Time Series Prediction via Aggregation: An Oracle Bound Including Numerical Cost

Andres Sanchez-Perez

Abstract We address the problem of forecasting a time series meeting the Causal Bernoulli Shift model, using a parametric set of predictors. The aggregation technique provides a predictor with well established and quite satisfying theoretical properties expressed by an oracle inequality for the prediction risk. The numerical computation of the aggregated predictor usually relies on a Markov chain Monte Carlo method whose convergence 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 risk. In this direction we present a fairly general result which can be seen as an oracle inequality including the numerical cost of the predictor computation. The numerical cost appears by letting the oracle inequality depend on the number of simulations required in the Monte Carlo approximation. Some numerical experiments are then carried out to support our findings.

1 Introduction

The objective of our work is to forecast a stationary time series $Y = (Y_t)_{t\in\mathbb{Z}}$ taking values in $\mathcal{X} \subseteq \mathbb{R}^r$ with $r \geq 1$. For this purpose we propose and study an aggregation scheme using exponential weights.

Consider a set of individual predictors giving their predictions at each moment t. An aggregation method consists of building a new prediction from this set, which is nearly as good as the best among the individual ones, provided a risk criterion (see [17]). This kind of result is established by oracle inequalities. The power and the beauty of the technique lie in its simplicity and versatility. The more basic and general context of application is individual sequences, where no assumption on the observations is made (see [9] for a comprehensive overview). Nevertheless, results need to be adapted if we set a stochastic model on the observations.

A. Sanchez-Perez (✉)
Institut Mines-Télécom; Télécom ParisTech; CNRS LTCI Télécom ParisTech, 37 rue Dareau, 75014 Paris, France
e-mail: andres.sanchez-perez@telecom-paristech.fr

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_13

The use of exponential weighting in aggregation and its links with the PAC-Bayesian approach has been investigated for example in [5, 8] and [11]. Dependent processes have not received much attention from this viewpoint, except in [1] and [2]. In the present paper we study the properties of the Gibbs predictor, applied to Causal Bernoulli Shifts (CBS). CBS are an example of dependent processes (see [12] and [13]).

Our predictor is expressed as an integral since the set from which we do the aggregation is in general not finite. Large dimension is a trending setup and the computation of this integral is a major issue. We use classical Markov chain Monte Carlo (MCMC) methods to approximate it. Results from Łatuszyński [15, 16] control the number of MCMC iterations to obtain precise bounds for the approximation of the integral. These bounds are in expectation and probability with respect to the distribution of the underlying Markov chain.

In this contribution we first slightly revisit certain lemmas presented in [2, 8] and [20] to derive an oracle bound for the prediction risk of the Gibbs predictor. We stress that the inequality controls the convergence rate of the exact predictor. Our second goal is to investigate the impact of the approximation of the predictor on the convergence guarantees described for its exact version. Combining the PAC-Bayesian bounds with the MCMC control, we then provide an oracle inequality that applies to the MCMC approximation of the predictor, which is actually used in practice.

The paper is organised as follows: we introduce a motivating example and several definitions and assumptions in Sect. 2. In Sect. 3 we describe the methodology of aggregation and provide the oracle inequality for the exact Gibbs predictor. The stochastic approximation is studied in Sect. 4. We state a general proposition independent of the model for the Gibbs predictor. Next, we apply it to the more particular framework delineated in our paper. A concrete case study is analysed in Sect. 5, including some numerical work. A brief discussion follows in Sect. 6. The proofs of most of the results are deferred to Sect. 7.

Throughout the paper, for $\boldsymbol{a} \in \mathbb{R}^q$ with $q \in \mathbb{N}^*$, $\|\boldsymbol{a}\|$ denotes its Euclidean norm, $\|\boldsymbol{a}\| = (\sum_{i=1}^q a_i^2)^{1/2}$ and $\|\boldsymbol{a}\|_1$ its 1-norm $\|\boldsymbol{a}\|_1 = \sum_{i=1}^q |a_i|$. We denote, for $\boldsymbol{a} \in \mathbb{R}^q$ and $\Delta > 0$, $B(\boldsymbol{a}, \Delta) = \{\boldsymbol{a}_1 \in \mathbb{R}^q : \|\boldsymbol{a} - \boldsymbol{a}_1\| \leq \Delta\}$ and $B_1(\boldsymbol{a}, \Delta) = \{\boldsymbol{a}_1 \in \mathbb{R}^q : \|\boldsymbol{a} - \boldsymbol{a}_1\|_1 \leq \Delta\}$ the corresponding balls centered at $\boldsymbol{a}$ of radius $\Delta > 0$. In general bold characters represent column vectors and normal characters their components; for example $\boldsymbol{y} = (y_i)_{i\in\mathbb{Z}}$. The use of subscripts with ':' refers to certain vector components $\boldsymbol{y}_{1:k} = (y_i)_{1\leq i\leq k}$, or elements of a sequence $X_{1:k} = (X_t)_{1\leq t\leq k}$. For a random variable U distributed as ν and a measurable function h, $\nu[h(U)]$ or simply $\nu[h]$ stands for the expectation of $h(U)$: $\nu[h] = \int h(u)\nu(\mathrm{d}u)$.

2 Problem Statement and Main Assumptions

Real stable autoregressive processes of a fixed order, referred to as AR(d) processes, are one of the simplest examples of CBS. They are defined as the stationary solution of

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

where the $(\xi_t)_{t\in\mathbb{Z}}$ are i.i.d. real random variables with $\mathbb{E}[\xi_t] = 0$ and $\mathbb{E}[\xi_t^2] = 1$.

We dispose of several efficient estimates for the parameter $\boldsymbol{\theta} = \begin{bmatrix}\theta_1 \ldots \theta_d\end{bmatrix}'$ which can be calculated via simple algorithms as Levinson-Durbin or Burg algorithm for example. From them we derive also efficient predictors. However, as the model is simple to handle, we use it to progressively introduce our general setup.

Denote

$$A(\boldsymbol{\theta}) = \begin{bmatrix} \theta_1 & \theta_2 & \ldots & \ldots & \theta_d \\ 1 & 0 & \ldots & \ldots & 0 \\ 0 & 1 & 0 & \ddots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 & 0 \end{bmatrix} ,$$

$\boldsymbol{X}_{t-1} = \begin{bmatrix}X_{t-1} \ldots X_{t-d}\end{bmatrix}'$ and $\boldsymbol{e}_1 = \begin{bmatrix}1\ 0 \ldots 0\end{bmatrix}'$ the first canonical vector of $\mathbb{R}^d$. M' represents the transpose of matrix M (including vectors). The recurrence (1) gives

$$X_t = \boldsymbol{\theta}'\boldsymbol{X}_{t-1} + \sigma\xi_t = \sigma \sum_{j=0}^{\infty} \boldsymbol{e}_1' A^j(\boldsymbol{\theta})\, \boldsymbol{e}_1 \xi_{t-j} \ . \tag{2}$$

The eigenvalues of $A(\boldsymbol{\theta})$ are the inverses of the roots of the autoregressive polynomial $\boldsymbol{\theta}(z) = 1 - \sum_{k=1}^{d} \theta_k z^k$, then at most δ for some $\delta \in (0, 1)$ due to the stability of X (see [7]). In other words $\boldsymbol{\theta} \in s_d(\delta) = \{\boldsymbol{\theta} \ : \ \boldsymbol{\theta}(z) \neq 0 \text{ for } |z| < \delta^{-1}\} \subseteq s_d(1)$. In this context (or even in a more general one, see [14]) for all $\delta_1 \in (\delta, 1)$ there is a constant $\bar{K}$ depending only on $\boldsymbol{\theta}$ and δ_1 such that for all $j \geq 0$

$$\left|\boldsymbol{e}_1' A^j(\boldsymbol{\theta})\, \boldsymbol{e}_1\right| \leq \bar{K}\delta_1^j \ , \tag{3}$$

and then, the variance of X_t, denoted γ_0, satisfies $\gamma_0 = \sigma^2 \sum_{j=0}^{\infty} |\boldsymbol{e}_1' A^j(\boldsymbol{\theta})\, \boldsymbol{e}_1|^2 \leq \bar{K}^2\sigma^2/(1-\delta_1^2)$.

The following definition allows to introduce the process which interests us.

Definition 1 Let $\mathcal{X}' \subseteq \mathbb{R}^{r'}$ for some $r' \geq 1$ and let $A = (A_j)_{j\geq 0}$ be a sequence of non-negative numbers. A function $H : (\mathcal{X}')^{\mathbb{N}} \to \mathcal{X}$ is said to be A-Lipschitz if

$$\|H(\boldsymbol{u}) - H(\boldsymbol{v})\| \leq \sum_{j=0}^{\infty} A_j \|u_j - v_j\| \ ,$$

for any $\boldsymbol{u} = (u_j)_{j\in\mathbb{N}}, \boldsymbol{v} = (v_j)_{j\in\mathbb{N}} \in (\mathcal{X}')^{\mathbb{N}}$.

Provided $A = (A_j)_{j \geq 0}$ with $A_j \geq 0$ for all $j \geq 0$, the i.i.d. sequence of $\mathcal{X}'$-valued random variables $(\xi_t)_{t \in \mathbb{Z}}$ and $H : (\mathcal{X}')^{\mathbb{N}} \to \mathcal{X}$, we consider that a time series $X = (X_t)_{t \in \mathbb{Z}}$ admitting the following property is a Causal Bernoulli Shift (CBS) with Lipschitz coefficients A and innovations $(\xi_t)_{t \in \mathbb{Z}}$.

(M) The process $X = (X_t)_{t \in \mathbb{Z}}$ meets the representation

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

where H is an A-Lipschitz function with the sequence A satisfying

$$\tilde{A}_* = \sum_{j=0}^{\infty} j A_j < \infty . \tag{4}$$

We additionally define

$$A_* = \sum_{j=0}^{\infty} A_j . \tag{5}$$

CBS regroup several types of nonmixing stationary Markov chains, real-valued functional autoregressive models and Volterra processes, among other interesting models (see [10]). Thanks to the representation (2) and the inequality (3) we assert that AR(d) processes are CBS with $A_j = \sigma \bar{K} \delta_1^j$ for $j \geq 0$.

