



HAL
open science

Time series prediction via aggregation: an oracle bound including numerical cost

Andres Sanchez-Perez

► **To cite this version:**

Andres Sanchez-Perez. Time series prediction via aggregation: an oracle bound including numerical cost. 2013. hal-00905418v2

HAL Id: hal-00905418

<https://hal.science/hal-00905418v2>

Preprint submitted on 18 Nov 2013 (v2), last revised 26 May 2014 (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.

TIME SERIES PREDICTION VIA AGGREGATION : AN ORACLE BOUND INCLUDING NUMERICAL COST

ANDRÉS SÁNCHEZ-PÉREZ

ABSTRACT. We study the problem of forecasting a time series for a Causal Bernoulli Shifts (CBS) model using a parametric family of predictors. The aggregation technique provides a forecaster of this parameter with well established and quite satisfying theoretical properties expressed in the form of an oracle inequality for the prediction risk. The main advantage of this result is that it does not require to specify a particular model on the data. The numerical computation of the aggregated predictor usually relies on a Markov chain Monte Carlo method whose performances should be evaluated. In particular, it is crucial to bound the number of simulations needed to achieve a numerical precision of the same order as the prediction error. In this direction we present a fairly general result which can be seen as an oracle inequality which includes the numerical cost of the predictor computation. Again it is not required to specify a particular model on the data. The numerical cost appears by letting the oracle inequality depend on the number of simulations required in the MCMC approximation. Using different priors, some numerical experiments are then carried out to support our findings.

1. INTRODUCTION

An aggregation method consists in building a new estimator or a new predictor from a collection of different ones (typically via an integration), which is nearly as good as the best among them, given a risk criterion (see [11]). The problem has been treated in different scenarios, with a few contributions in the dependent context, see [1] or [2], on which we shall rely in this work. The aggregated predictor is usually computed via a numerical procedure which raises an implementation issue. We will consider a widely used approach to deal with it, namely the Markov chain Monte Carlo method.

To evaluate the performance of this approach we proceed in two steps. First we establish an oracle inequality for the theoretical aggregated predictor in the general context of the Causal Bernoulli Shifts. We slightly revisit the results of [2] to derive an oracle bound for the prediction error of the theoretical aggregated predictor. Then we consider the practical predictor obtained by an MCMC approximation and derive an Oracle bound for it expressed with the number of simulations in the MCMC method. This is obtained using a result of Łatuszyński [9], [10], jointly with other properties of the basic MCMC algorithms that we use. Finally we treat the autoregressive process (with unknown order) as an illustrative example and we present some numerical results.

2. STATEMENT OF THE PROBLEM AND MAIN ASSUMPTIONS

Let us observe (X_1, \dots, X_n) from a stationary time series $X = (X_t)_{t \in \mathbb{Z}}$ valued in \mathbb{R}^r for some $r \geq 1$. In the following we denote by π_0 the probability (and the expectation associated to this probability) of the process $X = (X_t)_{t \in \mathbb{Z}}$.

Let \hat{X}_t be a given predictor, that is, a measurable function of the past of X ,

$\hat{X}_t = f((X_{t-i})_{i \geq 1})$. The prediction error is evaluated by

$$\tilde{R}(f) = \pi_0 \left[\ell(\hat{X}_t, X_t) \right] = \int_{\mathbb{R}^{\mathbb{Z}}} \ell(f((x_{t-i})_{i \geq 1}), x_t) \pi_0(d\mathbf{x}),$$

where ℓ be a loss function, which satisfies :

Assumption 1 (Lipschitz Loss). *For all $x, x' \in \mathbb{R}^r$,*

$$\ell(x, x') = g(x - x'),$$

for some convex function g which is non-negative, $g(0) = 0$ and K - Lipschitz :

$$|g(x) - g(y)| \leq K\|x - y\|.$$

Here and in the following, for $\mathbf{a} \in \mathbb{R}^d$, $\|\mathbf{a}\|$ denotes the Euclidean norm of \mathbf{a} ,

$$\|\mathbf{a}\| = \sqrt{\sum_{i=1}^d a_i^2}.$$

Consider a family of predictors $\{f_\theta, \theta \in \Theta\}$. For each $\theta \in \Theta$ there exists a unique $d = d(\theta) \in \mathbb{N}^*$ such that $f_\theta : (\mathbb{R}^r)^d \rightarrow (\mathbb{R}^r)$ is a function from which we define

$$\hat{X}_t^\theta = f_\theta(X_{t-1}, \dots, X_{t-d}),$$

as a possible predictor of X_t from its past.

A natural way to evaluate the performance of the predictor associated to θ is to compute the risk

$$R(\theta) = \tilde{R}(f_\theta) = \pi_0 \left[\ell(\hat{X}_t^\theta, X_t) \right].$$

The main goal of this work is to build a predictor function \hat{f}_n , possibly in the form $f_{\hat{\theta}_n}$, inferred from a sample (X_1, \dots, X_n) such that $\tilde{R}(\hat{f}_n)$ or $R(\hat{\theta}_n)$ is close to $\inf_{\theta \in \Theta} R(\theta)$ with π_0 -probability close to 1. The only assumptions that we shall suppose on the process X are the following :

Definition 1 (CBS). *A time series is defined as Causal Bernoulli Shifts (CBS) if it satisfies the representation*

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

where (ξ_s) is an i.i.d. sequence of $\mathbb{R}^{r'}$ -valued random variables called innovations, for some $r' \geq 1$ and $H : (\mathbb{R}^{r'})^{\mathbb{N}} \rightarrow \mathbb{R}^r$ is a function satisfying

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

for any $v = (v_j)_{j \in \mathbb{N}}, v' = (v'_j)_{j \in \mathbb{N}} \in \mathbb{R}^{r'}$, where $\sum_{j=0}^{\infty} j a_j(H) < +\infty$.

We denote $a(H) = \sum_{j=0}^{\infty} a_j(H)$, $\tilde{a}(H) = \sum_{j=0}^{\infty} j a_j(H)$.

Assumption 2. *For the CBS defined by $(\xi_s)_{s \in \mathbb{Z}}$ and H , the Laplace transform of ξ_0 at $a(H)$ is finite, i.e. $\Psi(a(H)) = \mathbb{E}[\exp(a(H)\|\xi_0\|)] < \infty$.*

Let $\bar{X}_t = H(\bar{\xi}_t, \bar{\xi}_{t-1}, \dots)$, for all t , where, for a fixed $C > 0$, $\bar{\xi}_t = (\xi_t \wedge C) \vee (-C)$. We note by $\bar{X} = \{\bar{X}_t\}$ and by \bar{r}_n and \bar{R} the risks associated to \bar{X} . This thresholding will be interesting because allows to the truncated CBS to enter in the class of weakly dependent processes. We just introduce a couple of definition before point out what we understand by weakly dependent process.

