorithm with gaussian proposal and 10000 iterations. To compare one simulation of $\hat{\theta} \sim \pi_{\hat{p}}\{-\hat{\lambda} r_n\}$ with $\hat{\theta}_{AIC}$ obtained by the classical R procedure we simulate independently $(X'_1, \ldots, X'_{500})$ distributed as $(X_1, \ldots, X_{500})$ and we compare

$$\mathrm{err}_1(\hat{\theta}) = \frac{1}{n-8} \sum_{i=9}^{500} \left| X'_i - \sum_{p=1}^{\hat{p}} (\hat{\theta})_p X'_{i-p} \right|$$

with $\mathrm{err}_1(\hat{\theta}_{AIC})$. As the classical R procedure is based on least square estimators, we also compare the quadratic prevision error

$$\mathrm{err}_2(\hat{\theta}) = \frac{1}{n-8} \sum_{i=9}^{500} \left(X'_i - \sum_{p=1}^{\hat{p}} (\hat{\theta})_p X'_{i-p} \right)^2$$

with $\mathrm{err}_2(\hat{\theta}_{AIC})$. The results of 20 experiments are reported in Table 1.

The results are coherent with the theory: in the gaussian cases, the optimal values of θ for the ℓ_1 and the quadratic risks of prediction are the same. Both procedures estimate efficiently the

TABLE 1. For each experiment, we report the median, mean and standard deviation of the $\mathrm{err}_i(.)$ quantities on the 20 experiments realized. The best results, for both $\mathrm{err}_1(.)$ and $\mathrm{err}_2(.)$, are highlighted for each serie.

ξ_t		$\mathrm{err}_1(\hat{\theta})$	$\mathrm{err}_1(\hat{\theta}_{AIC})$	$\mathrm{err}_2(\hat{\theta})$	$\mathrm{err}_2(\hat{\theta}_{AIC})$
$\mathcal{N}(0,1)$	median	**0.790**	0.792	**0.975**	**0.975**
	mean	**0.797**	0.798	**0.985**	0.988
	s.d.	0.023	0.024	0.054	0.054
$\mathcal{N}(0,3)$	median	2.433	**2.432**	0.918	**0.916**
	mean	**2.409**	2.412	**0.911**	0.912
	s.d.	0.078	0.065	0.496	0.412
$\frac{\delta_0+\mathcal{E}(1)}{2}$	median	**0.567**	0.592	0.819	**0.813**
	mean	**0.580**	0.589	0.836	**0.813**
	s.d.	0.047	0.043	0.153	0.150
$\frac{\delta_0+\mathcal{E}(1/\sqrt{12})}{2}$	median	**1.973**	2.000	9.525	**9.494**
	mean	**1.955**	1.997	9.733	**9.390**
	s.d.	0.158	0.162	1.656	1.522

same θ and their prediction risks are the same. In the other cases, the optimal values of θ for the ℓ_1 and the quadratic risks of prediction are not the same. We observe $\mathrm{err}_1(\hat{\theta}) < \mathrm{err}_1(\hat{\theta}_{AIC})$ and $\mathrm{err}_2(\hat{\theta}) > \mathrm{err}_2(\hat{\theta}_{AIC})$. The choice between the two procedures only depends on the prediction risk considered.

5.2. **Neural networks predictors.** Similarly than in [22], we present a procedure that approximates the best possible predictor using the best possible number of past values p for the one-step prediction. Given p, the best possible predictor for the $\mathbb{L}^1$-risk is $\mathrm{med}(X_0|X_{-1},\dots,X_{-p})$. We denote R_p^* the corresponding risk. For $\mathcal{X}=\mathbb{R}$ we use the abstract neural networks predictors defined in Barron [6] by the relation

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

for $a_i\in\mathbb{R}^p$ and $c_i,b_i\in\mathbb{R}$ for all $1\le i\le \ell$, the sigmoidal function $\phi(x) = (1+\exp(-x))^{-1}$ for all $x\in\mathbb{R}$ and $\theta=(c_0,a_{1,1},\dots,a_{1,p},b_1,c_1,\dots,a_{\ell,1},\dots,a_{\ell,p},b_\ell,c_\ell)$ in $\mathcal{B}^q_{c_{p,\ell}}$ for some $c_{p,\ell}>0$, $q=\ell(p+2)+1$ and $\ell\le n$. For any $p\ge 1$ we denote

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

and we assume that there exists a complex-valued function $\tilde{r}_p$ on $\mathbb{R}^p$ satisfying

$$\forall x\in\mathbb{R}^p \qquad r_p(x)-r_p(0) = \int_{\mathbb{R}^p}(e^{iwx}-1)\tilde{r}_p(w)dw \quad\text{and}\quad \int_{\mathbb{R}^p}\|w\|_1|\tilde{r}_p(w)|dw \le C'p^c$$

for some $C',c>0$. Then

Corollary 5.2. *Under* **(WDP)** *if for any* $(p,\ell)\in M$

$$\frac{q}{e} + 2\sqrt{\ell}\|X\|_\infty(C'p^c + \ell\log\ell) \le c_{p,\ell} \tag{5.2}$$

then, for all $n \geq \max_M c_{p,\ell}$, with probability at least $1-\varepsilon$,

$$R(\hat{\theta}) \leq \inf_{10(1+\log n)^2 p^{1+2c} \leq n} \left\{ R_p^* + C \frac{p^{1/4+c/2}\log^3 n}{n^{1/4}} \right\} + C \frac{\log \frac{1}{\varepsilon}}{\sqrt{n}}.$$

If (X_t) satisfies the Markov condition of order p_0, then $c=0$ and for n sufficiently large

$$R(\hat{\theta}) - R_{p_0}^* \leq C \left(\frac{\log^3 n}{n^{1/4}} + \frac{\log \frac{1}{\varepsilon}}{\sqrt{n}} \right).$$

Compared to the iid case, the loss is $\log^3 n$ and we do not need to know the order p_0 (our procedure is memory-universal). It is smaller than the loss of the memory-universal procedure given in [22].