We let ξ denote a random variable distributed as the ξ_ts. Results from [1] and [2] need a control on the exponential moment of ξ in $\zeta = A_*$, which is provided via the following hypothesis.

(I) The innovations $(\xi_t)_{t \in \mathbb{Z}}$ satisfy $\phi(\zeta) = \mathbb{E}\left[e^{\zeta \|\xi\|}\right] < \infty$.

Bounded or Gaussian innovations trivially satisfy this hypothesis for any $\zeta \in \mathbb{R}$.

Let π_0 denote the probability distribution of the time series Y that we aim to forecast. Observe that for a CBS, π_0 depends only on H and the distribution of ξ. For any $f : \mathcal{X}^{\mathbb{N}^*} \to \mathcal{X}$ measurable and $t \in \mathbb{Z}$ we consider $\hat{Y}_t = f\left((Y_{t-i})_{i \geq 1}\right)$, a possible predictor of Y_t from its past. For a given loss function $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, the prediction risk is evaluated by the expectation of $\ell(\hat{Y}_t, Y_t)$

$$R(f) = \mathbb{E}\left[\ell\left(\hat{Y}_t, Y_t\right)\right] = \pi_0\left[\ell\left(\hat{Y}_t, Y_t\right)\right] = \int_{\mathcal{X}^{\mathbb{Z}}} \ell\left(f\left((y_{t-i})_{i \geq 1}\right), y_t\right) \pi_0(\mathrm{d}\boldsymbol{y}) .$$

We assume in the following that the loss function ℓ fulfills the condition:

(L) For all $\boldsymbol{y}, \boldsymbol{z} \in \mathcal{X}$, $\ell(\boldsymbol{y}, \boldsymbol{z}) = g(\boldsymbol{y} - \boldsymbol{z})$, for some convex function g which is non-negative, $g(0) = 0$ and K- Lipschitz: $|g(\boldsymbol{y}) - g(\boldsymbol{z})| \leq K \|\boldsymbol{y} - \boldsymbol{z}\|$.

If $\mathcal{X}$ is a subset of $\mathbb{R}$, $\ell(y, z) = |y - z|$ satisfies 1 with $K = 1$.

From estimators of dimension d for $\boldsymbol{\theta}$ we can build the corresponding linear predictors $f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\theta}'\boldsymbol{y}_{1:d}$. Speaking more broadly, consider a set Θ and associated with it a set of predictors $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$. For each $\boldsymbol{\theta} \in \Theta$ there is a unique $d = d(\boldsymbol{\theta}) \in \mathbb{N}^*$ such that $f_{\boldsymbol{\theta}} : \mathcal{X}^d \to \mathcal{X}$ is a measurable function from which we define

$$\hat{Y}_t^{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(Y_{t-1}, \ldots, Y_{t-d}) \ ,$$

as a predictor of Y_t given its past. We can extend all functions $f_{\boldsymbol{\theta}}$ in a trivial way (using dummy variables) to start from $\mathcal{X}^{\mathbb{N}^*}$. A natural way to evaluate the predictor associated with $\boldsymbol{\theta}$ is to compute the risk $R(\boldsymbol{\theta}) = R(f_{\boldsymbol{\theta}})$. We use the same letter R by an abuse of notation.

We observe $X_{1:T}$ from $X = (X_t)_{t \in \mathbb{Z}}$, an independent copy of Y. A crucial goal of this work is to build a predictor function $\hat{f}_T$ for Y, inferred from the sample $X_{1:T}$ and Θ such that $R(\hat{f}_T)$ is close to $\inf_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta})$ with π_0- probability close to 1.

The set Θ also depends on T, we write $\Theta \equiv \Theta_T$. Let us define

$$d_T = \sup_{\boldsymbol{\theta} \in \Theta_T} d(\boldsymbol{\theta}) \ . \tag{6}$$

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

(P-1) The set $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ is such that for any $\boldsymbol{\theta} \in \Theta_T$ there are $b_1(\boldsymbol{\theta}), \ldots, b_{d(\boldsymbol{\theta})}(\boldsymbol{\theta}) \in \mathbb{R}_+$ satisfying for all $\boldsymbol{y} = (y_i)_{i \in \mathbb{N}^*}, \boldsymbol{z} = (z_i)_{i \in \mathbb{N}^*} \in \mathcal{X}^{\mathbb{N}^*}$,

$$||f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\boldsymbol{\theta}}(\boldsymbol{z})|| \le \sum_{j=1}^{d(\boldsymbol{\theta})} b_j(\boldsymbol{\theta}) \left|\left|y_j - z_j\right|\right| \ .$$

We assume moreover that $L_T = \sup_{\boldsymbol{\theta} \in \Theta_T} \sum_{j=1}^{d(\boldsymbol{\theta})} b_j(\boldsymbol{\theta}) < \infty$.

(P-2) The inequality $L_T + 1 \le \log T$ holds for all $T \ge 4$.

In the case where $\mathcal{X} \subseteq \mathbb{R}$ and $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ is such that $\boldsymbol{\theta} \in \mathbb{R}^{d(\boldsymbol{\theta})}$ and $f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\theta}'\boldsymbol{y}_{1:d(\boldsymbol{\theta})}$ for all $\boldsymbol{y} \in \mathbb{R}^{\mathbb{N}}$, we have

$$|f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\boldsymbol{\theta}}(\boldsymbol{z})| \le \sum_{j=1}^{d(\boldsymbol{\theta})} |\theta_j| \left|y_j - z_j\right| \ .$$

The last conditions are satisfied by the linear predictors when Θ_T is a subset of the ℓ_1-ball of radius $\log T - 1$ in $\mathbb{R}^{d_T}$.

3 Prediction via Aggregation

The predictor that we propose is defined as an average of predictors $f_{\boldsymbol{\theta}}$ based on the empirical version of the risk,

$$r_T(\boldsymbol{\theta}\,|X) = \frac{1}{T - d(\boldsymbol{\theta})} \sum_{t=d(\boldsymbol{\theta})+1}^{T} \ell\left(\hat{X}_t^{\boldsymbol{\theta}}, X_t\right) .$$

where $\hat{X}_t^{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}\left((X_{t-i})_{i\geq 1}\right)$. The function $r_T(\boldsymbol{\theta}\,|X)$ relies on $X_{1:T}$ and can be computed at stage T; this is in fact a statistic.

We consider a prior probability measure π_T on Θ_T. The prior serves to control the complexity of predictors associated with Θ_T. Using π_T we can construct one predictor 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 $\nu[\exp(h)] = \int \exp(h)\, \mathrm{d}\nu < \infty$, we denote by $\nu\{h\}$ the measure defined as

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

It is known as the Gibbs measure.

Definition 2 (Gibbs predictor) Given $\eta > 0$, called the temperature or the learning rate parameter, we define the Gibbs predictor as the expectation of $f_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ is drawn under $\pi_T\{-\eta r_T(\cdot\,|X)\}$, that is

$$\hat{f}_{\eta,T}(\boldsymbol{y}\,|X) = \pi_T\{-\eta r_T(\cdot\,|X)\}[f_\cdot(\boldsymbol{y})] = \int_{\Theta_T} f_{\boldsymbol{\theta}}(\boldsymbol{y}) \frac{\exp(-\eta r_T(\boldsymbol{\theta}\,|X))}{\pi_T[\exp(-\eta r_T(\cdot\,|X))]} \pi_T(\mathrm{d}\boldsymbol{\theta}) . \tag{7}$$

3.2 PAC-Bayesian Inequality

At this point more care must be taken to describe Θ_T. Here and in the following we suppose that

$$\Theta_T \subseteq \mathbb{R}^{n_T} \text{ for some } n_T \in \mathbb{N}^* . \tag{8}$$

Suppose moreover that Θ_T is equipped with the Borel σ-algebra $\mathcal{B}(\Theta_T)$.

A Lipschitz type hypothesis on $\boldsymbol{\theta}$ guarantees the robustness of the set $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ with respect to the risk R.

(P-3) There is $\mathcal{D} < \infty$ such that for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_T$,

$$\pi_0 \left[\left| f_{\boldsymbol{\theta}_1}\left((X_{t-i})_{i\geq 1}\right) - f_{\boldsymbol{\theta}_2}\left((X_{t-i})_{i\geq 1}\right) \right| \right] \leq \mathcal{D} d_T^{1/2} \, ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|| \ ,$$

where d_T is defined in (6).

Linear predictors satisfy this last condition with $\mathcal{D} = \pi_0\,[|X_1|]$.

Suppose that the $\boldsymbol{\theta}$ reaching the $\inf_{\boldsymbol{\theta}\in\Theta_T} R(\boldsymbol{\theta})$ has some zero components, i.e. $\text{supp}(\boldsymbol{\theta}) < n_T$. Any prior with a lower bounded density (with respect to the Lebesgue measure) allocates zero mass on lower dimensional subsets of Θ_T. Furthermore, if the density is upper bounded we have $\pi_T[B(\boldsymbol{\theta}, \Delta) \cap \Theta_T] = O(\Delta^{n_T})$ for Δ small enough. As we will notice in the proof of Theorem 1, a bound like the previous one would impose a tighter constraint to n_T. Instead we set the following condition.

(P-4) There is a sequence $(\boldsymbol{\theta}_T)_{T\geq 4}$ and constants $\mathcal{C}_1 > 0$, $\mathcal{C}_2, \mathcal{C}_3 \in (0, 1]$ and $\gamma \geq 1$ such that $\boldsymbol{\theta}_T \in \Theta_T$,

$$R(\boldsymbol{\theta}_T) \leq \inf_{\boldsymbol{\theta}\in\Theta_T} R(\boldsymbol{\theta}) + \mathcal{C}_1 \frac{\log^3(T)}{T^{1/2}} \ ,$$

$$\text{and} \quad \pi_T\left[B(\boldsymbol{\theta}_T, \Delta) \cap \Theta_T\right] \geq \mathcal{C}_2 \Delta^{n_T^{1/\gamma}}, \forall 0 \leq \Delta \leq \Delta_T = \frac{\mathcal{C}_3}{T} \ .$$

A concrete example is provided in Sect. 5.