Definition 2. Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a bounded variable Z in \mathbb{R}^q defined on the probability space, we denote, for any sub σ - algebra \mathfrak{S} of \mathcal{A}

$$\theta_\infty(\mathfrak{S}, Z) = \sup_{f \in \Lambda_1^q} \left\| E[f(Z)|\mathfrak{S}] - E[f(Z)] \right\| ,$$

where

$$\Lambda_1^q = \left\{ f : \mathbb{R}^q \rightarrow \mathbb{R}, \frac{|f(z_1, \dots, z_q) - f(z'_1, \dots, z'_q)|}{\sum_{i=1}^q \|z_i - z'_i\|} \leq 1 \right\} .$$

Definition 3. We introduce the σ - algebra $\mathfrak{S}_p = \sigma(X_t, t \leq p)$ and define the $\theta_{\infty, n}(1)$ coefficients as

$$\theta_{\infty, k}(1) = \sup \left\{ \theta_\infty(\mathfrak{S}_p, (X_{j_1}, \dots, X_{j_l})), p+1 \leq j_1 < \dots < j_l, 1 \leq l \leq k \right\} .$$

Assumption 3 (Weak Dependence). There exist finite constants \mathcal{B}, \mathcal{C} such that almost surely,

$$\begin{aligned} \sup_{t \in \mathbb{Z}} \|Z_t\| &\leq \mathcal{B} , \\ \theta_{\infty, k}(1) &\leq \mathcal{C}, \forall k \in \mathbb{N} . \end{aligned}$$

Under Assumption 3, $(Z_t)_{t \in \mathbb{Z}}$ will be called weakly dependent process (WDP), see [7] or [8]. It is straightforward to see that the truncated CBS is a WDP.

The main assumptions on the family of predictors are the following ones.

Assumption 4 (Lipschitz predictor). Let $\theta \in \Theta$ and $d = d(\theta)$. There exist $b_1(\theta), \dots, b_d(\theta) \in \mathbb{R}_+$ such that for all $(x_1, \dots, x_d), (y_1, \dots, y_d) \in \mathbb{R}^{rd}$,

$$\|f_\theta(x_1, \dots, x_d) - f_\theta(y_1, \dots, y_d)\| \leq \sum_{j=1}^d b_j(\theta) \|x_j - y_j\| .$$

Denote $L = \sup_{\theta \in \Theta} \sum_{j=1}^{d(\theta)} b_j(\theta)$. We assume that $L \leq \log(n) - 1$.

Assumption 5 (Uniform θ - Lipschitz). Define $D_n = \sup_{\theta \in \Theta} d(\theta)$. We assume that

$D_n \leq \frac{n}{2}$ and there exists $\mathcal{D} < +\infty$ such that,

$$\pi_0 \left[\left\| f_{\tilde{\theta}}(X_{t-1}, \dots, X_{t-D_n}) - f_\theta(X_{t-1}, \dots, X_{t-D_n}) \right\| \right] \leq \mathcal{D} \sqrt{D_n} \|\tilde{\theta} - \theta\| , \quad \forall \theta, \tilde{\theta} \in \Theta .$$

3. PREDICTION VIA AGGREGATION

The predictor that we shall propose will be defined as an average of predictors f_θ based on the empirical version of the risk,

$$r_n(\theta; X_1, \dots, X_n) = \frac{1}{n-d(\theta)} \sum_{t=d(\theta)+1}^n \ell(\hat{X}_t^\theta, X_t).$$

For the sake of simplicity we will identify $r_n(\theta) \equiv r_n(\theta; X_1, \dots, X_n)$ but without forgetting that it is a random variable which depends on n observations of the series.

We consider a probability measure π over Θ is labelled as the prior. It will serve to control the complexity of predictors associated to Θ and to construct one in particular, as detailed in the following.

3.1. Gibbs predictor. For a measure ν and a measurable function h (called “energy function”) such that $\nu[\exp(h)] = \int \exp(h) d\nu < +\infty$, we denote by $\nu\{h\}$ the measure defined by

$$\nu\{h\}(d\theta) = \frac{\exp(h(\theta))}{\nu[\exp(h)]} \nu(d\theta).$$

It is a particular Gibbs measure where the inverse temperature is equal to -1 .

Definition 4 (Gibbs predictor). *Given a $\lambda > 0$, called the temperature parameter, we define the Gibbs predictor as the expectation of f_θ , where θ is drawn under $\pi\{-\lambda r_n\}$, that is*

$$(1) \quad \hat{f}_{\lambda,n} = \pi\{-\lambda r_n\}[f] = \int_{\Theta} \hat{f}_\theta \frac{\exp(-\lambda r_n(\theta))}{\pi[\exp(-\lambda r_n(\theta))]} \pi(d\theta).$$

So far we have presented a quite general framework : a time series that we aim to predict using a parameter θ , and a generic setting for aggregating in a set where “good” candidates of θ are supposed to lie. All the needed assumptions are listed above (Assumption 1 to 5). We will only require one additional assumption below on Θ and the prior π (see Assumption 6).

3.2. Theoretical oracle bounds on CBS. The proof of main result of this section is based on the same tools as those used by [2] up to a point (Lemma 3). For a sake of completeness we quote the essential lemmas.

The first one can be found in [6].

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

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

with convention $\infty - \infty = -\infty$. $\mathcal{M}_+^1(E)$ is the space of probability measures on E . Moreover, as soon as h is upper-bounded on the support of ν , the supremum with respect to ρ in the right-hand side is reached for the Gibbs measure $\nu\{h\}$.

\mathcal{K} stands for the Kullback-Leibler divergence.

$$\mathcal{K}(\rho, \nu) = \begin{cases} \int \log \frac{d\rho}{d\nu}(\theta) \rho(d\theta) & , \text{ if } \rho \ll \nu \\ +\infty & , \text{ otherwise} \end{cases}$$

A Hoeffding type inequality introduced in [15] leads to :

Lemma 2 (Laplace transform of the risk). *Under CBS and Assumptions 1, 2 and 4, for any truncation level $C > 0$, $\lambda > 0$ and $\theta \in \Theta$ we have,*

$$(2) \quad \pi_0 \left[\exp \left(\lambda \left(\bar{R}(\theta) - \bar{r}_n(\theta) \right) \right) \right] \leq \exp \left(\frac{4\lambda^2 k_n^2(C)}{n} \right),$$

and

$$(3) \quad \pi_0 \left[\exp \left(\lambda \left(\bar{r}_n(\theta) - \bar{R}(\theta) \right) \right) \right] \leq \exp \left(\frac{4\lambda^2 k_n^2(C)}{n} \right),$$

where $k_n(C) = \sqrt{2}CK(1+L)(a(H) + \tilde{a}(H))$.

The following lemma is quoted from [2].

Lemma 3. *Under CBS and Assumptions 1, 2 and 4, for any truncation level $C > 0$ and any $0 \leq \lambda \leq \frac{n}{4(1+L)}$, we have,*

$$\pi_0 \left[\exp \left(\lambda \sup_{\theta \in \Theta} |r_n(\theta) - \bar{r}_n(\theta)| - \lambda \phi(C, \lambda) \right) \right] \leq 1,$$

where

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

We have the following result on the aggregated predictor defined in (1).