5.3. **Non-parametric auto-regressive predictors.** As in Baraud, Comte and Viennet [5], we assume that (X_t) is a solution of the equation:

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

where $\xi_t \sim \mathcal{N}(0,\sigma^2)$, the f_i are functions $[-1;1] \mapsto \mathbb{R}$ in Hölder class $H(s_i, L_i)$: f_i is derivable $\lfloor s_i \rfloor$ times and

$$\exists \mathcal{L}_i > 0, \forall (x,x') \in [-1,1]^2, \quad |f_i^{(\lfloor s_i \rfloor)}(x) - f_i^{(\lfloor s_i \rfloor)}(x')| \leq \mathcal{L}_i |x-x'|^{s_i - \lfloor s_i \rfloor}. \tag{5.3}$$

Consider the Fourier basis $(\phi_j(.))_{j\geq 1}$ on $[-1,1]$ composed by $\phi_{2k}(x) = \sqrt{2}\cos(2\pi k x)$ and $\phi_{2k+1}(x) = \sqrt{2}\sin(2\pi k x)$. Assumption 5.3 implies the existence of $\gamma_i > 0$ such that for any $m \geq 0$ it holds

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

Then natural predictors have the form

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

for any $p \in \{1,...,\lfloor n/2 \rfloor\}$ and any $\ell \in \{1,...,m_p = n\}$. We restrict the procedure on $\theta_{p,\ell}$ in the compact set

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

such that any f_θ is an L-Lipschitz function. Define also the coefficients $\overline{\theta}_{p,\ell} \in \mathbb{R}^{p\ell}$ by the relation

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

We obtain as a consequence of Theorem 3.1:

Corollary 5.3. *Under* **(CBS)**, *if for any $\ell \geq 1$ and any $p \geq 1$*

$$\frac{\ell p}{e} + \left(\sum_{i=1}^{p_0}\sum_{j=1}^{\ell} (\overline{\theta}_{p_0,\ell})_{i,j}^2 (2\lfloor j/2 \rfloor)^2 \right)^{\frac{1}{2}} \leq L,$$

then for all $n \geq 8e(1+L)$ *with probability at least* $1-\varepsilon$

$$R(\hat{\theta}) - \mu[|\xi_0|] \leq C\left(\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}} + \frac{\log\frac{1}{\varepsilon}}{\sqrt{n}}\right)$$

where s *denotes* $\min\{s_1,...,s_{p_0}\}$.

The (iid) minimax rate of convergence with respect to $s_1,...,s_{p_0}$ is for the ^{1}-risk achieved up to a logarithmic loss. In [5], the (iid) minimax rate of convergence for the quadratic risk is achieved for the empirical quadratic risk.

Aknowledgments

We would like to thank the anonymous referees for the various corrections and improvements they suggested.

6. Proofs

To present the proofs in a unified version whether we work under **(CBS)** or **(WDP)**, we truncate the observations if we are under **(CBS)**:

$$\overline{X}_t = H(\overline{\xi}_t, \overline{\xi}_{t-1}, \overline{\xi}_{t-2}, \ldots), \text{ for all } t \in \mathbb{Z},$$

where $\overline{\xi}_t = (\xi_t \wedge C) \vee (-C)$, under **(WDP)** we just take $\overline{X}_t = X_t$. We denote in the sequel $\overline{X} = (\overline{X}_t)$ and $\overline{r}$, $\overline{R}$ the risks associated with $\overline{X}$ under **(CBS)** and with X under **(WDP)**. For shorten the proofs, we denote $K_n = (1+L)\log^{3/2} n$ and $w_{p,\ell} = 1/(m_p\lfloor n/2 \rfloor)$ in the sequel. The proof of our main Theorem lies on estimates on Laplace transforms.

6.1. Preliminary lemmas: estimates on Laplace transforms.

The proofs of these lemmas are given in Section 6.4. The first Lemma is an estimate of the Laplace transforms of the risk of $\overline{X}$; it is a direct corollary of the result in Rio [24].

Lemma 6.1 (Laplace transform of the risk). *For any* $\lambda > 0$ *and* $\theta \in \Theta$ *we have:*

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

where $k_n = \sqrt{2}C(1+L)(a(H) + \tilde{a}(H))$ *under* **(CBS)** *and* $k_n = (1+L)(\|X_0\|_\infty + \theta_{\infty,n}(1))/\sqrt{2}$ *under* **(WDP)**.

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

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

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

Lemma 6.2 (Legendre transform of the Kullback divergence function)**.** *For any* $\pi \in \mathcal{M}_+^1(E)$*, for any measurable function* $h : E \to \mathbb{R}$ *such that* $\pi[\exp(h)] < +\infty$ *we have:*

$$\pi[\exp(h)] = \exp\left(\sup_{\rho\in\mathcal{M}_+^1(E)}\Big(\rho[h] - \mathcal{K}(\rho,\pi)\Big)\right), \tag{6.1}$$

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* (2.5)*.*

Using Lemmas 6.1 and 6.2 we get an upper-bound for the Laplace transform of the mean risk of Gibbs estimators in all sub-models:

Lemma 6.3. *Under the assumptions of Theorem 3.1 we have for any* $\lambda > 0$ *and* $(p, \ell) \in M$*:*

$$\pi_0\left[\exp\left(\sup_{\rho\in\mathcal{M}_+^1(\Theta_{p,\ell})}\left\{\lambda\rho[\overline{R} - \overline{r}_n] - \mathcal{K}\left(\rho, \pi_{p,\ell}\right)\right\} - \frac{\lambda^2 k_n^2}{n(1-p/n)^2}\right)\right] \le 1, \tag{6.2}$$

where k_n *has the same expression than in Lemma 6.1.*

Following the technique used by Catoni [7], we derive from Lemma 6.3 another upper-bound on the Laplace transform of the mean risk of any aggregation estimators of all Gibbs estimators:

Lemma 6.4. *For any measurable functions* $\hat{\rho}_{p,\ell} : \mathcal{X}^n \to \mathcal{M}_+^1(\Theta_{p,\ell})$ *for* $(p, \ell) \in M$*, under the assumptions of Theorem 3.1 we have:*

$$\pi_0\left[\sum_{(p,\ell)\in M}\sum_{\lambda\in\mathcal{G}_{p,\ell}}\hat{\rho}_{p,\ell}\left[\exp\left(\lambda(\overline{R} - \overline{r}_n) - \log\frac{d\hat{\rho}_{p,\ell}}{d\pi_{p,\ell}} - \frac{\lambda^2 k_n^2}{n(1-p/n)^2} + \log(w_{p,\ell}/n)\right)\right]\right] \le 1$$

and $$\pi_0\left[\sum_{(p,\ell)\in M}\sum_{\lambda\in\mathcal{G}_{p,\ell}}\exp\left(\lambda\hat{\rho}_{p,\ell}[\overline{r}_n - \overline{R}] - \mathcal{K}(\hat{\rho}_{p,\ell}, \pi_{p,\ell}) - \frac{\lambda^2 k_n^2}{n(1-p/n)^2} + \log(w_{p,\ell}/n)\right)\right] \le 1,$$

where we remind that k_n *is defined in Lemma 6.1.*

Finally, we use a Lemma that quantify the error in the risk due to the truncation under **(CBS)**:

Lemma 6.5. *Under* **(CBS)***, for any truncation level* $C > 0$ *and any* $0 \le \lambda \le n/(4(1 + L))$*, we have*

$$\pi_0\Big[\exp(\lambda\sup_{\theta\in\Theta}|r_n(\theta) - \overline{r}_n(\theta)| - \lambda 2(1+L)\Psi(a(H))\left(\frac{a(H)^2 C}{\exp(a(H)C) - 1} + \lambda\frac{4(1+L)}{n}\right)\Big] \le 1.$$

6.2. **Proof of Theorem 3.1.** Remark that **(WDP)** is satisfied, so $\overline{R} = R$ and $\overline{r} = r$. We apply the first inequality of Lemma 6.4 to $\hat{\rho}^{\lambda}_{p,\ell} = \pi_{p,\ell}\{-\lambda r_n\}$. Remembering that $(\hat{p}, \hat{\ell}, \hat{\lambda}) = \arg\min \hat{R}\,(p, \ell, \lambda)$, we obtain in particular:

$$\pi_0\hat{\rho}^{\hat{\lambda}}_{\hat{p},\hat{\ell}}\left[\exp\left(\hat{\lambda}(R - r_n) - \log\left(\frac{d\hat{\rho}^{\hat{\lambda}}_{\hat{p},\hat{\ell}}}{d\pi_{\hat{p},\hat{\ell}}}\right) - \frac{\hat{\lambda}^2 k_n^2}{n(1-\hat{p}/n)^2} + \log\left(\frac{w_{\hat{p},\hat{\ell}}}{n}\right)\right)\right] \le 1. \tag{6.3}$$

Remark that $\pi_0\,\hat\rho^{\hat\lambda}_{\hat p,\hat\ell}$ is a well defined probability measure as $\hat\rho$ are defined conditionally on the observations. Remark also that $\hat\theta\sim\hat\rho^{\hat\lambda}_{\hat p,\hat\ell}$ by definition, then using the classical Chernov bound we derive that with probability $1-\varepsilon$ it holds:

$$R(\hat\theta)\le r_n(\hat\theta)+\frac{\hat\lambda k_n^2}{n(1-\hat p/n)^2}+\frac{1}{\hat\lambda}\log\left(\frac{d\hat\rho^{\hat\lambda}_{\hat p,\hat\ell}}{d\pi_{\hat p,\hat\ell}}\right)+\frac{1}{\hat\lambda}\log\left(\frac{n}{w_{\hat p,\hat\ell}}\right)+\frac{1}{\hat\lambda}\log\frac{1}{\varepsilon}. \tag{6.4}$$

In order that the term $\hat R$ appears, we notice that (6.4) is equivalent to

$$\begin{aligned}R(\hat\theta)&\le-\frac{1}{\hat\lambda}\log\int_{\Theta_{\hat p,\hat\ell}}\exp\left(-\hat\lambda r_n(\theta)\right)\pi_{\hat p,\hat\ell}(d\theta)+\frac{\hat\lambda k_n^2}{n(1-\hat p/n)^2}+\frac{1}{\hat\lambda}\log\left(\frac{n}{w_{\hat p,\hat\ell}}\right)+\frac{1}{\hat\lambda}\log\frac{1}{\varepsilon}\\&\le\inf_{p,\ell,\lambda}\hat R\,(p,\ell,\lambda)+\frac{\hat\lambda(k_n^2-K_n^2)}{n(1-\hat p/n)^2}-\frac{1}{\hat\lambda}\log\varepsilon\end{aligned}$$

(remind that $K_n=(1+L)\log^{3/2}n$). Now, we upper bound the term $\hat R\,(p,\ell,\lambda)$, for any p, ℓ and λ. Using the second inequality of Lemma 6.4 we obtain for any $(p,\ell)\in M$, $\lambda\in\mathcal G$ and $\rho\in\mathcal M^1_+(\Theta_{p,\ell})$,

$$\int_{\Theta_{p,\ell}}r_n\,(\theta)\,\rho(d\theta)\le\int_{\Theta_{p,\ell}}R\,(\theta)\,\rho(d\theta)+\frac{\lambda k_n^2}{n(1-p/n)^2}+\frac{1}{\lambda}\mathcal K\,(\rho,\pi_{p,\ell})+\frac{1}{\lambda}\log\frac{n}{w_{p,\ell}}+\frac{1}{\lambda}\log\frac{1}{\varepsilon}. \tag{6.5}$$

From (6.5) and using Lemma 6.2 two times we derive that

$$\begin{aligned}-\frac{1}{\lambda}\log\int_{\Theta_{p,\ell}}\exp\left(-\lambda r_n(\theta)\right)\pi_{p,\ell}(d\theta)&=\inf_{\rho\in\mathcal M^1_+(\Theta_{p,\ell})}\left\{\int_{\Theta_{p,\ell}}r_n\,(\theta)\,\rho(d\theta)+\frac{1}{\lambda}\mathcal K\,(\rho,\pi_{p,\ell})\right\}\\&\le\inf_{\rho\in\mathcal M^1_+(\Theta_{p,\ell})}\left\{\int_{\Theta_{p,\ell}}R\,(\theta)\,\rho(d\theta)+\frac{2}{\lambda}\mathcal K\,(\rho,\pi_{p,\ell})\right\}+\frac{\lambda k_n^2}{n(1-p/n)^2}+\frac{1}{\lambda}\log\frac{n}{\varepsilon w_{p,\ell}}\\&=-\frac{2}{\lambda}\log\int_{\Theta_{p,\ell}}\exp\left(-\frac{\lambda}{2}R(\theta)\right)\pi_{p,\ell}(d\theta)+\frac{\lambda k_n^2}{n(1-p/n)^2}+\frac{1}{\lambda}\log\frac{n}{\varepsilon w_{p,\ell}}.\end{aligned}$$

Finally we obtain:

$$\hat R\,(p,\ell,\lambda)\le-\frac{2}{\lambda}\log\int_{\Theta_{p,\ell}}\exp\left(-\frac{\lambda}{2}R(\theta)\right)\pi_{p,\ell}(d\theta)+\frac{\lambda(k_n^2+K_n^2)}{n(1-p/n)^2}+\frac{1}{\lambda}\log\frac{n}{\varepsilon w_{p,\ell}}. \tag{6.6}$$

Under Assumption (2.3), as soon as $\lambda>2e$ it holds

$$-\log\pi_{p,\ell}\left[\exp\left(-\frac{\lambda}{2}\left(R-R(\overline\theta_{p,\ell})\right)\right)\right]\le d_{p,\ell}\log\frac{\lambda}{2}$$

and it easily follows that

$$-\log\pi_{p,\ell}\left[\exp\left(-\frac{\lambda}{2}\overline R\right)\right]\le d_{p,\ell}\log\frac{\lambda}{2}+\frac{\lambda}{2}R(\overline\theta_{p,\ell}).$$