We can now present the main result of this section, our PAC-Bayesian inequality concerning the predictor $\hat{f}_{\eta_T,T}(\cdot\,|X)$ built following (7) with the learning rate $\eta = \eta_T = T^{1/2}/(4\log T)$, provided an arbitrary probability measure π_T on Θ_T.

Theorem 1 *Let ℓ be a loss function such that Assumption (**L**) holds. Consider a process $X = (X_t)_{t\in\mathbb{Z}}$ satisfying Assumption (**M**) and let π_0 denote its probability distribution. Assume that the innovations fulfill Assumption (**I**) with $\zeta = A_*$; A_* is defined in (5). For each $T \geq 4$ let $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ be a set of predictors meeting Assumptions (**P-3**), (**P-4**) and (**P-3**) such that d_T, defined in (6), is at most $T/2$. Suppose that the set Θ_T is as in (8) with $n_T \leq \log^\gamma T$ for some $\gamma \geq 1$ and we let π_T be a probability measure on it such that Assumption (**P-4**) holds for the same γ. Then for any $\varepsilon > 0$, with π_0-probability at least $1 - \varepsilon$,*

$$R\left(\hat{f}_{\eta_T,T}(\cdot\,|X)\right) \leq \inf_{\boldsymbol{\theta}\in\Theta_T} R(f_{\boldsymbol{\theta}}) + \mathcal{E}\frac{\log^3 T}{T^{1/2}} + \frac{8\log T}{T^{1/2}} \log\left(\frac{1}{\varepsilon}\right) \ ,$$

where

$$\mathcal{E} = \mathcal{C}_1 + 8 + \frac{2}{\log 2} - \frac{2\log \mathcal{C}_2}{\log^2 2} - \frac{4\log \mathcal{C}_3}{\log 2} + \frac{8K^2\left(A_* + \tilde{A}_*\right)^2}{\tilde{A}_*^2} + \frac{K\mathcal{D}\mathcal{C}_3}{8\log^3 2} + \frac{4K\phi(A_*)}{\log 2} + \frac{2K^2\phi(A_*)}{\log^2 2}\,, \tag{9}$$

*with $\tilde{A}_*$ defined in (4), K, ϕ and $\mathcal{D}$ in Assumptions (**L**), (**I**) and (**P-3**), respectively, and $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ in Assumption (**P-4**).*

The proof is postponed to Sect. 7.1.

Here however we insist on the fact that this inequality applies to an exact aggregated predictor $\hat{f}_{\eta_T,T}(\cdot\,|X)$. We need to investigate how these predictors are computed and how practical numerical approximations behave compared to the properties of the exact version.

4 Stochastic Approximation

Once we have the observations $X_{1:T}$, we use the Metropolis – Hastings algorithm to compute $\hat{f}_{\eta,T}(\cdot\,|X) = \int f_{\boldsymbol{\theta}}(\cdot\,|X)\,\pi_T\{-\eta r_T(\boldsymbol{\theta}\,|X)\}(\mathrm{d}\boldsymbol{\theta})$. The Gibbs measure $\pi_T\{-\eta r_T(\cdot\,|X)\}$ is a distribution on Θ_T whose density $\pi_{\eta,T}(\cdot\,|X)$ with respect to π_T is proportional to $\exp(-\eta r_T(\cdot\,|X))$.

4.1 Metropolis: Hastings Algorithm

Given $X \in \mathcal{X}^{\mathbb{Z}}$, the Metropolis-Hastings algorithm generates a Markov chain $\Phi_{\eta,T}(X) = (\boldsymbol{\theta}_{\eta,T,n}(X))_{n\geq 0}$ with kernel $P_{\eta,T}$ (only depending on $X_{1:T}$) having the target distribution $\pi_T\{-\eta r_T(\cdot\,|X)\}$ as the unique invariant measure, based on the transitions of another Markov chain which serves as a proposal (see [21]). We consider a proposal transition of the form $Q_{\eta,T}(\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}) = q_{\eta,T}(\boldsymbol{\theta}_1, \boldsymbol{\theta})\pi_T(\mathrm{d}\boldsymbol{\theta})$ where the conditional density kernel $q_{\eta,T}$ (possibly also depending on $X_{1:T}$) on $\Theta_T \times \Theta_T$ is such that

$$\beta_{\eta,T}(X) = \inf_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\Theta_T\times\Theta_T} \frac{q_{\eta,T}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\pi_{\eta,T}(\boldsymbol{\theta}_2\,|X)} \in (0,1)\ . \tag{10}$$

This is the case of the independent Hastings algorithm, where the proposal is i.i.d. with density $q_{\eta,T}$. The condition gets into

$$\beta_{\eta,T}(X) = \inf_{\boldsymbol{\theta}\in\Theta_T} \frac{q_{\eta,T}(\boldsymbol{\theta})}{\pi_{\eta,T}(\boldsymbol{\theta}\,|X)} \in (0,1)\ . \tag{11}$$

In Sect. 5 we provide an example.

The relation (10) implies that the algorithm is uniformly ergodic, i.e. we have a control in total variation norm ($\| \cdot \|_{TV}$). Thus, the following condition holds (see [18]).

(A) Given $\eta, T > 0$, there is $\beta_{\eta,T} : \mathcal{X}^{\mathbb{Z}} \to (0,1)$ such for any $\boldsymbol{\theta}_0 \in \Theta_T$, $\boldsymbol{x} \in \mathcal{X}^{\mathbb{Z}}$ and $n \in \mathbb{N}$, the chain $\Phi_{\eta,T}(\boldsymbol{x})$ with transition law $P_{\eta,T}$ and invariant distribution $\pi_T\{-\eta r_T(\cdot|\boldsymbol{x})\}$ satisfies

$$\left\|P^n_{\eta,T}(\boldsymbol{\theta}_0, \cdot) - \pi_T\{-\eta r_T(\cdot|\boldsymbol{x})\}\right\|_{TV} \leq 2\left(1 - \beta_{\eta,T}(\boldsymbol{x})\right)^n .$$

4.2 Theoretical Bounds for the Computation

In [16, Theorem 3.1] we find a bound on the mean square error of approximating one integral by the empirical estimate obtained from the successive samples of certain ergodic Markov chains, including those generated by the MCMC method that we use.

A MCMC method adds a second source of randomness to the forecasting process and our aim is to measure it. Let $\boldsymbol{\theta}_0 \in \cap_{T \geq 1}\Theta_T$, we set $\boldsymbol{\theta}_{\eta,T,0}(\boldsymbol{x}) = \boldsymbol{\theta}_0$ for all $T, \eta > 0$, $\boldsymbol{x} \in \mathcal{X}^{\mathbb{Z}}$. We denote by $\mu_{\eta,T}(\cdot|X)$ the probability distribution of the Markov chain $\Phi_{\eta,T}(X)$ with initial point $\boldsymbol{\theta}_0$ and kernel $P_{\eta,T}$.

Let $\nu_{\eta,T}$ denote the probability distribution of $(X, \Phi_{\eta,T}(X))$; it is defined by setting for all sets $A \in (\mathcal{B}(\mathcal{X}))^{\otimes\mathbb{Z}}$ and $B \in (\mathcal{B}(\Theta_T))^{\otimes\mathbb{N}}$

$$\nu_{\eta,T}(A \times B) = \int \mathbb{1}_A(\boldsymbol{x})\, \mathbb{1}_B(\boldsymbol{\phi})\, \mu_{\eta,T}(\mathrm{d}\boldsymbol{\phi}|\boldsymbol{x})\, \pi_0(\mathrm{d}\boldsymbol{x}) \tag{12}$$

Given $\Phi_{\eta,T} = (\boldsymbol{\theta}_{\eta,T,n})_{n \geq 0}$, we then define for $n \in \mathbb{N}^*$

$$\bar{f}_{\eta,T,n} = \frac{1}{n}\sum_{i=0}^{n-1} f_{\boldsymbol{\theta}_{\eta,T,i}} . \tag{13}$$

Since our chain depends on X, we make it explicit by using the notation $\bar{f}_{\eta,T,n}(\cdot|X)$. The cited [16, Theorem 3.1] leads to a proposition that applies to the numerical approximation of the Gibbs predictor (the proof is in Sect. 7.2). We stress that this is independent of the model (CBS or any), of the set of predictors and of the theoretical guarantees of Theorem 1.

Proposition 1 *Let ℓ be a loss function meeting Assumption (**L**). Consider any process $X = (X_t)_{t \in \mathbb{Z}}$ with an arbitrary probability distribution π_0. Given $T \geq 2$, $\eta > 0$, a set of predictors $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ and $\pi_T \in \mathcal{M}^1_+(\Theta_T)$, let $\hat{f}_{\eta,T}(\cdot|X)$ be defined by (7) and let $\bar{f}_{\eta,T,n}(\cdot|X)$ be defined by (13). Suppose that $\Phi_{\eta,T}$ meets Assumption (**A**) for η and T with a function $\beta_{\eta,T} : \mathcal{X}^{\mathbb{Z}} \to (0,1)$. Let $\nu_{\eta,T}$ denote* *the probability distribution of* $(X, \Phi_{\eta,T}(X))$ *as defined in (14). Then, for all* $n \geq 1$ *and* $D > 0$*, with* $\nu_{\eta,T}$*- probability at least* $\max\{0, 1 - A_{\eta,T}/(Dn^{1/2})\}$ *we have* $|R(\bar{f}_{\eta,T,n}(\cdot\,|X)) - R(\hat{f}_{\eta,T}(\cdot\,|X))| \leq D$*, where*

$$A_{\eta,T} = 3K \int\limits_{\mathcal{X}^{\mathbb{Z}}} \frac{1}{\beta_{\eta,T}(\boldsymbol{x})} \int\limits_{\mathcal{X}^{\mathbb{Z}}} \sup_{\boldsymbol{\theta}\in\Theta_T} \left| f_{\boldsymbol{\theta}}(\boldsymbol{y}) - \hat{f}_{\eta,T}(\boldsymbol{y}\,|\boldsymbol{x}) \right| \pi_0(\mathrm{d}\boldsymbol{y})\, \pi_0(\mathrm{d}\boldsymbol{x}) \ . \tag{14}$$

We denote by $\nu_T = \nu_{\eta_T,T}$ the probability distribution of $(X, \Phi_{\eta,T}(X))$ setting $\eta = \eta_T = T^{1/2}/(4\log T)$. As Theorem 1 does not involve any simulation, it also holds in ν_T- probability. From this and Proposition 1 a union bound gives us the following.