Lemma 4. *Under CBS and Assumptions 1, 2 and 4, for any truncation level $C > 0$ and any $0 \leq \lambda \leq \frac{n}{4(1+L)}$, with probability at least $1 - \epsilon$,*

$$\tilde{R}(\hat{f}_{\lambda, n}) \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \rho[R] + \frac{2\mathcal{K}(\rho, \pi)}{\lambda} \right\} + \frac{16\lambda k_n^2(C)}{n} + \frac{2 \log \left(\frac{1}{2\epsilon} \right)}{\lambda} + 4\phi(C, 2\lambda).$$

See the Appendix for the proof.

We make an additional assumption on the prior π defined on Θ in order to obtain the main result of the section.

Assumption 6 (Balls of a minimizing sequence). *There exists a sequence $\{\mathbf{a}_n\}_{n \geq 1}$ and a constant C such that $\mathbf{a}_n \in \Theta$ (which depends on n),*

$$R(\mathbf{a}_n) \leq \inf_{\theta \in \Theta} R(\theta) + \frac{\log^4(n)}{\sqrt{n}},$$

$$\text{and } \pi[B(\mathbf{a}_n, \delta) \cap \Theta] \geq C\delta^{D_n}, \forall \delta \leq \delta_n^* = \frac{1}{\sqrt{n}},$$

where $B(\mathbf{a}_n, \delta)$ is the Euclidean ball centred at \mathbf{a}_n with radius δ .

Last assumption requires the prior to allocate a sufficiently large mass to low dimensional subsets of Θ . This is at the origin of the intuition of last condition. Imagine that the set Θ

can be expressed as $\Theta = \bigcup_{k=1}^{V(n)} \Theta_k$ with $\Theta_k \subset \mathbb{R}^{D_k}$ for all k and $\{D_k\}_{k \geq 1}$ an increasing sequence.

Here for simplicity we identify \mathbb{R}^{D_k} with the subspace $\left\{ \underbrace{(\mathbf{x}, 0, \dots, 0)}_{D_{V(n)} - D_k}, \mathbf{x} \in \mathbb{R}^{D_k} \right\}$ of $D^{V(n)}$.

Suppose that Θ_k is endowed with the prior probability measure π_k and that $\pi = \sum_{k=1}^{V(n)} c_k \pi_k$. Given $\mathbf{a} \in \Theta_k$ and such that its last $D_k - D_{k-1}$ coordinates are not all zero (it does not “belong” to Θ_{k-1}), we define a ball centred at \mathbf{a} with radius δ as

$$B(\mathbf{a}, \delta) = \bigcup_{j=k}^{V(n)} \left\{ \mathbf{u} \in \mathbb{R}^{D_j} : \|\mathbf{u} - \underbrace{(\mathbf{a}, 0, \dots, 0)}_{D_j - D_k}\|^T \leq \delta \right\},$$

and we set

$$\pi[B(\mathbf{a}_n, \delta) \cap \Theta] = \sum_{k=1}^{V(n)} c_k \mathbb{1}_{\Theta_k}(\mathbf{a}_n) \pi_k[B(\mathbf{a}_n, \delta) \cap \Theta_k].$$

Thanks to this decomposition, it will be possible to meet the condition of Assumption 6. See subsection 4.3 for a precise example.

Theorem 3.1. *In the context of CBS, if assumptions 1, 2, 4, 5 and 6 hold, with $D_n = O(\lfloor \log^3(n) \rfloor)$. Then there exists a constant \mathcal{E} , such that for all $\epsilon > 0$, with probability at least $1 - \epsilon$,*

$$\tilde{R}(\hat{f}_{\sqrt{n}, n}) \leq \inf_{\theta \in \Theta} \tilde{R}(\hat{f}_\theta) + \mathcal{E} \frac{\log^4(n)}{\sqrt{n}} + \frac{2}{\sqrt{n}} \log\left(\frac{1}{2\epsilon}\right).$$

The proof can be found in the Appendix.

Here however we shall focus on the fact that this inequality applies to a theoretical aggregated predictor $\hat{f}_{\sqrt{n}, n}$. One should indeed investigate how these predictors are computed in practice and how practical numerical approximations performs in comparison with the theoretical estimator.

4. COMPUTATION OF THE ESTIMATOR

We use the Metropolis - Hastings algorithm in order to compute the mean of a target probability whose density ρ , possibly unnormalised, is relatively easy to calculate. We will work over $\mathcal{X} \subseteq \mathbb{R}^r$ equipped with \mathcal{T} , the Borel σ - algebra. We will consider probability measures which are absolutely continuous, and have a known density with respect to the Lebesgue measure.

4.1. Metropolis - Hastings algorithm. The Metropolis-Hastings algorithm generates a Markov chain $\Phi = \{\Phi_i\}_{i \geq 0}$ with the target distribution as a unique invariant measure, based on another Markov chain which serves as a proposal (see [13] and [16]). We shall consider the two following classical setups for the proposal :

- The independent Hastings algorithm where the proposal is i.i.d. with density q such that for some $\beta > 0$

$$(4) \quad \inf_{y \in \mathcal{X}} \frac{q(y)}{\rho(y)} \geq \beta.$$

- The Metropolis-Hastings algorithm where the proposal is a Markov chain with conditional density kernel q on $\bar{\Theta} \times \bar{\Theta}$ such that

$$(5) \quad \beta = \inf_{(x,y) \in \bar{\mathcal{X}} \times \bar{\mathcal{X}}} \frac{\rho(y)}{\rho(x)} \inf_{(x,y) \in \bar{\mathcal{X}} \times \bar{\mathcal{X}}} q(x,y) > 0.$$

In both cases we can affirm that algorithm is uniformly ergodic (see [12]), i.e. for all $m \in \mathbb{N}$

$$\|P^m(x, \cdot) - \rho\| \leq (1 - \beta)^m.$$

4.2. Theoretical bounds for the computation. Theorem 4.1 from [10] allows to bound the amount of iterations needed by some ergodic Markov chains (included those generated by the MCMC method that we use) in order to control the error that we make in approximating the first moment of the stationary distribution by the empirical estimate obtained from the successive samples of the chain. Applying this result in our context (see the Appendix) we obtain the following result.