We plug this result into the inequality (6.6) to obtain:

$$\hat R\,(p,\ell,\lambda)\le R\left(\overline\theta_{p,\ell}\right)+\frac{1}{\lambda}\left(2d_{p,\ell}\log\frac{\lambda}{2}+\log\frac{n}{\varepsilon w_{p,\ell}}\right)+\frac{\lambda(k_n^2+K_n^2)}{n(1-p/n)^2}. \tag{6.7}$$

Collecting the inequalities (6.4) and (6.7), we obtain:

$$R(\hat{\theta}) \le \inf_{p,\ell,\lambda\in\mathcal{G}_{p,\ell}} \left\{ R\left(\overline{\theta}_{p,\ell}\right) + \frac{1}{\lambda}\left(2d_{p,\ell}\log\frac{\lambda}{2} + \log\frac{n}{\varepsilon w_{p,\ell}}\right) + \frac{\lambda(k_n^2+K_n^2)}{n(1-p/n)^2}\right\} + \frac{\hat{\lambda}(k_n^2-K_n^2)}{n(1-\hat{p}/n)^2} - \frac{1}{\hat{\lambda}}\log\varepsilon. \tag{6.8}$$

As for $\lambda \in \mathcal{G}_{p,\ell}$, we have, by definition of $\mathcal{G}_{p,\ell}$ that

$$\lambda \in \left[\check{c}\frac{\sqrt{d_{p,\ell}n}\log(d_{p,\ell}n)}{K_n}, \dots, \hat{c}\frac{\sqrt{d_{p,\ell}n}\log(d_{p,\ell}n)}{K_n}\right] \cap [2e, n]$$

then it holds

$$R(\hat{\theta}) \le \inf_{d_{p,\ell}\le n}\left\{ R\left(\overline{\theta}_{p,\ell}\right) + \frac{K_n}{\check{c}\sqrt{d_{p,\ell}n}\log(d_{p,\ell}n)}\left(2d_{p,\ell}\log\frac{n}{2} + \log\frac{n}{\varepsilon w_{p,\ell}}\right) + 4\hat{c}(k_n^2+K_n^2)\sqrt{\frac{d_{p,\ell}}{n}}\frac{\log(nd_{p,\ell})}{K_n}\right\} + 4(k_n^2-K_n^2)_+ + \frac{(1+L)\log\frac{1}{\varepsilon}}{\check{c}\sqrt{n}}. \tag{6.9}$$

For the sake of simplicity, we use rough estimates ($1 \le d_{p,\ell}$, $1 \le 1/\varepsilon$, $m_p \le n$, $\dots$) to obtain

$$R(\hat{\theta}) \le \inf_{d_{p,\ell}\le n}\left\{ R\left(\overline{\theta}_{p,\ell}\right) + (1+L)\left(\frac{6}{\check{c}} + 8\hat{c}\Big(1+\|X_0\|_\infty + \theta_{\infty,n}(1)\Big)^2\right)\sqrt{\frac{d_{p,\ell}}{n}}\log^{5/2}(n)\right\} + 4(k_n^2-K_n^2)_+ + \frac{7(1+L)\log\frac{n}{\varepsilon}}{\check{c}\sqrt{n}}.$$

This ends the proof as

$$k_n^2 - K_n^2 = (1+L)\left(\frac{(\|X_0\|_\infty + \theta_{\infty,n}(1))^2}{2} - \log^3(n)\right).$$

6.3. **Proof of Theorem 3.2.** As we work under **(CBS)**, we have to deal with the error of approximation of r and R by $\overline{R}$. To quantify it, we use Lemma 6.5. First remark that as $R = \pi_0[r]$ it holds

$$\exp\left(\lambda\sup_{\theta\in\Theta}|R(\theta) - \overline{R}(\theta)| - \lambda\phi(C,\lambda)\right) \le 1,$$

where

$$\phi(C,\lambda) = 2(1+L)\Psi(a(H))\left(\frac{a(H)^2C}{\exp(a(H)C)-1} + \lambda\frac{4(1+L)}{n}\right)$$

An immediate consequence is that

$$\pi_0\left[\exp\left(\lambda\sup_{\theta\in\Theta}|(r_n-R)(\theta) - (\overline{r}_n - \overline{R})(\theta)| - 2\lambda\phi(C,\lambda)\right)\right] \le 1.$$

As $R - r_n = \overline{r}_n - \overline{R} + (r_n - R) - (\overline{r}_n - \overline{R})$, for any measurable function $\rho_{p,\ell} : \mathcal{X}^n \to \mathcal{M}_+^1(\Theta_{p,\ell})$ the Cauchy-Schwartz inequality gives

$$\pi_0\rho[\exp(\lambda/2(R-r_n))] \le \sqrt{\pi_0\rho[\exp(\lambda(\overline{R}-\overline{r}_n))]\pi_0\rho\left[\exp\left(\lambda\sup_{\theta\in\Theta}|(r_n-R)(\theta) - (\overline{r}_n-\overline{R})(\theta)|\right)\right]}.$$

Using this remark and the same reasoning than in the proof of Theorem 3.1 that gives (6.3) from Lemma 6.4, we get the inequality

$$\pi_0\hat{\rho}^{\hat{\lambda}}_{\hat{p},\hat{\ell}}\left[\exp\left(\frac{\hat{\lambda}}{2}(R-r_n)-0,5\log\left(\frac{d\hat{\rho}^{\hat{\lambda}}_{\hat{p},\hat{\ell}}}{d\pi_{\hat{p},\hat{\ell}}}\right)-0,5\frac{\hat{\lambda}^2k_n^2}{n(1-\hat{p}/n)^2}+0,5\log\left(\frac{w_{\hat{p},\hat{\ell}}}{n}\right)\right.\right.$$
$$\left.\left.-\lambda\phi(C,\lambda)\right)\right]\leq 1.$$

As in the proof of Theorem 3.1, we derive an equivalent of (6.4), i.e. with probability $1-\varepsilon$ it holds:

$$R(\hat{\theta})\leq r_n(\hat{\theta})+\frac{\hat{\lambda}k_n^2}{n(1-\hat{p}/n)^2}+\frac{1}{\hat{\lambda}}\log\left(\frac{d\hat{\rho}^{\hat{\lambda}}_{\hat{p},\hat{\ell}}}{d\pi_{\hat{p},\hat{\ell}}}\right)+\frac{1}{\hat{\lambda}}\log\left(\frac{n}{w_{\hat{p},\hat{\ell}}}\right)+2\phi(C,\hat{\lambda})+\frac{2}{\hat{\lambda}}\log\frac{1}{\varepsilon}.$$