Theorem 2 *Under the hypothesis of Theorem 1, consider moreover that Assumption (**A**) is fulfilled by* $\Phi_{\eta,T}$ *for all* $\eta = \eta_T$ *and* T *with* $T \geq 4$*. Thus, for all* $\varepsilon > 0$ *and* $n \geq M(T, \varepsilon)$*, with* ν_T*- probability at least* $1 - \varepsilon$ *we have*

$$R\left(\bar{f}_{\eta_T,T,n}(\cdot\,|X)\right) \leq \inf_{\boldsymbol{\theta}\in\Theta_T} R(f_{\boldsymbol{\theta}}) + \left(\mathcal{E} + \frac{2}{\log 2} + 2\right) \frac{\log^3 T}{T^{1/2}} + \frac{8\log T}{T^{1/2}} \log\left(\frac{1}{\varepsilon}\right) \ ,$$

where $\mathcal{E}$ *is defined in (9) and* $M(T, \varepsilon) = A^2_{\eta_T,T} T/(\varepsilon^2 \log^6 T)$ *with* $A_{\eta,T}$ *as in (14).*

5 Applications to the Autoregressive Process

We carefully recapitulate all the assumptions of Theorem 2 in the context of an autoregressive process. After that, we illustrate numerically the behaviour of the proposed method.

5.1 Theoretical Considerations

Consider a real valued stable autoregressive process of finite order d as defined by (1) with parameter $\boldsymbol{\theta}$ lying in the interior of $s_d(\delta)$ and unit normally distributed innovations (Assumptions (**M**) and (**I**) hold). With the loss function $\ell(y, z) = |y - z|$ Assumption (**L**) holds as well. The linear predictors is the set that we test; they meet Assumption (**P-3**). Without loss of generality assume that $d_T = n_T$. In the described framework we have $\hat{f}_{\eta,T}(\cdot\,|X) = f_{\widehat{\boldsymbol{\theta}}_{\eta,T}(X)}$, where

$$\hat{\boldsymbol{\theta}}_{\eta,T}(X) = \int\limits_{\Theta_T} \boldsymbol{\theta} \frac{\exp(-\eta r_T(\boldsymbol{\theta}\,|X))}{\pi_T[\exp(-\eta r_T(\boldsymbol{\theta}\,|X))]} \pi_T(\mathrm{d}\boldsymbol{\theta}) \ .$$

This $\hat{\boldsymbol{\theta}}_{\eta,T}(X) \in \mathbb{R}^{d_T}$ is known as the Gibbs estimator.

Remark that, by (2) and the normality of the innovations, the risk of any $\hat{\boldsymbol{\theta}} \in \mathbb{R}^{d_T}$ is computed as the absolute moment of a centered Gaussian, namely

$$R\left(f_{\hat{\boldsymbol{\theta}}}\right) = R\left(\hat{\boldsymbol{\theta}}\right) = \frac{\left(2\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \Gamma_T \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) + 2\sigma^2\right)^{1/2}}{\pi^{1/2}}, \tag{15}$$

where $\Gamma_T = (\gamma_{i,j})_{0 \le i,j \le d_T - 1}$ is the covariance matrix of the process. In (15) the vector $\boldsymbol{\theta}$ originally in $\mathbb{R}^d$ is completed by $d_T - d$ zeros.

In this context $\arg\inf_{\boldsymbol{\theta} \in \mathbb{R}^{\mathbb{N}^*}} R(\boldsymbol{\theta}) \in s_d(1)$ gives the true parameter $\boldsymbol{\theta}$ generating the process. Let us verify Assumption (**P-4**) by setting conveniently Θ_T and π_T. Let $\Delta_{d*} > 0$ be such that $B(\boldsymbol{\theta}, \Delta_{d*}) \subseteq s_d(1)$.

We express $\Theta_T = \bigcup_{k=1}^{d_T} \Theta_{k,T}$ where $\boldsymbol{\theta} \in \Theta_{k,T}$ if and only if $d(\boldsymbol{\theta}) = k$. It is interesting to set $\Theta_{k,T}$ as the part of the stability domain of an AR(k) process satisfying Assumptions (**P-3**) and (**P-4**). Consider $\Theta_{1,T} = s_1(1) \times \{0\}^{d_T - 1} \cap B_1(\mathbf{0}, \log T - 1)$ and $\Theta_{k,T} = s_k(1) \times \{0\}^{d_T - k} \cap B_1(\mathbf{0}, \log T - 1) \backslash \Theta_{k-1,T}$ for $k \ge 2$. Assume moreover that $d_T = \lfloor \log^\gamma T \rfloor$.

We write $\pi_T = \sum_{k=1}^{d_T} c_{k,T} \pi_{k,T}$ where for all k, $c_{k,T}\pi_{k,T}$ is the restriction of π_T to $\Theta_{k,T}$ with $c_{k,T}$ a real non negative number and $\pi_{k,T}$ a probability measure on $\Theta_{k,T}$. In this setup $c_{k,T} = \pi_T[\Theta_{k,T}]$ and $\pi_{k,T}[A \cap \Theta_{k,T}] = \pi_T[A \cap \Theta_{k,T}] / c_{k,T}$ if $c_{k,T} > 0$ and $\pi_{k,T}[A \cap \Theta_{k,T}] = 0$ otherwise. The vector $\left[c_{1,T} \ldots c_{d_T,T}\right]$ could be interpreted as a prior on the model order. Set $c_{k,T} = c_k / (\sum_{i=1}^{d_T} c_i)$ where $c_k > 0$ is the k-th term of a convergent series ($\sum_{k=1}^{\infty} c_k = c^* < \infty$).

The distribution $\pi_{k,T}$ is inferred from some transformations explained below. Observe first that if $a \le b$ we have $s_k(a) \subseteq s_k(b)$. If $\boldsymbol{\theta} \in s_k(1)$ then $\left[\lambda\theta_1 \ldots \lambda^k\theta_k\right]' \in s_k(1)$ for any $\lambda \in (-1, 1)$. Let us set

$$\lambda_T(\boldsymbol{\theta}) = \min\left\{1, \frac{\log T - 1}{\|\boldsymbol{\theta}\|_1}\right\} .$$

We define $F_{k,T}(\boldsymbol{\theta}) = \left[\lambda_T(\boldsymbol{\theta})\theta_1 \ldots \lambda_T^k(\boldsymbol{\theta})\theta_k \; 0 \ldots 0\right]' \in \mathbb{R}^{d_T}$. Remark that for any $\boldsymbol{\theta} \in s_k(1)$, $\|F_{k,T}(\boldsymbol{\theta})\|_1 \le \lambda_T(\boldsymbol{\theta})\|\boldsymbol{\theta}\|_1 \le \log T - 1$. This gives us an idea to generate vectors in $\Theta_{k,T}$. Our distribution $\pi_{k,T}$ is deduced from:

Algorithm 1: $\pi_{k,T}$ generation

```
input an effective dimension $k$, the number of observations $T$ and $F_{k,T}$;
generate a random $\boldsymbol{\theta}$ uniformly on $s_k(1)$;
return $F_{k,T}(\boldsymbol{\theta})$
```

The distribution $\pi_{k,T}$ is lower bounded by the uniform distribution on $s_k(1)$.

Provided any $\gamma \ge 1$, let $T_* = \min\{T : d_T \ge d^\gamma, \log T \ge d^{1/2}2^d\}$. Since $s_k(1) \subseteq B(\mathbf{0}, 2^k - 1)$ (see [19, Lemma 1]) and $k^{1/2}\|\boldsymbol{\theta}\| \ge \|\boldsymbol{\theta}\|_1$ for any $\boldsymbol{\theta} \in \mathbb{R}^k$, the constraint $\|\boldsymbol{\theta}\|_1 \le \log T - 1$ becomes redundant in $\Theta_{k,T}$ for $1 \le k \le d$ and $T \ge T_*$, i.e.

$\Theta_{1,T} = s_1(1) \times \{0\}^{d_T-1}$ and $\Theta_{k,T} = s_k(1) \times \{0\}^{d_T-k} \backslash \Theta_{k-1,T}$ for $2 \leq k \leq d$. We define the sequence of Assumption (**P-4**) as $\boldsymbol{\theta}_T = \mathbf{0}$ for $T < T_*$ and $\boldsymbol{\theta}_T = \arg\inf_{\boldsymbol{\theta} \in \Theta_T} R(\boldsymbol{\theta})$ for $T \geq T_*$. Remark that the first d components of $\boldsymbol{\theta}_T$ are constant for $T \geq T_*$ (they correspond to the $\boldsymbol{\theta} \in \mathbb{R}^d$ generating the AR(d) process), and the last $d_T - d$ are zero. Let $\Delta_{1*} = 2\log 2 - 1$. Then, we have for $T < T_*$ and all $\Delta \in [0, \Delta_{1*}]$

$$\pi_T \left[B\left(\boldsymbol{\theta}_T, \Delta\right) \cap \Theta_T\right] \geq c_{1,T} \pi_{1,T} \left[B\left(\mathbf{0}, \Delta\right) \cap s_1(1) \times \{0\}^{d_T-1}\right] \geq \frac{c_1}{c^*} \Delta \ .$$

Furthermore, for $T \geq T_*$ and $\Delta \in [0, \Delta_{d*}]$

$$\pi_T \left[B\left(\boldsymbol{\theta}_T, \Delta\right) \cap \Theta_T\right] \geq c_{d,T} \pi_{d,T} \left[B\left(\boldsymbol{\theta}_T, \Delta\right) \cap s_d(1) \times \{0\}^{d_T-d}\right] \geq \frac{c_d}{2^{d^2} c^*} \Delta^d \ .$$

Assumption (**P-4**) is then fulfilled for any $\gamma \geq 1$ with

$$\begin{aligned} \mathcal{C}_1 &= \max\left\{0, (R\left(0\right) - \inf_{\boldsymbol{\theta} \in \Theta_T} R\left(\boldsymbol{\theta}\right)) T^{1/2} \log^{-3} T, 4 \leq T < T_*\right\} \\ \mathcal{C}_2 &= \min\left\{1, \frac{c_1}{c^*}, \frac{c_d}{2^{d^2} c^*}\right\} \\ \mathcal{C}_3 &= \min\left\{1, 4\Delta_{1*}, T_* \Delta_{d*}\right\} \ . \end{aligned}$$