Corollary 1 (Confidence Estimation). *Let $\{\Phi_i\}_{i \geq 0}$ be the chain generated by the independence Hastings algorithm under hypothesis (4) or by the Metropolis-Hastings algorithm under (5). Denote by*

$$\begin{aligned} \bar{\Phi}_m &= \frac{1}{m} \sum_{i=0}^{m-1} \Phi_i, \\ \bar{\Phi} &= \rho[f]. \end{aligned}$$

Let $\alpha > 0$ and $0 < \epsilon < 1$ be arbitrary. For any $m \geq M(\alpha, \beta, \epsilon, X)$, with probability at least $1 - \epsilon$,

$$|\bar{\Phi}_m - \bar{\Phi}| \leq \alpha,$$

where :

$$\begin{aligned} M(\alpha, \beta, \epsilon, X) &= \frac{2(\text{diam}(X))^2}{\alpha^2 \beta \epsilon} + 2 \sqrt{\frac{(\text{diam}(X))^4}{\alpha^4 \beta^2 \epsilon^2} + \frac{(\text{diam}(X))^2}{\alpha^2 \beta^2 \epsilon}}, \\ \text{diam}(X) &= \sup_{x, y \in X} \|x - y\|. \end{aligned}$$

By setting α appropriately, this result says how many iterations of the MCMC method are required in order to be reach a precisions of the same order as the prediction error enjoyed by the target Gibbs predictor.

Theorem 4.1. *Under the hypothesis of Theorem 3.1, using a numerical method described by Corollary 1, we conclude that there exists a constant \mathcal{F} such that for all*

$$m \geq M\left(\frac{\log^3(n)}{\sqrt{n}}, \beta_{\sqrt{n}, n}, \epsilon, X\right), \text{ with probability at least } (1 - \epsilon)^2,$$

$$\tilde{R}(\bar{f}_{\sqrt{n}, n, m}) \leq \inf_{\theta \in \bar{\Theta}} \tilde{R}(f_\theta) + \mathcal{F} \frac{\log^4(n)}{\sqrt{n}} + \frac{2}{\sqrt{n}} \log\left(\frac{1}{\epsilon}\right).$$

We have noted by $\bar{f}_{\sqrt{n}, n, m}$ the MCMC approximation of $\hat{f}_{\sqrt{n}, n}$ after m iterations. In particular, and as specified also in Theorem 3.1, $\lambda = \sqrt{n}$. Remark that, when we target the distribution $\pi\{-\lambda r_n\}$ with a suitable MCMC method, the convergence rate depends on a β which is specific of that distribution, i.e., it is a function of λ and n . That is why parameter β has two sub-indexes: one corresponding to λ and the other to n .

Proof of Theorem 4.1

Considering that assumptions 1 and 5 hold we get

$$|R(\theta_2) - R(\theta_1)| \leq K\mathcal{D}\sqrt{D_{V(n)}}\|\theta_2 - \theta_1\|.$$

Our bounds for the convergence of MCMC algorithm are independent of the observations. In consequence, the probability of having the risk of the computed predictor as near as needed of the Gibbs predictor and also that the risk of Gibbs predictor be as near as needed of the infimum, is the product of both probabilities. Finally, using Corollary 1 and plugging last inequality in Theorem 3.1, we get the result. ■

4.3. The example of the autoregressive process. We study the autoregressive model of order p or simply the $\text{AR}(p)$, defined as the stationary solution of

$$X_t = \sum_{j=1}^p \theta_j X_{t-j} + \sigma \xi_t,$$

where the ξ_t are i.i.d. with $\mathbb{E}\xi_t = 0$.

We denote $s_d(\rho) = \left\{ (\theta_1, \dots, \theta_d) : 1 - \sum_{k=1}^d \theta_k z^k \neq 0 \text{ for } |z| < \rho^{-1} \right\}$ the set of θ s for which the autoregressive polynomial $\theta(z) = 1 - \sum_{k=1}^d \theta_k z^k$ has all its roots outside the circle of radius ρ^{-1} .

In this context, the CBS assumption implies that the true parameter $\bar{\theta} = (\theta_1, \dots, \theta_p) \in s_p(1)$ and the process is stable (see [5]).

Let $\Theta = \bigcup_{d=1}^{\lfloor \log(n) \rfloor} \Theta_d$. Suppose that $\Theta_d \subset \mathbb{R}^{D_d}$ for all d where a prior π_k is considered.

Regarding the predictors of the form

$$f_\theta(X_{t-1}, \dots, X_{t-\lfloor \log(n) \rfloor}) = \theta^T (X_{t-1}, \dots, X_{t-\lfloor \log(n) \rfloor})^T,$$

remark that, the Gibbs predictor can be expressed as $\hat{f}_{\lambda,n} = f_{\hat{\theta}_{\lambda,n}}$ where $\hat{\theta}_{\lambda,n}$ is the Gibbs estimator defined as

$$\hat{\theta}_{\lambda,n} = \pi\{-\lambda r_n\}[\text{Id}] = \int_{\Theta} \theta \frac{\exp(-\lambda r_n(\theta))}{\pi[\exp(-\lambda r_n(\theta))]} \pi(d\theta).$$

The MCMC conditions are easier verified on Θ , thus we develop in this subsection the estimation, but knowing that it does not change any previous result. All are straightforward applicable.

Without any information about the order of the process (i.e. p) it would be convenient to favor those $\theta \in \Theta_d$ with d small. Let

$$\pi(d\theta) = \sum_{d=1}^{\lfloor \log(n) \rfloor} c_d 1_{\Theta_d}(\theta) \pi_d(d\theta),$$

where $(c_1, \dots, c_{\lfloor \log(n) \rfloor})$ is the prior on the order and π_d is the prior on Θ_d , for $1 \leq d \leq \lfloor \log(n) \rfloor$.

In the following we moreover assume that :

- $r = 1$.
- The innovations $\{\xi_t\}$ have compact support and denote by \mathcal{B} a constant such that $X_t \in [-\mathcal{B}, \mathcal{B}]$ for all t . Assumption 2 is then satisfied.

- $\Theta_d \subset \mathbb{R}^d \simeq \mathbb{R}^d \times \{0\}^{[\log(n)]-d} \subset \mathbb{R}^{[\log(n)]}$. They are open and bounded by B .
 - ℓ is the square, thus assumption 1 holds because X_t and θ are bounded.
 - $\pi_d \propto 1_{\Theta_d}$.
 - $f_{\theta}(x_1, \dots, x_d(\theta)) = \sum_{i=1}^{d(\theta)} \theta_i x_i$.
- Assumption 5 holds because (Using Cauchy-Schwarz and Jensen's)

$$\begin{aligned} \pi_0 \left[\left\| f_{\tilde{\theta}}(X_{t-1}, \dots, X_{t-[\log(n)]}) - f_{\theta}(X_{t-1}, \dots, X_{t-[\log(n)]}) \right\| \right] &= \pi_0 \left[\left\| \sum_{i=1}^{[\log(n)]} (\tilde{\theta}_i - \theta_i) X_{[\log(n)]-i} \right\| \right] \\ &\leq \sqrt{\pi_0 \left[\sum_{i=1}^{[\log(n)]} X_i^2 \right]} \|\tilde{\theta} - \theta\| \\ &\leq \mathcal{B} \sqrt{[\log(n)]} \|\tilde{\theta} - \theta\|. \end{aligned}$$

Also assumption 6 holds because $\bar{\theta} = \arg \inf_{\theta \in \Theta} R(\theta)$, then we could take $\mathbf{a}_n = \bar{\theta}$, and for n big enough $B_p(\mathbf{a}_n, \delta) \subset \Theta_p$, and $\forall \delta \leq \delta_n^* = n^{-\frac{1}{2}}$ we have

$$\begin{aligned} \pi[B(\mathbf{a}_n, \delta) \cap \Theta] &\geq c_p \pi_p[B_p(\mathbf{a}_n, \delta)] \\ &= c_p \frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2} + 1\right)} \delta^p \\ &\geq C \delta^{[\log(n)]}, \end{aligned}$$

with $C = c_p \frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2} + 1\right)}$ and Γ the gamma function.

