Aggregation of predictors for non stationary sub-linear processes and application to online adaptive forecasting of locally stationary time varying autoregressive processes

Christophe Giraud, François Roueff, Andres Sanchez-Perez

Aggregation of predictors for non stationary sub-linear processes and application to online adaptive forecasting of locally stationary time varying autoregressive processes

Christophe Giraud [*1], François Roueff[†2], and Andres Sanchez-Perez[‡2]

[1]Université Paris Sud ; Département de Mathématiques
[2]Institut Mines-Télécom ; Télécom ParisTech ; CNRS LTCI

April 27, 2014

Abstract

In this work, we study the problem of aggregating a finite number of predictors for non stationary sub-linear processes. We provide oracle inequalities relying essentially on three ingredients: 1) a uniform bound of the ℓ^1 norm of the time-varying sub-linear coefficients, 2) a Lipschitz assumption on the predictors and 3) moment conditions on the noise appearing in the linear representation. Two kinds of aggregations are considered giving raise to different moment conditions on the noise and more or less sharp oracle inequalities. We apply this approach for deriving an adaptive predictor for locally stationary time varying autoregressive (TVAR) processes. It is obtained by aggregating a finite number of well chosen predictors, each of them enjoying an optimal minimax rate under specific smoothness conditions on the TVAR coefficients. We show that the obtained aggregated estimator achieves a minimax rate while adapting to the unknown smoothness. To prove this result, a lower bound is established for the minimax rate of the prediction risk for the TVAR process. An important feature of this approach is that the aggregated predictor can be computed recursively and is thus applicable in an online prediction context.

1 Introduction

In many applications where high frequency data are observed, we wish to forecast the next values of this time series through an online prediction learning algorithm able to

[*]christophe.giraud@math.u-psud.fr
[†]francois.roueff@telecom-paristech.fr
[‡]andres.sanchez-perez@telecom-paristech.fr

process a large amount of data. The classical stationarity assumption on the distribution of the observations has to be weakened to take into account some smooth evolution of the environment. In order to sequentially track a time-evolving parameter from high-frequency data, the algorithms must require few operations and a low storage capacity to update the parameter estimation and the process' forecast after each new observation. The most common online methods are least mean squares (LMS), normalised least mean squares (NLMS), regularised least squares (RLS) or Kalman. All of them rely on the choice of a gradient step, a forgetting factor, or, more generally on a tuning parameter corresponding to some *a priori* on the smoothness of the time evolution of the statistical distribution of the data. To adapt automatically to this smoothness, usually unknown in practice, we propose to use an exponentially weighted aggregation of several such predictors, with various tuning parameters. We emphasize that to meet the online constraint, we cannot use methods that require a large amount of computations (such as cross validation).

The exponential weighting technique in aggregation have been parallel developed in the machine learning community (see the seminal paper [18]) and in the statistical community (see [19, 12], or more recently [9, 15]). The book [4] provides a complete survey in sequential prediction of individual sequences. It includes the exponentially weighted average forecaster. In contrast with the classical statistical assumption, the observations are not assumed to be generated by an underlying stochastic process. A more recent contribution in the same context is [16]. The link between the theory of individual sequences prediction and the classical statistical setting such as the one evoked just above is analyzed in [10]. In that dissertation the regression model is considered with both fixed and random design.

Exponential weighting has also been investigated in the case of dependent stationary data in [1]. More recently, an approach inspired from individual sequences prediction have been studied in [2] for bounded ARMA processes under some specific conditions on the (constant) ARMA coefficients.

In this contribution, we consider aggregation schemes based on exponential weights which can be computed recursively. We provide oracle inequalities applying to the aggregated predictor under the two main assumptions that 1) the observations are sublinearly depending of an independent process with possibly time varying linear coefficients and 2) the predictors to be aggregated are Lipschitz functions of the past. An important feature of our observation model is that it embeds the well known class of *local stationarity* processes. We refer to [6, 8] and the references therein for a recent general view about statistical inference for locally stationary processes. As an application, we focus on a particular locally stationary model, that of the time-varying autoregressive (TVAR) process. The estimation of the parameters generating a TVAR process is closely related to the prediction problem. The minimax rate of some recursive estimators of the TVAR coefficients is studied in [14]. To our knowledge, there is not a well-established method on the automatic choice of the gradient step when the smoothness index is unknown. We show that the proposed aggregation method provides a solution to this question, in the sense that it gives raise to a recursive adaptive minimax predictor.

The paper is organized as follows. In Section 2, we provide oracle inequalities for the aggregation of predictors under general conditions applying to non-stationary sub-

linear processes. TVAR processes are introduced in Section 3 in a non-parametric setting based on Hölder smoothness assumptions on the TVAR coefficients. A lower bound of the prediction risk is given in this setting and this result is used to show that the proposed aggregation methods achieve the minimax adaptive rate. Section 4 contains the proofs of the oracle inequalities and their application to the non-parametric TVAR setting. The proof of the lower bound of the minimax prediction risk is presented in Section 5. Two appendices complete this paper. Appendix A explains how to build non-adaptive minimax predictors which can be used in the aggregation step and Appendix B contains some postponed proofs and useful lemmas.