Let $q_{\eta,T}$ be the constant function 1, this means that the proposal has the same distribution π_T. Let us bound the ratio (11).

$$\begin{aligned} \beta_{\eta,T}\left(X\right) = \inf_{\boldsymbol{\theta} \in \Theta_T} \frac{q_{\eta,T}\left(\boldsymbol{\theta}\right)}{\pi_{\eta,T}\left(\boldsymbol{\theta} \left| X\right.\right)} &= \inf_{\boldsymbol{\theta} \in \Theta_T} \frac{\sum_{k=1}^{d_T} c_{k,T} \int_{\Theta_{k,T}} \exp\left(-\eta r_T\left(z \left| X\right.\right)\right) \pi_{k,T}\left(\mathrm{d}z\right)}{\exp\left(-\eta r_T\left(\boldsymbol{\theta} \left| X\right.\right)\right)} \\ &\geq \sum_{k=1}^{d_T} c_{k,T} \int_{\Theta_{k,T}} \exp\left(-\eta r_T\left(z \left| X\right.\right)\right) \pi_{k,T}\left(\mathrm{d}z\right) > 0 \ . \end{aligned} \tag{16}$$

Now note that

$$\left|x_t - f_{\boldsymbol{\theta}}\left(\left(x_{t-i}\right)_{i \geq 1}\right)\right| \leq \left|x_t\right| + \sum_{j=1}^{d(\boldsymbol{\theta})} \left|\theta_j\right| \left|x_{t-j}\right| \leq \log T \max_{j=0,\ldots,d(\boldsymbol{\theta})} \left|x_{t-j}\right| \ . \tag{17}$$

Plugging the bound (17) on (16) with $\eta = \eta_T$

$$\beta_{\eta_T,T}\left(\boldsymbol{x}\right) \geq \sum_{k=1}^{d_T} c_k \int_{\Theta_k} \exp\left(-\eta_T r_T\left(z \left| \boldsymbol{x}\right.\right)\right) \pi_k\left(\mathrm{d}z\right) \geq \exp\left(-\frac{T^{1/2}}{4} \max_{j=0,\ldots,d_T} \left|x_{t-j}\right|\right) \ ,$$

we deduce that

$$\frac{1}{\beta_{\eta_T,T}(\boldsymbol{x})} \le \sum_{k=0}^{d_T} \exp\left(\frac{T^{1/2}\left|x_{t-j}\right|}{4}\right) . \tag{18}$$

Taking (18) into account, setting $\gamma = 1$ (thus $d_T = \lfloor \log T \rfloor$), using Assumption (**P-3**), that $K = 1$ and applying the Cauchy-Schwarz inequality we get

$$\begin{aligned} A_{\eta_T,T} &= 3K \int_{\mathcal{X}^{\mathbb{Z}}} \frac{1}{\beta_{\eta_T,T}(\boldsymbol{x})} \int_{\mathcal{X}^{\mathbb{Z}}} \sup_{\boldsymbol{\theta}\in\Theta_T} \left| f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\hat{\boldsymbol{\theta}}_{\eta_T,T}(\boldsymbol{x})}(\boldsymbol{y}) \right| \pi_0(\mathrm{d}\boldsymbol{y})\,\pi_0(\mathrm{d}\boldsymbol{x}) \\ &\le 3\,(d_T+1)\,d_T^{1/2}\pi_0\left[\exp\left(\frac{T^{1/2}\,|X_1|}{4}\right)\right]\pi_0\left[|X_1|\right] \sup_{\boldsymbol{\theta}\in\Theta_T} ||\boldsymbol{\theta}|| \\ &\le 6\log^{3/2} T\pi_0\left[\exp\left(\frac{T^{1/2}\,|X_1|}{4}\right)\right]\pi_0\left[|X_1|\right] . \end{aligned}$$

As X_1 is centered and normally distributed of variance γ_0, $\pi_0\left[|X_1|\right] = (2\gamma_0/\pi)^{1/2}$ and $\pi_0[\exp(T^{1/2}\,|X_1|\,/4)] = \gamma_0 T^{1/2}\exp(\gamma_0 T/32)/4$.

From $n \ge M^*(T,\varepsilon) = 9\gamma_0^3 T^2 \exp(\gamma_0 T/16)/(2\pi\varepsilon^2\log^3 T)$ the result of Theorem 2 is reached. This bound of $M(T,\varepsilon)$ is prohibitive from a computational viewpoint. That is why we limit the number of iterations to a fixed n^*.

What we obtain from MCMC is $\bar{f}_{\eta_T,T,n}(\boldsymbol{y}\,|X) = \bar{\boldsymbol{\theta}}'_{\eta_T,T,n}(X)\boldsymbol{y}_{1:d_T}$ with $\bar{\boldsymbol{\theta}}_{\eta_T,T,n}(X) = \sum_{i=0}^{n-1}\boldsymbol{\theta}_{\eta_T,T,i}(X)/n$. Remark that $\bar{f}_{\eta_T,T,n}(\cdot\,|X) = f_{\bar{\boldsymbol{\theta}}_{\eta_T,T,n}(X)}$. The risk is expressed as

$$R\left(\bar{f}_{\eta_T,T,n}(\cdot\,|X)\right) = \frac{\left(2\left(\bar{\boldsymbol{\theta}}_{\eta_T,T,n}(X) - \boldsymbol{\theta}\right)'\Gamma(Y)\left(\bar{\boldsymbol{\theta}}_{\eta_T,T,n}(X) - \boldsymbol{\theta}\right) + 2\sigma^2\right)^{1/2}}{\pi^{1/2}} .$$

5.2 Numerical Work

Consider 100 realisations of an autoregressive processes X simulated with the same $\boldsymbol{\theta} \in s_d(\delta)$ for $d = 8$ and $\delta = 3/4$ and with $\sigma = 1$. Let $\boldsymbol{c}^{(i)}$, $i = 1, 2$ the sequences defining two different priors in the model order:

1. $c_k^{(1)} = k^{-2}$, the sparsity is favoured,
2. $c_k^{(2)} = \mathrm{e}^{-k}$, the sparsity is strongly favoured.

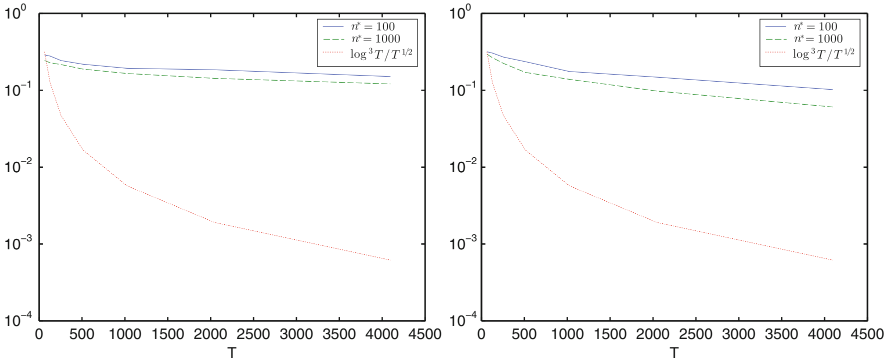

Fig. 1 The plots represent the 0.9-quantiles in data $R(\bar{\boldsymbol{\theta}}_{\eta_T,T,n^*}(X)) - (2/\pi)^{1/2}\sigma^2$ for $T = 32, 64, \ldots, 4{,}096$. The graph on the *left corresponds* to the order prior $c_k^{(1)} = k^{-2}$ while that on the *right corresponds* to $c_k^{(2)} = \mathrm{e}^{-k}$. The *solid curves* were plotted with $n^* = 100$, the *dashed ones* with $n^* = 1{,}000$ and as a reference, the *dotted curve* is proportional to $\log^3 T/T^{1/2}$

For each sequence $\boldsymbol{c}$ and for each value of $T \in \{2^j, j = 6, \ldots, 12\}$ we compute $\bar{\boldsymbol{\theta}}_{\eta_T,T,n^*}$, the MCMC approximation of the Gibbs estimator using Algorithm 2 with $\eta = \eta_T$.

```
Algorithm 2: Independent Hastings Sampler
  input the sample $X_{1:T}$ of $X$, the prior $\boldsymbol{c}$, the learning rate $\eta$, the generators $\pi_{k,T}$
        for $k = 1, \ldots, d_T$ and a maximum iterations number $n^*$;
  initialization $\boldsymbol{\theta}_{\eta,T,0} = \mathbf{0}$;
  for $i$=1 to $n^* - 1$ do
      generate $k \in \{1, \ldots, d_T\}$ using the prior $\boldsymbol{c}$;
      generate $\boldsymbol{\theta}_{candidate} \sim \pi_{k,T}$;
      generate $U \sim \mathcal{U}(0,1)$;
      if $U \le \alpha_{\eta,T,X}(\boldsymbol{\theta}_{\eta,T,i-1}, \boldsymbol{\theta}_{candidate})$ then
          $\boldsymbol{\theta}_{\eta,T,i} = \boldsymbol{\theta}_{candidate}$ else
              $\boldsymbol{\theta}_{\eta,T,i} = \boldsymbol{\theta}_{\eta,T,i-1}$;
  return $\bar{\boldsymbol{\theta}}_{\eta,T,n^*}(X) = \sum_{i=0}^{n^*-1} \boldsymbol{\theta}_{\eta,T,k}(X)/n^*$.
```

The acceptance rate is computed as $\alpha_{\eta,T,X}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \exp(\eta r_T(\boldsymbol{\theta}_1|X) - \eta r_T(\boldsymbol{\theta}_2|X))$.

Algorithm 1 used by the distributions $\pi_{k,T}$ generates uniform random vectors on $s_k(1)$ by the method described in [6]. It relies in the Levinson-Durbin recursion algorithm. We also implemented the numerical improvements of [3].

Set $\varepsilon = 0.1$. Figure 1 displays the $(1-\varepsilon)$-quantiles in data $R(\bar{\boldsymbol{\theta}}_{\eta_T,T,n^*}(X)) - (2/\pi)^{1/2}\sigma^2$ for $\boldsymbol{c}^{(1)}$ and $\boldsymbol{c}^{(2)}$ using different values of n^*.