Since $B_d\left(\frac{1}{\sqrt{d}}\right) \subseteq s_d(1) \subseteq B_d(2^d - 1)$, (see [14]), the prior π could be defined on $\Theta = \bigcup_{d=1}^{[\log(n)]} \Theta_d$ with, for example, $\Theta_d = B_d\left(\frac{1}{\sqrt{d}}\right)$, $\Theta_d = s_d(1)$ or $B_d(2^d - 1)$. Clearly $\Theta_d = B_d\left(\frac{1}{\sqrt{d}}\right)$ is a more restrictive setting. If p is unknown and $\Theta_d = s_d(1)$ or $B_d(2^d - 1)$ we would face the problem of the size of set Θ because assumption 4 would not be verified. On the contrary, if $\Theta_d \subset B_d(B)$, for all $1 \leq d \leq [\log(n)]$, it is easy to see that assumption 4 holds.

With the aim of applying the oracle inequality given by Theorem 4.1 that applies to $\bar{\theta}_{\sqrt{n}, n, m}$, the numerical approximation of the estimator, we will see different priors combined with a proposal in the Metropolis-Hasting algorithm. As proposal chain we will use the uniform distribution over the entire Θ_d (independent of current state).

See that

$$\begin{aligned} \left| X_t - \sum_{j=1}^{d(\theta)} \theta_j X_{t-j} \right| &\leq |X_t| + \sum_{j=1}^d |\theta_j| |X_{t-j}| \leq \sqrt{d+1} \mathcal{B} \sqrt{1+B^2} \Rightarrow \\ r_n(z; X_1, \dots, X_n) &\leq \sqrt{d+1} \mathcal{B} \sqrt{1+B^2}, \end{aligned}$$

and then

$$\begin{aligned}
\frac{q(\boldsymbol{\theta})}{\rho(\boldsymbol{\theta})} &= \frac{\sum_{d=1}^{\lfloor \log(n) \rfloor} c_d \mathbf{1}_{\Theta_d}(\boldsymbol{\theta})}{\sum_{d=1}^{\lfloor \log(n) \rfloor} c_d \int_{\Theta_d} \mathbf{dz}} \frac{\sum_{d=1}^{\lfloor \log(n) \rfloor} c_d \int_{\Theta_d} \exp(-\lambda r_n(z; X_1, \dots, X_n)) \mathbf{dz}}{\sum_{d=1}^{\lfloor \log(n) \rfloor} c_d \mathbf{1}_{\Theta_d}(\boldsymbol{\theta}) \exp(-\lambda r_n(\boldsymbol{\theta}; X_1, \dots, X_n))} \\
&\geq \frac{\min_{1 \leq d \leq \lfloor \log(n) \rfloor} \{c_d\}}{\sum_{d=1}^{\lfloor \log(n) \rfloor} c_d V(\Theta_d)} \sum_{d=1}^{\lfloor \log(n) \rfloor} c_d \exp\left(-\lambda \sqrt{d+1} \mathcal{B} \sqrt{1+B^2}\right) V(\Theta_d) \\
&\geq \min_{1 \leq d \leq \lfloor \log(n) \rfloor} \{c_d\} n^{-2\lambda \mathcal{B} \sqrt{1+B^2}}.
\end{aligned}$$

This bound of β , maybe pessimistic, guaranties anyway that given $c_d > 0$ for all d , the independent Hastings algorithm converges and also allows to compute the number of iterations needed in order to reach the theoretical rate.

5. NUMERICAL WORK

Concretely two types of sets Θ_d were considered :

- $\Theta_d = B_d\left(\frac{1}{\sqrt{d}}\right)$

In order to generate uniform random vectors over the d - ball of radius R we use following algorithm from [17]:

- (1) Generate a random vector $\mathbf{Y} = (Y_1, \dots, Y_d)$ with i.i.d. $\mathcal{N}(0, 1)$ components
- (2) Generate $r = U^{\frac{1}{d}}$, with $U \sim \mathcal{U}(0, 1)$
- (3) Return $Z = Rr \frac{\mathbf{Y}}{\|\mathbf{Y}\|}$

- $\Theta_d = s_d(1)$.

In [4] it is described a method for sampling uniformly from $s_d(1)$ using Levinson-Durbin recursion algorithm. It was numerically improved by [3].

And for each one we run experiments as if :

- p would be known : $\Theta = \Theta_p$ and $c_p = 1$,
- or not : $\Theta = \bigcup_{d=1}^{\lfloor \log(n) \rfloor} \Theta_d$ and $c_d = \frac{e^{-d}}{\sum_{k=1}^{\lfloor \log(n) \rfloor} e^{-k}} \geq e^{-d} (e - 1)$.

We iterate the algorithm with $m = 1000$ times for the four schemes at $n \in \{64, 128, 256, 512, 1024, 2048, 4096\}$ with $p \in \{2, 4, 6, 8\}$. Twenty realisations of autoregressive processes were simulated for each order. The following graphs resume the behavior of the algorithm these time series in each case.

5.1. **Known order.**

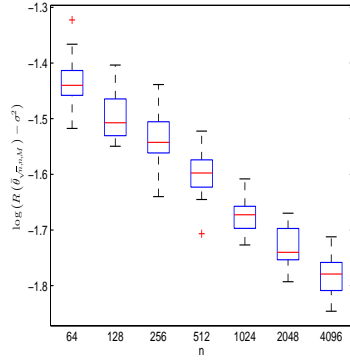


Figure 1: Prediction error. Uniform proposal, $p = 8$, $\Theta = s_8(1)$.

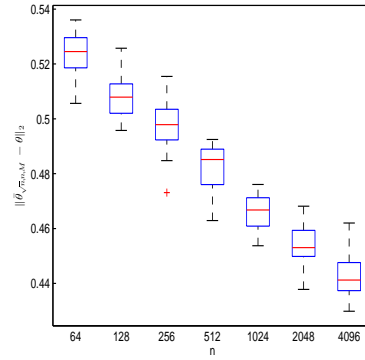


Figure 2: Estimation error. Uniform proposal, $p = 8$, $\Theta = s_8(1)$.