2 Online aggregation of predictors for non-stationary processes

2.1 General model

In this section, we consider a time series $(X_t)_{t\in\mathbb{Z}}$ admitting the following *non-stationary* sub-linear property.

(**M**-1) The process $(X_t)_{t\in\mathbb{Z}}$ satisfies

$$|X_t| \leq \sum_{j\in\mathbb{Z}} A_t(j)\, Z_{t-j}\ , \tag{2.1}$$

where $(Z_t)_{t\in\mathbb{Z}}$ is a sequence of non-negative independent random variables and $(A_t(j))_{t,j\in\mathbb{Z}}$ are non-negative coefficients such that

$$A_* := \sup_{t\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} A_t(j) < \infty\ . \tag{2.2}$$

The condition on A_* in (2.2) guarantees that, if $(Z_t)_{t\in\mathbb{Z}}$ has a uniformly bounded L^p-norm, the convergence of the infinite sum in (2.1) holds almost surely and in the L^p-sense (with both convergences defining the same limit). It follows that $(X_t)_{t\in\mathbb{Z}}$ also has uniformly bounded L^p moments. However, because the sequence $(A_t(j))_{j\in\mathbb{Z}}$ may vary with t, such condition applies for processes that may be neither weakly nor strongly stationary. The class of linear processes with time varying coefficients is such an example. In this case we have

$$X_t = \sum_{j\in\mathbb{Z}} a_t(j)\, \xi_{t-j}\ ,$$

where (ξ_t) is a sequence of centered independent random variables with unit variance and $(a_t(j))_{t,j}$ is supposed to satisfies (2.2) with $A_t(j) = |a_t(j)|$, so that (**M**-1) holds with $Z_t = |\xi_t|$. For this general class of processes, statistical inference is not easily carried out : each new observation X_t comes with a new unknown sequence $(a_t(j))_{j\in\mathbb{Z}}$. However additional assumptions on these set of sequences allows to derive and study appropriate statistical inference procedures. A sensible approach in this direction is to consider a *locally stationary* model as introduced in [5]. In this framework, the

set of sequences $\{(a_t(j))_{j\in\mathbb{Z}},\ 1 \le t \le T\}$ is controlled as $T \to \infty$ by artificially (but meaningfully) introducing a dependence in T, hence is written as $(a_{t,T}(j))_{j\in\mathbb{Z},1\le t\le T}$, and by approximating it with a set of sequences rescaled on the time interval $[0,1]$, $a(u,j)$, $u \in [0,1]$, $j \in \mathbb{Z}$, for example in the following way

$$\sup_{T\ge 1}\sup_{j\in\mathbb{Z}}\sum_{t=1}^{T}\left|a_{t,T}(j) - a(t/T, j)\right| < \infty\,.$$

Then various interesting statistical inference problems based on $X_1, \ldots, X_T$ can be tackled by assuming some smoothness on the mapping $u \mapsto a(u,j)$ and, possibly, additional assumptions on the structure of the sequence $(a(u,j))_{j\in\mathbb{Z}}$ for each $u \in [0,1]$, see [6] and the references therein. A focus on the specific TVAR model will be treated in Section 3. Let us stress, however, that our general condition (**M**-1) includes all the models treated in [6].

Our goal in this section is to derive oracle bounds for the aggregation of predictors that hold for the general model (**M**-1) with one of the two following additional assumptions on $(Z_t)_{t\in\mathbb{Z}}$.

(**N**-1) The non-negative process $(Z_t)_{t\in\mathbb{Z}}$ satisfies

$$m_p := \sup_{t\in\mathbb{Z}}\mathbb{E}\left[Z_t^p\right] < \infty\,.$$

(**N**-2) The non-negative process $(Z_t)_{t\in\mathbb{Z}}$ satisfies

$$\phi(\zeta) := \sup_{t\in\mathbb{Z}}\mathbb{E}\left[\mathrm{e}^{\zeta Z_t}\right] < \infty\,.$$

2.2 Aggregation of predictors

Let $(x_t)_{t\in\mathbb{Z}}$ be a real valued sequence. We say that $\hat{x}_t$ is a predictor of x_t if it is a measurable function of $(x_s)_{s\le t-1}$. Throughout this paper, the quality of a sequence of predictors $(\hat{x}_t)_{1\le t\le T}$ is evaluated for some $T \ge 1$ using the ℓ^2 loss averaged over the time period $\{1,\ldots,T\}$