Note that, for the proposed algorithm the prediction risk decreases very slowly when the number T of observations grows and the number of MCMC iterations remains constant. If $n^* = 1{,}000$ the decaying rate is faster than if $n^* = 100$ for smaller values of T. For $T \geq 2{,}000$ we observe that both rates are roughly the same in the logarithmic scale. This behaviour is similar in both cases presented in Fig. 1. As expected, the risk of the approximated predictor does not converge as $\log^3 T/T^{1/2}$.

6 Discussion

There are two sources of error in our method: prediction (of the exact Gibbs predictor) and approximation (using the MCMC). The first one decays when T grows and the obtained guarantees for the second one explode. We found a possibly pessimistic upper bound for $M(T,\epsilon)$. The exponential growing of this bound is the main weakness of our procedure. The use of a better adapted proposal in the MCMC algorithm needs to be investigated. The Metropolis Langevin Algorithm (see [4]) gives us an insight in this direction. However it is encouraging to see that, in the analysed practical case, the risk of $\bar{f}_{\eta_T,T,n^*}(\cdot\,|X)$ does not increase with T.

7 Technical Proofs

7.1 Proof of Theorem 1

The proof of Theorem 1 is based on the same tools used by [2] up to Lemma 3. For the sake of completeness we quote the essential ones.

We denote by $\mathcal{M}_+^1(F)$ the set of probability measures on the measurable space $(F, \mathcal{F})$. Let $\rho, \nu \in \mathcal{M}_+^1(F)$, $\mathcal{K}(\rho,\nu)$ stands for the Kullback-Leibler divergence of ν from ρ.

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

The first lemma can be found in [8, Equation 5.2.1].

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

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

with the convention $\infty - \infty = -\infty$. 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 by the Gibbs measure $\nu\{h\}$.

For a fixed $C > 0$, let $\tilde{\xi}_t^{(C)} = \max\{\min\{\xi_t, C\}, -C\}$. Consider $\tilde{X}_t = H(\tilde{\xi}_t^{(C)}, \tilde{\xi}_{t-1}^{(C)}, \ldots)$.

Denote $\tilde{X} = (\tilde{X}_t)_{t\in\mathbb{Z}}$ and by $\tilde{R}(\boldsymbol{\theta})$ and $\tilde{r}_T\left(\boldsymbol{\theta}\,\middle|\tilde{X}\right)$ the respective exact and empirical risks associated with $\tilde{X}$ in $\boldsymbol{\theta}$.

$$\tilde{R}(\boldsymbol{\theta}) = \mathbb{E}\left[\ell\left(\widehat{\tilde{X}}_t^{\boldsymbol{\theta}}, \tilde{X}_t\right)\right],$$

$$\tilde{r}_T\left(\boldsymbol{\theta}\,\middle|\tilde{X}\right) = \frac{1}{T - d(\boldsymbol{\theta})} \sum_{t=d(\boldsymbol{\theta})+1}^{T} \ell\left(\widehat{\tilde{X}}_t^{\boldsymbol{\theta}}, \tilde{X}_t\right),$$

where $\widehat{\tilde{X}}_t^{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}((\tilde{X}_{t-i})_{i\geq 1})$.

This thresholding is interesting because truncated CBS are weakly dependent processes (see [2, Section 4.2]).

A Hoeffding type inequality introduced in [20, Theorem 1] provides useful controls on the difference between empirical and exact risks of a truncated process.

Lemma 2 (Laplace transform of the risk) *Let ℓ be a loss function meeting Assumption (**L**) and $X = (X_t)_{t\in\mathbb{Z}}$ a process satisfying Assumption (**M**). For all $T \geq 2$, any $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ satisfying Assumption (**P-1**), Θ_T such that d_T, defined in (6), is at most $T/2$, any truncation level $C > 0$, $\eta \geq 0$ and $\boldsymbol{\theta} \in \Theta_T$ we have,*

$$\mathbb{E}\left[\exp\left(\eta\left(\tilde{R}(\boldsymbol{\theta}) - \tilde{r}_T\left(\boldsymbol{\theta}\,\middle|\tilde{X}\right)\right)\right)\right] \leq \exp\left(\frac{4\eta^2 k^2(T, C)}{T}\right), \qquad (19)$$

and

$$\mathbb{E}\left[\exp\left(\eta\left(\tilde{r}_T\left(\boldsymbol{\theta}\,\middle|\tilde{X}\right) - \tilde{R}(\boldsymbol{\theta})\right)\right)\right] \leq \exp\left(\frac{4\eta^2 k^2(T, C)}{T}\right), \qquad (20)$$

*where $k(T, C) = 2^{1/2} C K (1 + L_T)\left(A_* + \tilde{A}_*\right)$. The constants $\tilde{A}_*$ and A_* are defined in (4) and (5) respectively, K and L_T in Assumptions (**L**) and (**P-1**) respectively.*

The following lemma is a slight modification of [2, Lemma 6.5]. It links the two versions of the empirical risk: original and truncated.

Lemma 3 *Suppose that Assumption (**L**) holds for the loss function ℓ, Assumption (**P-1**) holds for $X = (X_t)_{t\in\mathbb{Z}}$ and Assumption (**I**) holds for the innovations with $\zeta = A_*$; A_* is defined in (5). For all $T \geq 2$, any $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ meeting Assumption (**P-1**) with Θ_T such that d_T, defined in (6), is at most $T/2$, any truncation* *level $C > 0$ and any $0 \leq \eta \leq T/4\,(1 + L_T)$ we have,*

$$\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta} \in \Theta_T} \left| r_T\left(\boldsymbol{\theta}\left|X\right.\right) - \tilde{r}_T\left(\boldsymbol{\theta}\left|\tilde{X}\right.\right)\right|\right)\right] \leq \exp\left(\eta\varphi\left(T, C, \eta\right)\right) \ ,$$

where

$$\varphi(T, C, \eta) = 2K(1 + L_T)\phi(A_*)\left(\frac{A_*C}{\exp\left(A_*C\right) - 1} + \eta\frac{4K(1 + L_T)}{T}\right) \ ,$$

*with K and L_T defined in Assumptions (**L**) and (**P-1**) respectively.*

Finally we present a result on the aggregated predictor defined in (7). The proof is partially inspired by that of [2, Theorem 3.2].

Lemma 4 *Let ℓ be a loss function such that Assumption (**L**) holds and let $X = (X_t)_{t\in\mathbb{Z}}$ a process satisfying Assumption (**M**) with probability distribution π_0. For each $T \geq 2$ let $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ be a set of predictors and $\pi_T \in \mathcal{M}_+^1\left(\Theta_T\right)$ any prior probability distribution on Θ_T. We build the predictor $\hat{f}_{\eta,T}\left(\cdot\left|X\right.\right)$ following (7) with any $\eta > 0$. For any $\varepsilon > 0$ and any truncation level $C > 0$, with π_0-probability at least $1 - \varepsilon$ we have,*

$$\begin{aligned} R\left(\hat{f}_{\eta,T}\left(\cdot\left|X\right.\right)\right) \leq & \inf_{\rho \in \mathcal{M}_+^1(\Theta_T)} \left\{\rho\left[R\right] + \frac{2\mathcal{K}\left(\rho, \pi_T\right)}{\eta}\right\} + \frac{2\log\left(\frac{2}{\varepsilon}\right)}{\eta} \\ & + \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right]\right) + \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\tilde{r}_T - \tilde{R}\right)\right)\right]\right) \\ & + \frac{2}{\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T\left(\boldsymbol{\theta}\left|X\right.\right) - \tilde{r}_T\left(\boldsymbol{\theta}\left|\tilde{X}\right.\right)\right|\right)\right]\right) \ . \end{aligned}$$

Proof We use Tonelli's theorem and Jensen's inequality with the convex function g to obtain an upper bound for $R\left(\hat{f}_{\eta,T}\left(\cdot\left|X\right.\right)\right)$

$$\begin{aligned} R\left(\hat{f}_{\eta,T}\left(\cdot\left|X\right.\right)\right) &= \int_{\mathcal{X}^{\mathbb{Z}}} g\left(\int_{\Theta_T}\left(f_{\boldsymbol{\theta}}\left((y_{t-i})_{i\geq1}\right) - y_t\right)\pi_T\left\{-\eta r_T\left(\cdot\left|X\right.\right)\right\}(\mathrm{d}\boldsymbol{\theta})\right)\pi_0\left(\mathrm{d}\mathbf{y}\right) \\ &\leq \int_{\mathcal{X}^{\mathbb{Z}}}\left[\int_{\Theta_T} g\left(f_{\boldsymbol{\theta}}\left((y_{t-i})_{i\geq1}\right) - y_t\right)\pi_T\left\{-\eta r_T\left(\cdot\left|X\right.\right)\right\}(\mathrm{d}\boldsymbol{\theta})\right]\pi_0\left(\mathrm{d}\mathbf{y}\right) \\ &= \int_{\Theta_T}\left[\int_{\mathcal{X}^{\mathbb{Z}}} g\left(f_{\boldsymbol{\theta}}\left((y_{t-i})_{i\geq1}\right) - y_t\right)\pi_0\left(\mathbf{y}\right)\right]\pi_T\left\{-\eta r_T\left(\cdot\left|X\right.\right)\right\}(\mathrm{d}\boldsymbol{\theta}) = \pi_T\left\{-\eta r_T\left(\cdot\left|X\right.\right)\right\}\left[R\right] \ . \end{aligned}$$

In the remainder of this proof we search for upper bounding $\pi_T\left\{-\eta r_T\left(\cdot\left|X\right.\right)\right\}\left[R\right]$.