With similar arguments we derive an equivalent of (6.5):

$$\int_{\Theta_{p,\ell}}r_n(\theta)\,\rho(d\theta)\leq\int_{\Theta_{p,\ell}}R(\theta)\,\rho(d\theta)+\frac{\lambda k_n^2}{n(1-p/n)^2}+\frac{1}{\lambda}\mathcal{K}(\rho,\pi_{p,\ell})+\frac{1}{\lambda}\log\frac{n}{w_{p,\ell}}$$
$$+2\phi(C,\lambda)+\frac{2}{\lambda}\log\frac{1}{\varepsilon}$$

and also

$$R(\hat{\theta})\leq\inf_{p,\ell,\lambda}\left\{R\left(\overline{\theta}_{p,\ell}\right)+\frac{1}{\lambda}\left(2d_{p,\ell}\log\frac{\lambda}{2}+\log\frac{n}{\varepsilon w_{p,\ell}}\right)+\frac{\lambda(k_n^2+K_n^2)}{n(1-p/n)^2}+2\phi(C,\lambda)\right\}$$
$$+\frac{\hat{\lambda}(k_n^2-K_n^2)}{n(1-p/n)^2}+2\phi(C,\hat{\lambda})-\frac{2}{\hat{\lambda}}\log\varepsilon. \tag{6.10}$$

We still have

$$-\frac{2}{\hat{\lambda}}\log\varepsilon\leq\frac{2(1+L)}{\check{c}\sqrt{n}}\log\frac{1}{\varepsilon}$$

so we now have to upper bound $2\phi(C,\hat{\lambda})$. As $\hat{\lambda}\leq n/(4(1+L))$ by definition of the $\mathcal{G}_{p,\ell}$, fixing $C=a(H)^{-1}\log n$ we obtain:

$$\phi(C,\hat{\lambda})\leq\frac{4a(H)(1+L)\Psi(a(H))[2\hat{\lambda}(1+L)+a(H)\log(n)]}{n}.$$

As $\hat{\lambda}\leq\hat{c}d_{\hat{p},\hat{\ell}}\log(d_{\hat{p},\hat{\ell}}n)/(1+L)$ by definition of $\mathcal{G}_{p,\ell}$, we obtain

$$\frac{\hat{\lambda}(k_n^2-K_n^2)}{n(1-p/n)^2}+2\phi(C,\hat{\lambda})\leq\frac{8a(H)^2(1+L)\Psi(a(H))\log(n)}{n}$$
$$+\sqrt{\frac{d_{\hat{p},\hat{\ell}}}{n}}\log(d_{\hat{p},\hat{\ell}}n)4(1+L)\hat{c}\left(4a(H)\Psi(a(H))+2\log^2(n)(1+\tilde{a}(H)/a(H))^2-\log^3(n)\right)_+.$$

We now plug this result into (6.10) to end the proof.

6.4. **Proofs of Lemmas 6.1, 6.3, 6.4, 6.5 and of Proposition 2.1.**

Proof of Lemma 6.1. The proof of this Lemma is based on the following result of Rio [24] on $\overline{X}$:

Theorem 6.6. *Let $Y = (Y_t)_{t\in\mathbb{Z}}$ be a bounded stationary time series bounded distributed as π_0 on $\mathcal{X}^{\mathbb{Z}}$. Let h be a 1-Lipschitz function of $\mathcal{X}^n \to \mathbb{R}$, i.e. such that:*

$$\forall (x_1, y_1, ..., x_n, y_n) \in \mathcal{X}^{2n}, \quad |h(x_1, ..., x_n) - h(y_1, ..., y_n)| \le \sum_{i=1}^{n} \|x_i - y_i\| . \tag{6.11}$$

Then for any $t \ge 0$ we have:

$$\pi_0 \left[\exp(t(\pi_0[h(X_1, ..., X_n)] - h(X_1, ..., X_n)))\right] \le \exp(t^2 n(\|X_0\|_\infty + \theta_{\infty,n}(1))^2/2).$$

Proof of Theorem 6.6. This version of Theorem 1 of [24] comes rewriting the inequality (3) in [24] as, for any 1-Lipschitz function g:

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

The result is proved as $\sup_{1\le r\le n} \theta_{\infty,r}(1) \le \theta_{\infty,n}(1)$. □

We now apply the result of Theorem 6.6 on $Y = \overline{X}$ to obtain the result of Lemma 6.1. 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:

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

We easily check that h satisfies condition (6.11):

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

The direct application of Theorem 6.6 ends the proof under **(WDP)**. Under **(CBS)** k_n follows from the estimates of $\|X_0\|_\infty$ and $\theta_{\infty,n}(1)$ obtained in Proposition 4.1. □

Proof of Lemma 6.3. Integrate the inequality in Lemma 6.1 with respect $\pi_{p,\ell}$ on $\Theta_{p,\ell}$ (then $p(\theta) = p$) for any $(p,\ell) \in M$ in order to obtain:

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

Fubini's Theorem implies that

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

Applying Lemma 6.2 for $\pi = \pi_{p,\ell}$ and $h = \lambda(\overline{R} - \overline{r}_n) - \lambda^2 k_n^2/(n(1-p/n)^2)$ on $\mathcal{M}_+^1(\Theta_{p,\ell})$ leads to the inequality:

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

This ends the proof. $\square$

Proof of Lemma 6.4. First, let us choose $\lambda \in \Lambda$. Let $h_{p,\ell}^{\lambda}$ denotes, for any $(p,\ell) \in M$:

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

From Lemma 6.3 applied on the different $\mathcal{M}_+^1(\Theta_{p,\ell})$ we have, for any $(p,\ell) \in M$:

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

Now we apply Inequality (6.1) in Lemma 6.2 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

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

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

$$\pi_0\left[\sup_{\sum_{(p,\ell)\in M} w'_{p,\ell}=1} \left\{\sum_{(p,\ell)\in M} w'_{p,\ell} \sup_{\rho_{p,\ell} \in \mathcal{M}_+^1(\Theta_{p,\ell})} \exp\left(\lambda\rho_{p,\ell}\left[\lambda(\overline{R} - \overline{r}_n) - \log\frac{d\rho_{p,\ell}}{d\pi_{p,\ell}}\right] - \frac{\lambda^2 k_n^2}{n(1-p/n)^2} + \log\frac{w_{p,\ell}}{w'_{p,\ell}}\right)\right\}\right] \leq 1. \tag{6.12}$$

By Jensen again, we obtain a bound for the first term in the sum bounded in Lemma 6.4:

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

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 (6.12) such that $w'_{p,\ell} = 1$ for $(p,\ell) = \arg\max_M\{h_{p,\ell}\}$:

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

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

Proof of Lemma 6.5. From the proof of the Lemma 6.1, we already know that $|\overline{r}_n(\theta) - r_n(\theta)| \le (1+L)/(n-p)\sum_{i=1}^n \|X_i - \overline{X}_i\|$. This bound holds uniformly on Θ. As $p \le n/2$ it remains to estimate $\pi_0[\exp(\lambda 2(1+L)/n\sum_{i=1}^n \|X_i - \overline{X}_i\|])$. From the assumption (4.3), the stationarity of X and as the ξ_is are iid we have:

$$\begin{aligned}\pi_0\Big[\exp\Big(\lambda 2(1+L)/n\sum_{i=1}^n\|X_i-\overline{X}_i\|\Big]\Big) &\le \pi_0\Big[\exp\Big(\lambda 2(1+L)/n\sum_{i=1}^n\sum_{j=0}^\infty a_j(H)\|\xi_{i-j}-\overline{\xi}_{i-j}\|\Big)\Big]\\ &\le \pi_0\Big[\exp\Big(\lambda 2(1+L)/n\sum_{j=0}^\infty\sum_{i=1\vee(n-j)}^n a_{n-i+j}(H)\|\xi_{n-j}-\overline{\xi}_{n-j}\|\Big)\Big]\\ &\le \prod_{j=0}^\infty \pi_0\Big[\exp(\lambda 2(1+L)/n\sum_{i=1\vee(n-j)}^n a_{n-i+j}(H)\|\xi_0\|\mathbb{1}_{\|\xi_0\|>C})\Big].\end{aligned}$$

Denoting $c_j = \lambda 2(1+L)\sum_{i=1\vee(n-j)}^n a_{n-i+j}(H)/n$, we develop for all $j \ge 0$

$$\pi_0\Big[\exp(c_j\|\xi_0\|\mathbb{1}_{\|\xi_0\|>C})\Big] = 1 + c_j\pi_0\Big[\|\xi_0\|\mathbb{1}_{\|\xi_0\|>C}\Big] + \sum_{k\ge 2}\frac{c_j^k\pi_0\Big[\|\xi_0\|^k\mathbb{1}_{\|\xi_0\|>C}\Big]}{k!}.$$

There exists $\delta > 0$ such that the complex function $\Psi(z) = \pi_0[\exp(z\|\xi_0\|)]$ is holomorphic on the open disk $D(0, a(H)+\delta)$. From the Cauchy estimates we obtain:

$$\pi_0\Big[\|\xi_0\|^k\mathbb{1}_{\|\xi_0\|>C}\Big] \le \pi_0\Big[\|\xi_0\|^k\Big] \le \Psi^{(k)}(0) \le \frac{k!\max_{c\in\overline{D(0,a(H))}}\Psi(\lambda)}{a(H)^k} \le \frac{k!\Psi(a(H))}{a(H)^k} \quad \forall k \ge 2$$

As $\lambda < n/(4(1+L))$ then $2c_j \le a(H)$ for all $j \ge 0$ and then we derive that for all $j \ge 0$:

$$\begin{aligned}\pi_0\Big[\exp(c_j\|\xi_0\|\mathbb{1}_{\|\xi_0\|>C})\Big] &\le 1 + c_j\pi_0\Big[\|\xi_0\|\mathbb{1}_{\|\xi_0\|>C}\Big] + \Psi(a(H))\sum_{k\ge 2}(c_j/a(H))^k\\ &\le 1 + c_j\pi_0\Big[\|\xi_0\|\mathbb{1}_{\|\xi_0\|>C}\Big] + \frac{\Psi(a(H))c_j^2}{a(H)(a(H)-c_j)} \le 1 + c_j\pi_0\Big[\|\xi_0\|\mathbb{1}_{\|\xi_0\|>C}\Big] + c_j^2\frac{2\Psi(a(H))}{a(H)^2}.\end{aligned}$$

As $\phi(x) = (\exp(x)-1)/x$ is an increasing function for $x > 0$, then $\mathbb{1}_{\|\xi_0\|>C} \le \phi(a(H)\|\xi_0\|)/\phi(a(H)C)$ and the Markov formula gives for all $j \ge 0$

$$\pi_0\Big[\exp(c_j\|\xi_0\|\mathbb{1}_{\|\xi_0\|>C})\Big] \le 1 + c_j\frac{\Psi(a(H))a(H)C}{\exp(a(H)C)-1} + c_j^2\frac{2\Psi(a(H))}{a(H)^2}.$$

Collecting those bounds we obtain

$$\pi_0\Big[\exp\Big(\lambda\sup_{\theta\in\Theta}|\overline{r}_n(\theta)-r_n(\theta)|\Big)\Big]\leq\prod_{j=0}^{\infty}\left(1+c_j\frac{\Psi(a(H))a(H)C}{\exp(a(H)C)-1}+c_j^2\frac{2\Psi(a(H))}{a(H)^2}\right).$$

Using that $\log(1+x)\leq x$ for all $x>0$, we finally obtain:

$$\log\left(\pi_0\Big[\exp\Big(\lambda\sup_{\theta\in\Theta}|\overline{r}_n(\theta)-r_n(\theta)|\Big)\Big]\right)\leq\sum_{j=0}^{\infty}c_j\frac{\Psi(a(H))a(H)C}{\exp(a(H)C)-1}+\sum_{j=0}^{\infty}c_j^2\frac{2\Psi(a(H))}{a(H)^2}.$$

The desired result follows from the estimates $\sum_{j=0}^\infty c_j\leq\lambda a(H)2(1+L)$ and $\sum_{j=0}^\infty c_j^2\leq\lambda^2a(H)^24(1+L)^2/n$. $\square$

Now give the proof of the useful Proposition 2.1.

Proof of Proposition 2.1. Let us introduce a parameter $\zeta>0$ then we have

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

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

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

So

$$\zeta\gamma-q\log\frac{\zeta}{Cc_{p,\ell}}\leq q\wedge\gamma C(c_{p,\ell}-\|\overline{\theta}_{p,\ell}\|)+q\log\left(\frac{Cc_{p,\ell}\gamma}{q}\vee\frac{c_{p,\ell}}{c_{p,\ell}-\|\overline{\theta}_{p,\ell}\|}\right).$$

Now if $q\leq\gamma C(c_{p,\ell}-\|\overline{\theta}_{p,\ell}\|)$ then we get the estimate $q(1+\log(Cc_{p,\ell}\gamma/q))/\log\gamma$ which decreases with γ. We then get the desired bound when the supremum is established for $\gamma=e\vee q/(C(c_{p,\ell}-\|\overline{\theta}_{p,\ell}\|))$. If $q\geq\gamma C(c_{p,\ell}-\|\overline{\theta}_{p,\ell}\|)$ then we get the estimate $(\gamma C(c_{p,\ell}-\|\overline{\theta}_{p,\ell}\|)+q\log(c_{p,\ell}/(c_{p,\ell}-\|\overline{\theta}_{p,\ell}\|)))/\log\gamma$ which increases with γ. We have to consider γ as large as possible, i.e. when $q=\gamma C(c_{p,\ell}-\|\overline{\theta}_{p,\ell}\|)$ and we are going back to the case treated above. $\square$

6.5. **Proofs of the results given in Section 4.** After proving Proposition 4.1, we give Lemma 6.7 that introduces a coupling argument used to estimate the coefficients $\theta_{\infty,n}(1)$ in Propositions 4.2 and 4.3.