5.2. **Unknown order.**

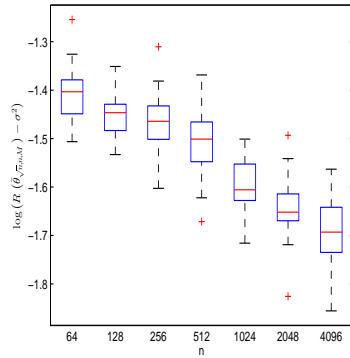


Figure 3: Prediction error. Uniform proposal, $p = 8$, $\Theta = \bigcup_{d=1}^{\lfloor \log(n) \rfloor} s_8(1)$.

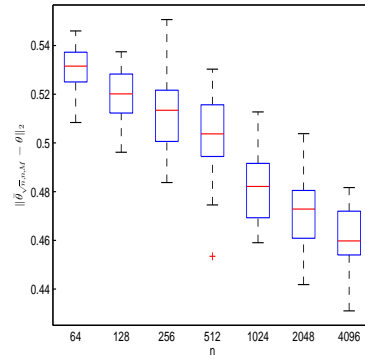


Figure 4: Estimation error. Uniform proposal, $p = 8$, $\Theta = \bigcup_{d=1}^{\lfloor \log(n) \rfloor} s_8(1)$.

6. CONCLUSION

The use of aggregated techniques determining a forecaster with almost minimal prediction risk has been considered in this work in the context of stationary time series. An approximation of the Gibbs predictor can be computed using the Metropolis Hastings algorithm. This allows us to obtain guaranties on the numerical approximation, that we expressed by a new oracle inequality. We have illustrated this approach through simulations in the case where AR predictor with weighted order are aggregated.

ACKNOWLEDGEMENTS

The author is thankful to François Roueff for fruitful discussion and his thoughtful review of this paper. This work has been partially supported by the Conseil régional d'Île-de-France under a doctoral allowance of its program Réseau de Recherche Doctoral en Mathématiques de l'Île de France (RDM-IdF) for the period 2012 - 2015 and by the Labex LMH (ANR-11-IDEX-003-02).

APPENDIX

Proof of Lemma 4

We use the relationship :

$$(6) \quad R - r_n = (\bar{R} - \bar{r}_n) + (R - r_n) - (\bar{R} - \bar{r}_n).$$

For any measure $\mu \in \mathcal{M}_+^1(\mathbb{R}^n \times \Theta)$, (6) and the Cauchy-Schwarz inequality lead to

$$(7) \quad \begin{aligned} \mu \left[\exp \left(\frac{\lambda}{2} (R - r_n) \right) \right] &= \mu \left[\exp \left(\frac{\lambda}{2} (\bar{R} - \bar{r}_n) \right) \exp \left(\frac{\lambda}{2} ((R - r_n) - (\bar{R} - \bar{r}_n)) \right) \right] \\ &\leq \sqrt{\mu \left[\exp(\lambda(\bar{R} - \bar{r}_n)) \right] \mu \left[\exp(\lambda((R - r_n) - (\bar{R} - \bar{r}_n))) \right]} \\ &\leq \sqrt{\mu \left[\exp(\lambda(\bar{R} - \bar{r}_n)) \right] \mu \left[\exp \left(\lambda \sup_{\theta \in \Theta} |(R - r_n)(\theta) - (\bar{R} - \bar{r}_n)(\theta)| \right) \right]}. \end{aligned}$$

Jensen's Inequality for the exponential function and Lemma 3 give that

$$\begin{aligned} \exp \left(\lambda \sup_{\theta \in \Theta} |R(\theta) - \bar{R}(\theta)| \right) &= \exp \left(\lambda \sup_{\theta \in \Theta} |\pi_0 [r_n(\theta) - \bar{r}_n(\theta)]| \right) \\ &\leq \pi_0 \left[\exp \left(\lambda \sup_{\theta \in \Theta} |r_n(\theta) - \bar{r}_n(\theta)| \right) \right] \\ &\leq \exp(\lambda\phi(C, \lambda)), \end{aligned}$$

and thanks to Lemma 3

$$(8) \quad \frac{\sqrt{\mu \left[\exp \left(\lambda \sup_{\theta \in \Theta} |(r_n - R)(\theta) - (\bar{r}_n - \bar{R})(\theta)| \right) \right]}}{\mu \left[\exp(\lambda\phi(C, \lambda)) \right]} \leq 1.$$

with $\mu = \pi_0 \otimes \pi$ and any $\pi \in \mathcal{M}_+^1(\Theta)$.

Lemma 2 implies that

$$(9) \quad \sqrt{\pi_0 \otimes \pi \left[\exp(\lambda(\bar{R} - \bar{r}_n)) \right]} \leq \exp \left(\frac{2\lambda^2 k_n^2(C)}{n} \right).$$

Multiplying (7), (8) and (9) with $\mu = \pi_0 \otimes \pi$ we obtain

$$\pi_0 \otimes \pi \left[\exp \left(\frac{\lambda}{2} (R - r_n) - \frac{2\lambda^2 k_n^2(C)}{n} - \lambda \phi(C, \lambda) \right) \right] \leq 1.$$

Changing λ by 2λ and thanks to Lemma 1 we get

$$\pi_0 \left[\exp \left(\sup_{\rho \in \mathcal{M}_+^1(\Theta)} (\lambda \rho [R - r_n] - \mathcal{K}(\rho, \pi)) - \frac{8\lambda^2 k_n^2(C)}{n} - 2\lambda \phi(C, 2\lambda) \right) \right] \leq 1.$$

Then, Markov's Inequality implies that for all $\epsilon > 0$,

$$\pi_0 \left(\sup_{\rho \in \mathcal{M}_+^1(\Theta)} (\lambda \rho [R - r_n] - \mathcal{K}(\rho, \pi_{p,\ell})) - \frac{8\lambda^2 k_n^2(C)}{n} - 2\lambda \phi(C, 2\lambda) - \log \left(\frac{1}{\epsilon} \right) \leq 0 \right) \geq 1 - \epsilon.$$

Hence with π_0 - probability at least $1 - \epsilon$, for all $\rho \in \mathcal{M}_+^1(\Theta)$

$$(10) \quad \rho [R - r_n] - \frac{1}{\lambda} \mathcal{K}(\rho, \pi) - \frac{8\lambda k_n^2(C)}{n} - 2\phi(C, 2\lambda) - \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \leq 0.$$

Setting $\rho = \pi\{-\lambda r_n\}$ and relying on Lemma 1, we have

$$\begin{aligned} \mathcal{K}(\pi\{-\lambda r_n\}, \pi) &= \pi\{-\lambda r_n\} \left[\log \frac{d\pi\{-\lambda r_n\}}{d\pi} \right] \\ &= \pi\{-\lambda r_n\} \left[\log \frac{\exp(-\lambda r_n)}{\pi[\exp(-\lambda r_n)]} \right] \\ &= \pi\{-\lambda r_n\} [-\lambda r_n] - \log(\pi[\exp(-\lambda r_n)]) \\ &= \pi\{-\lambda r_n\} [-\lambda r_n] + \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \{\rho[\lambda r_n] + \mathcal{K}(\rho, \pi)\} \end{aligned}$$