First, we use the relationship:

$$R - r_T(\cdot\,|X) = \left(\tilde{R} - \tilde{r}_T\left(\cdot\,|\tilde{X}\right)\right) + \left(R - \tilde{R}\right) - \left(r_T(\cdot\,|X) - \tilde{r}_T\left(\cdot\,|\tilde{X}\right)\right)\ . \tag{21}$$

For the sake of simplicity and while it does not disrupt the clarity, we lighten the notation of r_T and $\tilde{r}_T$. We now suppose that in the place of $\boldsymbol{\theta}$ we have a random variable distributed as $\pi_T \in \mathcal{M}_+^1(\Theta_T)$. This is taken into account in the following expectations. The identity (21) and the Cauchy-Schwarz inequality lead to

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\eta}{2}(R - r_T)\right)\right] &= \mathbb{E}\left[\exp\left(\frac{\eta}{2}\left(\tilde{R} - \tilde{r}_T\right)\right)\exp\left(\frac{\eta}{2}\left(\left(R - \tilde{R}\right) - (r_T - \tilde{r}_T)\right)\right)\right] \\ &\leq \left(\mathbb{E}\left[\exp\left(\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right]\mathbb{E}\left[\exp\left(\eta\left(\left(R - \tilde{R}\right) - (r_T - \tilde{r}_T)\right)\right)\right]\right)^{1/2} \\ &\leq \left(\mathbb{E}\left[\exp\left(\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right]\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|\left(R - \tilde{R}\right)(\boldsymbol{\theta}) - (r_T - \tilde{r}_T)(\boldsymbol{\theta})\right|\right)\right]\right)^{1/2}\ .\end{aligned} \tag{22}$$

Observe now that $R(\boldsymbol{\theta}) = \mathbb{E}\left[r_T(\boldsymbol{\theta}\,|X)\right]$ and $\tilde{R}(\boldsymbol{\theta}) = \mathbb{E}[\tilde{r}_T(\boldsymbol{\theta}|\tilde{X})]$. Jensen's inequality for the exponential function gives that

$$\begin{aligned}\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|R(\boldsymbol{\theta}) - \tilde{R}(\boldsymbol{\theta})\right|\right) &\leq \exp\left(\eta\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}\,|X) - \tilde{r}_T\left(\boldsymbol{\theta}\,|\tilde{X}\right)\right|\right]\right) \\ &\leq \mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}\,|X) - \tilde{r}_T\left(\boldsymbol{\theta}\,|\tilde{X}\right)\right|\right)\right]\ .\end{aligned} \tag{23}$$

From (23) we see that

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|\left(R - \tilde{R}\right)(\boldsymbol{\theta}) - (r_T - \tilde{r}_T)(\boldsymbol{\theta})\right|\right)\right] \\ &\quad\leq \mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|R(\boldsymbol{\theta}) - \tilde{R}(\boldsymbol{\theta})\right|\right)\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}\,|X) - \tilde{r}_T\left(\boldsymbol{\theta}\,|\tilde{X}\right)\right|\right)\right] \\ &\quad\leq \left(\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}\,|X) - \tilde{r}_T\left(\boldsymbol{\theta}\,|\tilde{X}\right)\right|\right)\right]\right)^2\ .\end{aligned} \tag{24}$$

Combining (22) and (24) we obtain

$$\mathbb{E}\left[\exp\left(\frac{\eta}{2}\left(R - r_T\left(\cdot\,|X\right)\right)\right)\right] \leq \left(\mathbb{E}\left[\exp\left(\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right]\right)^{1/2}$$
$$\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T\left(\boldsymbol{\theta}\,|X\right) - \tilde{r}_T\left(\boldsymbol{\theta}\,\left|\tilde{X}\right.\right)\right|\right)\right] . \tag{25}$$

Let $L_{\eta,T,C} = \log((\mathbb{E}[\exp(\eta(\tilde{R}-\tilde{r}_T))])^{1/2}\mathbb{E}[\exp(\eta\sup_{\boldsymbol{\theta}\in\Theta_T}|r_T(\boldsymbol{\theta}\,|X)-\tilde{r}_T(\boldsymbol{\theta}\,|\tilde{X})|)])$. Remark that the left term of (25) is equal to the integral of the expression enclosed in brackets with respect to the measure $\pi_0 \times \pi_T$. Changing η by 2η and thanks to Lemma 1 we get

$$\pi_0\left[\exp\left(\sup_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left(\eta\rho[R - r_T\left(\cdot\,|X\right)] - \mathcal{K}\left(\rho,\pi_T\right)\right)\right)\right] \leq \exp\left(L_{2\eta,T,C}\right) .$$

Markov's inequality implies that for all $\varepsilon > 0$, with π_0- probability at least $1-\varepsilon$

$$\sup_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left(\eta\rho\left[R - r_T\left(\cdot\,|X\right)\right] - \mathcal{K}\left(\rho,\pi_T\right)\right) - \log\left(\frac{1}{\varepsilon}\right) - L_{2\eta,T,C} \leq 0 .$$

Hence, for any $\pi_T \in \mathcal{M}_+^1\left(\Theta_T\right)$ and $\eta > 0$, with π_0- probability at least $1-\varepsilon$, for all $\rho \in \mathcal{M}_+^1\left(\Theta_T\right)$

$$\rho\left[R - r_T\left(\cdot\,|X\right)\right] - \frac{1}{\eta}\mathcal{K}\left(\rho,\pi_T\right) - \frac{1}{\eta}\log\left(\frac{1}{\varepsilon}\right) - \frac{L_{2\eta,T,C}}{\eta} \leq 0 . \tag{26}$$

By setting $\rho = \pi_T\{-\eta r_T\left(\cdot\,|X\right)\}$ and relying on Lemma 1, we have

$$\begin{aligned}\mathcal{K}\left(\pi_T\left\{-\eta r_T\right\},\pi_T\right) &= \pi_T\left\{-\eta r_T\right\}\left[\log\frac{\mathrm{d}\pi_T\left\{-\eta r_T\right\}}{\mathrm{d}\pi_T}\right] = \pi_T\left\{-\eta r_T\right\}\left[\log\frac{\exp\left(-\eta r_T\right)}{\pi_T\left[\exp\left(-\eta r_T\right)\right]}\right]\\ &= \pi_T\left\{-\eta r_T\right\}\left[-\eta r_T\right] - \log\left(\pi_T\left[\exp\left(-\eta r_T\right)\right]\right)\\ &= \pi_T\left\{-\eta r_T\right\}\left[-\eta r_T\right] + \inf_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left\{\rho\left[\eta r_T\right] + \mathcal{K}\left(\rho,\pi_T\right)\right\}\end{aligned}$$

Using (26) with $\rho = \pi_T\{-\eta r_T\left(\cdot\,|X\right)\}$ it follows that, with π_0- probability at least $1-\varepsilon$,

$$\pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}[R] \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left\{\rho\left[r_T\left(\cdot\,|X\right)\right] + \frac{\mathcal{K}\left(\rho,\pi_T\right)}{\eta}\right\} + \frac{\log\left(\frac{1}{\varepsilon}\right)}{\eta} + \frac{L_{2\eta,T,C}}{\eta} .$$

To upper bound $\rho[r_T(\cdot|X)]$ we use an upper bond on $\rho\left[r_T(\cdot|X) - R\right]$. We obtain an inequality similar to (26) with $\rho\left[R - r_T(\cdot|X)\right]$ replaced by $\rho\left[r_T(\cdot|X) - R\right]$ and $L_{\eta,T,C}$ replaced by $L'_{\eta,T,C} = \log((\mathbb{E}[\exp(\eta(\tilde{r}_T - \tilde{R}))])^{1/2}\mathbb{E}[\exp(\eta \sup_{\boldsymbol{\theta}\in\Theta_T} |r_T(\boldsymbol{\theta}|X) - \tilde{r}_T(\boldsymbol{\theta}|\tilde{X})|)])$. This provides us another inequality satisfied with π_0- probability at least $1-\varepsilon$. To obtain a π_0- probability of the intersection larger than $1-\varepsilon$ we apply previous computations with $\varepsilon/2$ instead of ε and hence,

$$\pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}[R] \le \inf_{\rho\in\mathcal{M}^1_+(\Theta_T)}\left\{\rho\left[R\right] + \frac{2\mathcal{K}\left(\rho,\pi_T\right)}{\eta}\right\} + \frac{2\log\left(\frac{2}{\varepsilon}\right)}{\eta}$$

$$+ \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right]\right) + \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\tilde{r}_T - \tilde{R}\right)\right)\right]\right)$$

$$+ \frac{2}{\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T\left(\boldsymbol{\theta}\,|X\right) - \tilde{r}_T\left(\boldsymbol{\theta}\,\middle|\tilde{X}\right)\right|\right)\right]\right).$$

We can now proof Theorem 1.

Proof Let $\pi_{0,C}$ denote the distribution on $\mathcal{X}^{\mathbb{Z}} \times \mathcal{X}^{\mathbb{Z}}$ of the couple $(X, \tilde{X})$. Fubini's theorem and (19) of Lemma 2 imply that

$$\mathbb{E}\left[\exp\left(2\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right] = \pi_{0,C}\times\pi_T\left[\exp\left(2\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right] = \pi_T\times\pi_{0,C}\left[\exp\left(2\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right]$$
$$\le \exp\left(\frac{16\eta^2 k^2(T,C)}{T}\right). \tag{27}$$

Using (20), we analogously get

$$\mathbb{E}\left[\exp\left(2\eta\left(\tilde{r}_T - \tilde{R}\right)\right)\right] \le \exp\left(\frac{16\eta^2 k^2(T,C)}{T}\right). \tag{28}$$

Consider the set of probability measures $\{\rho_{\boldsymbol{\theta}_T,\Delta}, T \ge 2, 0 \le \Delta \le \Delta_T\} \subset \mathcal{M}^1_+\left(\Theta_T\right)$, where $\boldsymbol{\theta}_T$ is the parameter defined by Assumption (**P-4**) and $\rho_{\boldsymbol{\theta}_T,\Delta}\left(\boldsymbol{\theta}\right) \propto \pi_T\left(\boldsymbol{\theta}\right)\mathbb{1}_{B(\boldsymbol{\theta}_T,\Delta)\cap\Theta_T}\left(\boldsymbol{\theta}\right)$. Lemma 4, together with Lemma 3, (27) and (28) guarantee that for all $0 < \eta \le T/8\left(1+L_T\right)$

$$R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\right) \le \inf_{0\le\Delta\le\Delta_T}\left\{\rho_{\boldsymbol{\theta}_T,\Delta}\left[R\right] + \frac{2\mathcal{K}\left(\rho_{\boldsymbol{\theta}_T,\Delta},\pi_T\right)}{\eta}\right\} + \frac{16\eta k^2(T,C)}{T} + \frac{2\log\left(\frac{2}{\varepsilon}\right)}{\eta} +$$
$$4\varphi(T,C,2\eta). \tag{29}$$