Proof of Proposition 4.1. The theorem 3.1 of Doukhan and Wintenberger [14] gives the existence of a unique stationary solution and the existence of an H such that $X_t=H(\xi_t,\xi_{t-1},\xi_{t-2},\dots)$. We prove that conditions (3.1) and (3.2) are automatically satisfied. Let (x_i) and (y_i) be two sequences such that there exists $j\in\mathbb{N}$ with $x_i=y_i$ for all $i\neq j$. Then $H(x)=u_0^\infty$ where $u_0^\infty=\lim_{k\to\infty}u_0^k$ for $(u_{-i}^k)_{i\in\mathbb{N}}$ defined recursively by

$$u_{-i}^k=F(u_{-i-1}^k,u_{-i-2}^k,\dots,u_{1-k}^k,u_{-k}^k,0,\dots;x_i)\qquad\forall 0\leq i\leq k.$$

Similarly, we denote $H(y) = v_0^\infty$ such that $\|H(x) - H(y)\| = \|u_0^\infty - v_0^\infty\|$. For $j = 0$, using (4.3) $\|u_0^k - v_0^k\| \le u\|x_j - y_j\|$ for all k. For $j \ge 1$, as $x_i = y_i$ for $i > j$, for k sufficiently large it holds (with the convention $\sum_{\ell=1}^{-k} = 0$ for $k \ge 0$):

$$\|u_0^k - v_0^k\| \le \sum_{\ell_1=1}^{j} a_{\ell_1}(F) \sum_{\ell_2=1}^{j-\ell_1} a_{\ell_2}(F) \cdots \sum_{\ell_j=1}^{j-\ell_1-\cdots-\ell_{j-1}} a_{\ell_j}(F)\|u_{-j}^k - v_{-j}^k\|.$$

By definition $\|u_{-j}^k - v_{-j}^k \le u\|x_j - y_j\|$ and we obtain $\|u_0^k - v_0^k\| \le ua(F)^{j-1}\|x_j - y_j\|$ for sufficiently large k. As the estimate does not depends on k, we derive that (3.1) holds with $a_j(H) = ua(F)^{j-1}$ and that (3.2) follows from the condition (4.4). □

Now we state a useful coupling Lemma; (X_t^*) is said to be a coupling version of (X_t) if it is similarly distributed and such that $(X_t^*)_{t>0}$ is independent of $\mathfrak{S}_0 = \sigma(X_t, t \le 0)$. From a version of the Kantorovitch-Rubinstein duality, see Dedecker and Prieur [12] for more details, we obtain an estimate of $\theta_{\infty,n}(1)$:

Lemma 6.7. *For any version (X_t^*) we have*

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

For the sake of completeness, we give the proof of this already known result.

Proof of Lemma 6.7. As we equipped $\mathcal{X}^n$ with the norm $\|(x_1, \ldots, x_n)\| = \sum_{i=1}^n \|x_i\|$ we immediately get the inequality

$$\tau_{\infty,n}(1) \le \|\mathbb{E}(\|(X_1, \ldots, X_n) - (X_1^*, \ldots, X_n^*)\||\mathfrak{S}_0)\|_\infty \le \sum_{i=1}^{u} \|\mathbb{E}(\|X_i - X_i^*\||\mathfrak{S}_0)\|_\infty. \qquad □$$

The proof of Propositions 4.2 and 4.3 are simple applications of this Lemma:

Proof of Proposition 4.2. Let us consider the coupling version of the causal Bernoulli shift (X_t) given by

$$X_t^* = H(\xi_t, \xi_{t-1}, \ldots, \xi_1, \xi_0^*, \xi_{-1}^*, \ldots), \qquad \forall t \in \mathbb{Z}$$

where (ξ_t^*) is similarly distributed than (ξ_t) and the two processes are independent. Then from Lemma 6.7 and condition (3.1) we obtain:

$$\theta_{\infty,n}(1) \le \sum_{i=1}^{n} \Big\| \sum_{j=i}^{\infty} a_j(H)\mathbb{E}(\|\xi_{i-j} - \xi_{i-j}^*\|/\mathfrak{S}_0) \Big\|_\infty \le \sum_{j=i}^{\infty} ja_j(H)\|\mathbb{E}(\|\xi_{i-j} - \xi_{i-j}^*\|/\mathfrak{S}_0)\Big\|_\infty$$

and the desired result follows. □

Proof of Proposition 4.3. Here we will consider the maximal coupling scheme of Goldstein [15]: there exists a version (X_t^*) such that

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

As $\|Y - Z\| \le 2\|X_0\|_\infty \mathbb{1}_{Y \ne Z}$ for any variables Y, Z bounded by $\|X_0\|_\infty$, we have:

$$\|\mathbb{E}(\|X_i - X_i^*\|/\mathfrak{S}_0)\|_\infty \le 2\|X_0\|_\infty \|\mathbb{E}(\mathbb{1}_{X_i \ne X_i^*}/\mathfrak{S}_0)\|_\infty \le 2\|X_0\|_\infty \|\mathbb{P}(X_i \ne X_i^*/\mathfrak{S}_0)\|_\infty.$$

As $\mathbb{P}(X_i \ne X_i^*/\mathfrak{S}_0) \le \mathbb{P}(X_t \ne X_t^* \text{ for some } t \ge r/\mathfrak{S}_0)$, we conclude using Lemma 6.7. □

6.6. **Proofs of the results given in Section 5.** We proof the Corollaries 5.2 and 5.3 of Theorem 3.1 applied in the context of Neural Networks and projection in the Fourier basis predictors.