From (10) it follows that, with π_0 - probability at least $1 - \epsilon$,

$$\pi\{-\lambda r_n\}[R] \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \rho[r_n] + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\} + \frac{8\lambda k_n^2(C)}{n} + \frac{\log \left(\frac{1}{\epsilon} \right)}{\lambda} + 2\phi(C, 2\lambda).$$

We also have that with π_0 - probability at least $1 - \epsilon$,

$$\pi\{-\lambda r_n\}[r_n] \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \rho[R] + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\} + \frac{8\lambda k_n^2(C)}{n} + \frac{\log \left(\frac{1}{\epsilon} \right)}{\lambda} + 2\phi(C, 2\lambda).$$

Similarly to (10), but using (3) instead of (2), we obtain the same inequality with $\rho[R - r_n]$ replaced by $\rho[r_n - R]$ and hence, from a union bound, with π_0 - probability at least $1 - \epsilon$,

$$\pi\{-\lambda r_n\}[R] \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \rho[R] + \frac{2\mathcal{K}(\rho, \pi)}{\lambda} \right\} + \frac{16\lambda k_n^2(C)}{n} + \frac{2\log \left(\frac{1}{2\epsilon} \right)}{\lambda} + 4\phi(C, 2\lambda).$$

Using Tonelli's Theorem and Jensen's Inequality with the convex function g

$$\begin{aligned}
\pi\{-\lambda r_n\}[R] &= \int_{\Theta} \left[\int_{\mathbb{R}^Z} g(f_{\theta}(x_{n-1}, i \geq 0) - x_n) \pi_0(\mathbf{x}) \right] \pi\{-\lambda r_n\}(\mathbf{d}\theta) \\
&= \int_{\mathbb{R}^Z} \left[\int_{\Theta} g(f_{\theta}(x_{n-1}, i \geq 0) - x_n) \pi\{-\lambda r_n\}(\mathbf{d}\theta) \right] \pi_0(\mathbf{x}) \\
&\geq \int_{\mathbb{R}^Z} g \left(\int_{\Theta} (f_{\theta}(x_{n-1}, i \geq 0) - x_n) \pi\{-\lambda r_n\}(\mathbf{d}\theta) \right) \pi_0(\mathbf{x}) \\
&= \pi_0 \left[g(\hat{X}_{\lambda, n+1} - X_{n+1}) \right] \\
&= \tilde{R}(\hat{X}_{\lambda, n+1}).
\end{aligned}$$

This, with the previous bound, concludes the proof. ■

Proof of Theorem 3.1

We consider the set of probability measures $\{\rho_{a_n, \delta}, n \in \mathbb{N}, 0 \leq \delta \leq \delta^*\} \subset \mathcal{M}_+^1(\Theta)$, where $\rho_{a_n, \delta}(\tilde{\theta}) \propto \pi(\tilde{\theta}) 1_{\{\tilde{\theta} \in B(a_n, \delta) \cap \Theta\}}$.

The result above guarantees that

$$(11) \quad \tilde{R}(\hat{f}_n) \leq \inf_{0 \leq \delta \leq \delta^*} \left\{ \rho_{a_n, \delta}[R] + 2 \frac{\mathcal{K}(\rho_{a_n, \delta}, \pi)}{\lambda} \right\} + \frac{16\lambda k_n^2(C)}{n} + 2 \frac{\log\left(\frac{1}{2\epsilon}\right)}{\lambda} + 4\phi(C, 2\lambda).$$

$$(12) \quad \tilde{R}(\hat{f}_n) \leq \inf_{0 \leq \delta \leq \delta^*} \left\{ \rho_{a_n, \delta}[R] + 2 \frac{\mathcal{K}(\rho_{a_n, \delta}, \pi)}{\lambda} \right\} + \frac{16\lambda k_n^2(C)}{n} + 2 \frac{\log\left(\frac{1}{2\epsilon}\right)}{\lambda} + 4\phi(C, 2\lambda).$$

Thanks to assumptions 1 and 5, for any $n \in \mathbb{N}$ and $\tilde{\theta} \in B(a_n, \delta)$

$$R(\tilde{\theta}) - R(a_n) \leq K\pi_0 \left[\left\| f_{\tilde{\theta}}(X_{t-1}, \dots, X_{t-D_n(n)}) - f_{a_n}(X_{t-1}, \dots, X_{t-D_n}) \right\| \right] \leq K\mathcal{D} \sqrt{D_n} \delta.$$

Clearly

$$\mathcal{K}(\rho_{\theta, \delta}, \pi) = \log \left(\frac{1}{\pi[B(a_n, \delta) \cap \Theta]} \right) \leq -D_n \log(\delta) - \log(C)$$

Plugging these two expressions into (12)

$$\begin{aligned}
 \tilde{R}(\hat{f}_n) &\leq \inf_{0 \leq \delta \leq \delta_n^*} \left\{ \int_{\Theta} R(\theta) \rho_{a_n, \delta}(\mathrm{d}\theta) + \frac{16\lambda k_n^2(C)}{n} + 2 \frac{\mathcal{K}(\rho_{a_n, \delta}, \pi) + \log\left(\frac{1}{2\epsilon}\right)}{\lambda} + 4\phi(C, 2\lambda) \right\} \\
 &\leq \inf_{0 \leq \delta \leq \delta_n^*} \left\{ R(a_n) + \mathcal{E}_1 \sqrt{D_n} \delta + \frac{\mathcal{E}_2 \lambda (1+L)^2 C^2}{n} + \right. \\
 &\quad \left. - D_n \log(\delta) - \log(C) + \log\left(\frac{1}{2\epsilon}\right) + 2 \frac{\mathcal{E}_3 (1+L)C}{\lambda} + \frac{\mathcal{E}_4 (1+L)^2 \lambda}{\exp(a(H)C) - 1} \right\} \\
 &\leq \inf_{\theta \in \Theta} \tilde{R}(\hat{X}_{\sqrt{n}, n+1}^\theta) + \frac{\log^4(n)}{\sqrt{n}} + \frac{\mathcal{E}_2 \lambda (1+L)^2 C^2}{n} - \frac{2 \log(C)}{\lambda} + \frac{2 \log\left(\frac{1}{2\epsilon}\right)}{\lambda} + \frac{\mathcal{E}_3 (1+L)C}{\exp(a(H)C) - 1} + \\
 &\quad + \frac{\mathcal{E}_4 (1+L)^2 \lambda}{n} + \inf_{0 \leq \delta \leq \delta_n^*} \left\{ \mathcal{E}_1 \sqrt{D_n} \delta - \frac{2D_n \log(\delta)}{\lambda} \right\},
 \end{aligned}$$

where $\mathcal{E}_1 = K\mathcal{D}$, $\mathcal{E}_2 = 32K^2(a(H) + \tilde{a}(H))^2$, $\mathcal{E}_3 = 8K\Psi(a(H))a(H)$ and $\mathcal{E}_4 = 64K^2\Psi(a(H))$.