Thanks to Assumptions (**L**) and (**P-3**), for any $T \geq 2$ and $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_T, \Delta)$

$$R(\boldsymbol{\theta}) - R(\boldsymbol{\theta}_T) \leq K\pi_0 \left[\left\| f_{\boldsymbol{\theta}}\left((Y_{t-i})_{i\geq 1}\right) - f_{\boldsymbol{\theta}_T}\left((Y_{t-i})_{i\geq 1}\right)\right\|\right] \leq K\mathcal{D}d_T^{1/2}\Delta \ . \tag{30}$$

For $T \geq 4$ Assumption (**P-4**) gives

$$\mathcal{K}(\rho_{\boldsymbol{\theta}_T,\Delta}, \pi_T) = \log\left(\frac{1}{\pi_T\left[B(\boldsymbol{\theta}_T, \Delta) \cap \Theta_T\right]}\right) \leq -n_T^{1/\gamma}\log(\Delta) - \log(\mathcal{C}_2) \ . \tag{31}$$

Plugging (30) and (31) into (29) and using again Assumption (**P-4**)

$$\begin{aligned} R\left(\hat{f}_{\eta,T}(\cdot\,|X)\right) \leq R(\boldsymbol{\theta}_T) + \inf_{0\leq\Delta\leq\Delta_T}\left\{\mathcal{E}_1 d_T^{1/2}\Delta - \frac{2n_T^{1/\gamma}\log(\Delta)}{\eta}\right\} + \frac{\mathcal{E}_2\eta(1+L_T)^2C^2}{T} \\ + \frac{\mathcal{E}_3(1+L_T)C}{\exp(A_*C)-1} + \frac{2\log\left(\frac{2}{\varepsilon}\right) - 2\log(\mathcal{C}_2)}{\eta} + \frac{\mathcal{E}_4(1+L_T)^2\eta}{T} \end{aligned} \tag{32}$$

where $\mathcal{E}_1 = K\mathcal{D}$, $\mathcal{E}_2 = 32K^2\left(A_* + \tilde{A}_*\right)^2$, $\mathcal{E}_3 = 8K\phi(A_*)A_*$ and $\mathcal{E}_4 = 32K^2\phi(A_*)$.

We upper bound d_T by $T/2$, n_T by $\log^\gamma T$ and substitute $\Delta_T = \mathcal{C}_3/T$. Since it is difficult to minimize the right term of (32) with respect to η and C at the same time, we evaluate them in certain values to obtain a convenient upper bound.

At a fixed ε, the convergence rate of $\left[2\log(2/\varepsilon) - 2\log(\mathcal{C}_2)\right]/\eta + \mathcal{E}_4(1+L_T)^2\eta/T$ is at best $\log T/T^{1/2}$, and we get it doing $\eta \propto T^{1/2}/\log T$. As $\eta \leq T/8(1+L_T)$ we set $\eta = \eta_T = T^{1/2}/(4\log T)$.

The order of the already chosen terms is $\log^3 T/T^{1/2}$, doing $C = \log T/A_*$ we preserve it. Taking into account that $R(\boldsymbol{\theta}_T) \leq \inf_{\boldsymbol{\theta}\in\Theta_T} R(\boldsymbol{\theta}) + \mathcal{C}_1\log^3 T/T^{1/2}$ the result follows.

7.2 *Proof of Proposition 1*

Considering that Assumption (**L**) holds we get

$$\left|R\left(\bar{f}_{\eta,T,n}(\cdot\,|X)\right) - R\left(\hat{f}_{\eta,T}(\cdot\,|X)\right)\right| \leq K\int_{\mathcal{X}^{\mathbb{Z}}}\left|\bar{f}_{\eta,T,n}(\mathbf{y}\,|X) - \hat{f}_{\eta,T}(\mathbf{y}\,|X)\right|\pi_0(\mathrm{d}\mathbf{y})$$

Observe that the last expression depends on $X_{1:T}$ and $\Phi_{\eta,T}(X)$. We bound the expectation to infer a bound in probability.

Tonelli's theorem and Jensen's inequality lead to

$$\nu_{\eta,T}\left[\left|R\left(\bar{f}_{\eta,T,n}\left(\cdot\,|X\right)\right)-R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\right)\right|\right]\leq$$
$$K\int_{\mathcal{X}^{\mathbb{Z}}}\int_{\mathcal{X}^{\mathbb{Z}}}\left(\int_{\Theta_T^{\mathbb{N}}}\left|\bar{f}_{\eta,T,n}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right|^2\mu_{\eta,T}\left(\mathrm{d}\boldsymbol{\phi}\,|\boldsymbol{x}\right)\right)^{1/2}\pi_0\left(\mathrm{d}\boldsymbol{y}\right)\pi_0\left(\mathrm{d}\boldsymbol{x}\right)\ . \tag{33}$$

We are then interested in upper bounding the expression under the square root. To that end, we use [16, Theorem 3.1] which implies that for any $\boldsymbol{x}$

$$\int_{\Theta_T^{\mathbb{N}}}\left|\bar{f}_{\eta,T,n}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right|^2\mu_{\eta,T}\left(\mathrm{d}\boldsymbol{\phi}\,|\boldsymbol{x}\right)\leq$$
$$\sup_{\boldsymbol{\theta}\in\Theta_T}\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{y}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right)^2\left(\frac{4}{\beta_{\eta,T}\left(\boldsymbol{x}\right)}-3\right)\left(\frac{1}{n}+\frac{2}{n^2\beta_{\eta,T}\left(\boldsymbol{x}\right)}\right)\ .$$

Plugging this on (33), using that $n\geq 1$ and that

$$\left(\left(4-3\beta_{\eta,T}\left(\boldsymbol{x}\right)\right)\left(2+\beta_{\eta,T}\left(\boldsymbol{x}\right)\right)\right)^{1/2}\leq 3\ ,$$

we obtain the following

$$\nu_{\eta,T}\left[\left|R\left(\bar{f}_{\eta,T,n}\left(\cdot\,|X\right)\right)-R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\right)\right|\right]\leq$$
$$\frac{3K}{n^{1/2}}\int_{\mathcal{X}^{\mathbb{Z}}}\frac{1}{\beta_{\eta,T}\left(\boldsymbol{x}\right)}\int_{\mathcal{X}^{\mathbb{Z}}}\sup_{\boldsymbol{\theta}\in\Theta_T}\left|f_{\boldsymbol{\theta}}\left(\boldsymbol{y}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right|\pi_0\left(\mathrm{d}\boldsymbol{y}\right)\pi_0\left(\mathrm{d}\boldsymbol{x}\right)\ .$$

The result follows from Markov's inequality.

Acknowledgements The author is specially thankful to François Roueff, Christophe Giraud, Peter Weyer-Brown and the two referees for their extremely careful readings and highly pertinent remarks which substantially improved the 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).

References

1. Alquier, P., & Li, X. (2012). Prediction of quantiles by statistical learning and application to GDP forecasting. In J.-G. Ganascia, P. Lenca, & J.-M. Petit (Eds.), *Discovery science* (Volume 7569 of Lecture notes in computer science, pp. 22–36). Berlin/Heidelberg: Springer.
2. Alquier, P., & Wintenberger, O. (2012). Model selection for weakly dependent time series forecasting. *Bernoulli, 18*(3), 883–913.
3. Andrieu, C., & Doucet, A. (1999). An improved method for uniform simulation of stable minimum phase real ARMA (p,q) processes. *IEEE Signal Processing Letters, 6*(6), 142–144.
4. Atchadé, Y. F. (2006). An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in Applied Probability, 8*(2), 235–254.
5. Audibert, J.-Y. (2004). PAC-bayesian statistical learning theory. PhD thesis, Université Pierre et Marie Curie-Paris VI.
6. Beadle, E. R., & Djurić, P. M. (1999). Uniform random parameter generation of stable minimum-phase real ARMA (p,q) processes. *IEEE Signal Processing Letters, 4*(9), 259–261.
7. Brockwell, P. J., & Davis, R. A. (2006). *Time series: Theory and methods* (Springer series in statistics). New York: Springer. Reprint of the second (1991) edition.
8. Catoni, O. (2004). *Statistical learning theory and stochastic optimization* (Volume 1851 of Lecture notes in mathematics). Berlin: Springer. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, 8–25 July 2001.
9. Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge: Cambridge University Press.
10. Coulon-Prieur, C., & Doukhan, P. (2000). A triangular central limit theorem under a new weak dependence condition. *Statistics and Probability Letters, 47*(1), 61–68.
11. Dalalyan, A. S., & Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp PAC-bayesian bounds and sparsity. *Machine Learning, 72*(1–2), 39–61.
12. Dedecker, J., Doukhan, P., Lang, G., León R, J. R., Louhichi, S., & Prieur, C. (2007). *Weak dependence: With examples and applications* (Volume 190 of Lecture notes in statistics). New York: Springer.
13. Dedecker, J., & Prieur, C. (2005). New dependence coefficients. Examples and applications to statistics. *Probability Theory and Related Fields, 132*(2), 203–236.
14. Künsch, H. R. (1995). A note on causal solutions for locally stationary AR-processes. Note from ETH Zürich, available on line here: ftp://ftp.stat.math.ethz.ch/U/hkuensch/localstat-ar.pdf.
15. Łatuszyński, K., Miasojedow, B., & Niemiro, W. (2013). Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli, 19*, 2033–2066.
16. Łatuszyński, K., & Niemiro, W. (2011). Rigorous confidence bounds for MCMC under a geometric drift condition. *Journal of Complexity, 27*(1), 23–38.
17. Leung, G., & Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory, 52*(8), 3396–3410.
18. Mengersen, K. L., & Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics, 24*(1), 101–121.
19. Moulines, E., Priouret, P., & Roueff, F. (2005). On recursive estimation for time varying autoregressive processes. *The Annals of Statistics, 33*(6), 2610–2654.
20. Rio, E. (2000). Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Academie des Sciences Paris Series I Mathematics, 330*(10), 905–908.
21. Roberts, G. O., & Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys, 1*, 20–71.