Proof of Corollary 5.2. Let us check that all the predictors are L-Lipschitz functions of the observations. For any $x, y \in \mathbb{R}^p$, as the function ϕ is 1-Lipschitz, we have

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

For $\theta \in \mathcal{B}^q_{c_{p,\ell}}$ then $L = (c_{p,\ell} \vee 1)^3$ is convenient. Next, using Jensen to estimate $\mathbb{L}_1$-risk by $\mathbb{L}_2$-risk, we obtain from the theorem 1 of Barron [6] the existence of $C > 0$ such that

$$\pi_0\left[\left|\mathrm{med}(X_0 \,|\, X_{-1}, \dots, X_{-p}) - f_{\overline{\theta}_{p,\ell}}(X_{-1}, \dots, X_{-p})\right|\right] \le C\frac{p^c\|X_0\|_\infty}{\sqrt{\ell}}$$

where $\overline{\theta}_{p,\ell}$ belongs to the compact set

$$\mathcal{B}'_{p,\ell} = \left\{\theta \in \mathbb{R}^{\ell(p+2)+1}; \sum_{i=1}^{\ell}|c_i| \le C'c^p; \max_{1\le i\le \ell}\|a_i\| \le \sqrt{\ell}\log \ell;\; \max_{1\le i\le\ell}|b_i| \le \|X_0\|_\infty\sqrt{\ell}\log\ell\right\}.$$

Remark that under the assumptions of Corollary 5.2 we have $c_{p,\ell} - \|\overline{\theta}_{p,\ell}\| \ge q/e$. It implies by Proposition 2.1 that $d_{p,\ell} \le 3q(1 + \log(c_{p,\ell}))$ when $c_{p,\ell} \ge 1$. From Theorem 3.1 there exists $C > 0$ satisfying

$$R(\hat{\theta})] \le \inf_{d_{p,\ell}\le n}\left\{R^*_p + C\left(\frac{p^c}{\sqrt{\ell}} + \log^3(n)\sqrt{\frac{p\ell}{n}}\right)\right\} + C\frac{\log\frac{1}{\varepsilon}}{\sqrt{n}}.$$

The result follows from considering $\ell = \sqrt{n}p^{c-1/2}$. □

Proof of Proposition 5.3. Let us apply Theorem 3.2: there exists $C > 0$ such that

$$R(\hat{\theta}) \le \inf_{p,\ell:d_{p,\ell}\le n}\left\{\min_{\theta\in\Theta_{p,\ell}} R(\theta) + C\sqrt{\frac{d_{p,\ell}}{n}}\log^{5/2}(n)\right\} + C\frac{\log\frac{1}{\varepsilon}}{\sqrt{n}}$$
$$\le \inf_{\ell:d_{p_0,\ell}\le n}\left\{\min_{\theta\in\Theta_{p_0,\ell}} R(\theta) + C\sqrt{\frac{d_{p_0,\ell}}{n}}\log^{5/2}(n)\right\} + C\frac{\log\frac{1}{\varepsilon}}{\sqrt{n}}.$$

Remarking that

$$R\left(\overline{\theta}_{p_0,\ell}\right) = \inf_{\theta\in\Theta}\pi_0\left[\left|X_{p+1} - f_{\overline{\theta}_{p_0,\ell}}(X_p, ..., X_1)\right|\right]$$
$$\le \pi_0\left[\left|X_{p+1} - \sum_{i=1}^{p_0} f_i(X_{p-i})\right|\right) + \inf_{\theta\in\Theta}\pi_0\left[\left|\sum_{i=1}^{p_0} f_i(X_{p-i}) - \sum_{i=1}^{p_0}\sum_{j=1}^{n}\theta_{i,j}\varphi_j(X_{p-i})\right|\right]$$
$$\le \mu[|\xi_0|] + \inf_{\theta\in\Theta}\sum_{i=1}^{p_0}\pi_0\left[\left|f_i(X_1) - \sum_{j=1}^{n}\theta_{i,j}\varphi_j(X_1)\right|\right].$$

Note also that under our hypothesis X_1 has a density upper bounded by $1/\sqrt{2\pi\sigma^2}$. It then holds

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

$$\leq \mu[|\xi_0|] + \frac{1}{\sqrt{2\pi\sigma^2}} \inf_{\theta\in\Theta} \sum_{i=1}^{p_0} \left(\int \left[f_i(x) - \sum_{j=1}^{n} \theta_{i,j}\varphi_j(x) \right]^2 dx \right)^{\frac{1}{2}}$$

$$\leq \mu[|\xi_0|] + \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{i=1}^{p_0} \gamma_i \ell^{-s_i} \leq \mu[|\xi_0|] + \frac{\sum_{i=1}^{p_0} \gamma_i}{\sqrt{2\pi\sigma^2}} \ell^{-s}.$$

Then we have

$$\pi_0[R(\hat{\theta})] \leq \mu[|\xi_0|] + \inf_{\ell} \left\{ \ell^{-s} \frac{\sum_{i=1}^{p_0} \gamma_i}{\sqrt{2\pi\sigma^2}} + C\sqrt{\frac{d_{p_0,\ell}}{n}} \log^{5/2}(n) \right\} + C\frac{\log\frac{1}{\varepsilon}}{\sqrt{n}}. \tag{6.13}$$

The estimate of $d_{p_0,\ell}$ from Propostition 2.1 is plugged into (6.13) to obtain for some $C > 0$

$$\pi_0[R(\hat{\theta})] \leq \mu[|\xi_0|] + \inf_{\ell} \left\{ \ell^{-s} \frac{\sum_{i=1}^{p_0} \gamma_i}{\sqrt{2\pi\sigma^2}} + C\sqrt{\frac{p_0\ell}{n}} \log^{5/2}(n) \right\} + C\frac{\log\frac{1}{\varepsilon}}{\sqrt{n}}.$$

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

References

[1] Akaike, H. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory (Tsahkadsor, 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 Aéatoires* (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] Dalalyan, A., and Tsybakov, A. Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *Machine Learning 72*, 1-2 (2008), 39–61.

[11] 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.

[12] Dedecker, J., and Prieur, C. New dependence coefficients: Examples and applications to statistics. *Probability Theory and Related Fields 132* (2005), 203–235.

[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] Lacour, C. Nonparametric estimation of the stationary density and the transition density of a markov chain. *Stoch. Process. and Appli. 118*, 2 (2008), 232–260.
[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] Modha, D. S., and Masry, E. Memory-universal prediction of stationary random processes. *IEEE transactions on information theory 44*, 1 (1998), 117–133.
[23] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
[24] Rio, E. Ingalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Académie des Sciences de Paris, Srie 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] Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics 6* (1978), 461–464.
[27] Shawe-Taylor, J., and Williamson, R. A pac analysis of a bayes estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT'97* (1997), ACM, pp. 2–9.
[28] Stoltz, G. *Information incomplète et regret interne en prédiction de suites individuelles*. PhD thesis, University Paris Sud, 2005.
[29] 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 Maréchal De Lattre De Tassigny 75775 PARIS CEDEX 16 FRANCE
E-mail address: wintenberger@ceremade.dauphine.fr