At a fixed ϵ , the rate of convergence of $\frac{2 \log\left(\frac{1}{2\epsilon}\right)}{\lambda} + \frac{\mathcal{E}_4 (1+L)^2 \lambda}{n}$ is at best $\frac{\log^2(n)}{\sqrt{n}}$, and we get it doing $\lambda = \sqrt{n}$. Regarding $\mathcal{E}_1 \sqrt{D_n} \delta - \frac{2D_n \log(\delta)}{\lambda}$, if we don't want to lose the rate $\frac{1}{\sqrt{n}}$ (up to a power of $\log(n)$) we should pick $D_n = O([\log^3(n)])$. Finally we do $\delta = \frac{1}{\sqrt{n}}$ and $C = \frac{\log(n)}{a(H)}$ and the result follows. ■

Proof of Corollary 1

Let us first introduce some additional notation. Given the functions $h : \mathcal{X} \rightarrow \mathbb{R}$ and $V : \mathcal{X} \rightarrow [1, \infty)$,

$$\begin{aligned}
 |h|_V &= \sup_{x \in \mathcal{X}} \frac{|h(x)|}{V(x)}, \\
 h_c(x) &= h(x) - \pi[h],
 \end{aligned}$$

and for any signed measure μ the V - norm is defined as

$$\|\mu\|_V = \sup_{|g| \leq V} \left| \int g(y) \mu(\mathrm{d}y) \right|.$$

Conditions of the cited theorem are satisfied under both circumstances because

- (1) \mathcal{X} is $(1, \beta, \pi)$ - small.
- (2) Foster-Lyapunov drift condition holds with $V(x) = 1, \forall x, K = 1$ and $\lambda \in [0, 1)$.
- (3) Strong aperiodicity follows from the fact that the whole set \mathcal{X} is small.

See that

$$\begin{aligned}
 \|P^m(x, \cdot) - \rho\|_V &= \sup_{|h| \leq 1} \left| \int h(y) P^m(x, \mathrm{d}y) - \int h(y) \rho(\mathrm{d}y) \right| \\
 &\leq (1 - \beta)^m.
 \end{aligned}$$

If the initial distribution is $\xi = \delta_x$ (i.e. we start always at $\Phi_0 = x$),

$$\begin{aligned}\delta_x[V] &= V(x) = 1, \\ \|\delta_x - \rho\|_V &= \sup_{|h| \leq 1} \left| \int h(y) \delta_x(dy) - \int h(y) \rho(dy) \right| = 1,\end{aligned}$$

$\Rightarrow \min\{\delta_x[V], \|\delta_x - \rho\|_V\} = 1$, then we can take $M = 1$ and $\gamma = 1 - \beta$ in the relation

$$\|\xi P^m - \rho\|_V \leq \min\{\xi[V], \|\xi - \rho\|_V\} M \gamma^m.$$

Taking also into account that

$$\|f_c\|_V^2 = \sup_{x \in \mathcal{X}} (x - \rho[f_c])^2 \leq (\text{diam}(\mathcal{X}))^2,$$

we can bound the quantities

$$\begin{aligned}b &= \frac{\rho[V] \|f_c\|_V^2}{\alpha^2 \epsilon} \left(1 + \frac{2M\gamma^{\frac{1}{2}}}{1 - \gamma^{\frac{1}{2}}} \right) \leq \frac{4(\text{diam}(\mathcal{X}))^2}{\alpha^2 \beta \epsilon}, \\ c &= \frac{M^2 \min\{\xi[V], \|\xi - \rho\|_V\} \|f_c\|_V^2}{\alpha^2 \epsilon (1 - \gamma)} \left(1 + \frac{2M\gamma^{\frac{1}{2}}}{1 - \gamma^{\frac{1}{2}}} \right) \leq \frac{4(\text{diam}(\mathcal{X}))^2}{\alpha^2 \beta^2 \epsilon}.\end{aligned}$$

Then $M(\alpha, \beta, \epsilon, \mathcal{X})$ corresponds to the upper bound of $\frac{b + \sqrt{b^2 + 4c}}{2}$.

■

REFERENCES

- [1] Pierre Alquier and Xiaoyin Li. Prediction of quantiles by statistical learning and application to gdp forecasting. In Jean-Gabriel Ganascia, Philippe Lenca, and Jean-Marc Petit, editors, *Discovery Science*, volume 7569 of *Lecture Notes in Computer Science*, pages 22–36. Springer Berlin Heidelberg, 2012.
- [2] Pierre Alquier and Olivier Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913, 2012.
- [3] Christophe Andrieu and Arnaud Doucet. An improved method for uniform simulation of stable minimum phase real arma (p,q) processes. *Signal Processing Letters, IEEE*, 6(6):142–144, june 1999.
- [4] Edward R. Beadle and Petar M. Djurić. Uniform random parameter generation of stable minimum-phase real arma (p,q) processes. *Signal Processing Letters, IEEE*, 4(9):259–261, september 1999.
- [5] Peter J. Brockwell and Richard A. Davis. *Time series: theory and methods*. Springer Series in Statistics. Springer, New York, 2006. Reprint of the second (1991) edition.
- [6] Olivier Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.
- [7] Jérôme Dedecker, Paul Doukhan, Gabriel Lang, José Rafael León R., Sana Louhichi, and Clémentine Prieur. *Weak dependence: with examples and applications*, volume 190 of *Lecture Notes in Statistics*. Springer, New York, 2007.
- [8] Jérôme Dedecker and Clémentine Prieur. New dependence coefficients. Examples and applications to statistics. *Probab. Theory Related Fields*, 132(2):203–236, 2005.
- [9] Krzysztof Łatuszyński, Blazej Miasojedow, and Wojciech Niemiro. Nonasymptotic bounds on the estimation error of mcmc algorithms. *Bernoulli*, 2013.

- [10] Krzysztof Łatuszyński and Wojciech Niemirow. Rigorous confidence bounds for MCMC under a geometric drift condition. *J. Complexity*, 27(1):23–38, 2011.
- [11] Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.
- [12] K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24(1):101–121, 1996.
- [13] Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. With a prologue by Peter W. Glynn.
- [14] Eric Moulines, Pierre Priouret, and François Roueff. On recursive estimation for time varying autoregressive processes. *Ann. Statist.*, 33(6):2610–2654, 2005.
- [15] Emmanuel Rio. Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(10):905–908, 2000.
- [16] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probab. Surv.*, 1:20–71, 2004.
- [17] Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo method*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2008.

INSTITUT MINES-TÉLÉCOM; TÉLÉCOM PARISTECH; CNRS LTCI, TÉLÉCOM PARISTECH, 37 RUE DAREAU, 75014 PARIS, FRANCE

E-mail address: andres.sanchez-perez@telecom-paristech.fr