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{x}_t - x_t)^2\,.$$

Now, given a collection of N sequences of predictors $\left\{(\hat{x}_t^{(j)})_{1\le t\le T},\ 1 \le j \le N\right\}$, we wish to sequentially derive a new predictor which predicts almost as or more accurately than the best of them. In this context the $\hat{x}_t^{(j)}$s are called expert's predictions or forecasts. Aggregating the experts amounts to compute a convex combination of them at each time t. This corresponds to choose an element of the simplex

$$\mathcal{S}_N = \left\{\boldsymbol{\nu} = (\nu_1,\ldots,\nu_N) \in \mathbb{R}_+^N : \sum_{i=1}^{N}\nu_i = 1\right\}. \tag{2.3}$$

Given a finite collection of predictors $\left\{(\hat{x}_t^{(j)})_{1\le t\le T},\ j=1,\dots,N\right\}$ and $\nu\in\mathcal{S}_N$, we denote the resulting aggregated predictor at time t by

$$\hat{x}_t^{[\nu]}=\sum_{j=1}^{N}\nu_j\,\hat{x}_t^{(j)}\ .$$

We consider two strategies of aggregation, which are studied in the context of bounded sequences in [4, 3]. More recent contributions and extensions can be found in [10], see also [16] for a pedagogical introduction. These strategies are sequential and online, which respectively mean that, to compute the aggregation weights at time t, only the values of $\{\hat{x}_s^{(j)},\ 1\le j\le N\}$ and x_s up to time $s=t-1$ are used and the computation can be done recursively by updating a number of quantities which does not depend on t, see Algorithm 1 detailed below.
The two considered strategies are defined trough as weighted combinations of experts

$$\hat{x}_t=\hat{x}_t^{[\widehat{\alpha}_t]}=\sum_{i=1}^{N}\widehat{\alpha}_{i,t}\hat{x}_t^{(i)},\quad 1\le t\le T\ ,$$

with specific weights $\widehat{\alpha}_{i,t}$ defined as follows.

Strategy 1: building weights from the gradient of the quadratic loss

The first strategy is defined by the following formula on the weights. For all $i=1,\dots,N$ and $t=1,\dots,T$, set

$$\widehat{\alpha}_{i,t}\ =\ \frac{\exp\left(-2\eta\sum_{s=1}^{t-1}\left(\sum_{j=1}^{N}\widehat{\alpha}_{j,s}\hat{x}_s^{(j)}-x_s\right)\hat{x}_s^{(i)}\right)}{\sum_{k=1}^{N}\exp\left(-2\eta\sum_{s=1}^{t-1}\left(\sum_{j=1}^{N}\widehat{\alpha}_{j,s}\hat{x}_s^{(j)}-x_s\right)\hat{x}_s^{(k)}\right)}\ , \tag{2.4}$$

with the convention that a sum over no element is null, so $\widehat{\alpha}_{i,1}=1/N$ for all i. The parameter $\eta>0$ is known as the *learning rate* and will be specified later.

Strategy 2: building weights from the quadratic loss

The second strategy is defined by the following formula on the weights. For all $i=1,\dots,N$ and $t=1,\dots,T$, set

$$\widehat{\alpha}_{i,t}\ =\ \frac{\exp\left(-\eta\sum_{s=1}^{t-1}\left(\hat{x}_s^{(i)}-x_s\right)^2\right)}{\sum_{k=1}^{N}\exp\left(-\eta\sum_{s=1}^{t-1}\left(\hat{x}_s^{(k)}-x_s\right)^2\right)}\ , \tag{2.5}$$

with again the convention that a sum over no element is null, so that $\widehat{\alpha}_{i,1} = 1/N$ for all i.

Algorithm 1: Recursive algorithm for Strategies 1 and 2.

parameters the learning rate η;
initialization $v_{i,1} = 1$ for $i = 1, \ldots, N$;
input the experts advices $\hat{x}_1^{(i)}$ for $i = 1, \ldots, N$;
return $\widehat{\alpha}_1 = \left(v_{i,1} / \sum_{k=1}^N v_{k,1}\right)_{i=1,\ldots,N}$ *and* $\hat{x}_1 = \hat{x}_1^{[\widehat{\alpha}_1]}$;
while input *a new* x_{t-1} *and the experts advices* $\hat{x}_t^{(i)}$ *for* $i = 1, \ldots, N$*;*
do
  for $i \leftarrow 1$ **to** N **do**
    switch *strategy* **do**
      case *1*
        $v_{i,t} = v_{i,t-1} \exp\left(-2\eta \left(\hat{x}_{t-1}^{[\widehat{\alpha}_{t-1}]} - x_{t-1}\right) \hat{x}_{t-1}^{(i)}\right)$;
      case *2*
        $v_{i,t} = v_{i,t-1} \exp\left(-\eta \left(\hat{x}_{t-1}^{(i)} - x_{t-1}\right)^2\right)$;
  return $\widehat{\alpha}_t = \left(v_{i,t} / \sum_{k=1}^N v_{k,t}\right)_{i=1,\ldots,N}$ *and* $\hat{x}_t = \hat{x}_t^{[\widehat{\alpha}_t]} = \sum_{i=1}^N \widehat{\alpha}_{i,t} \hat{x}_t^{(i)}$;

2.3 Oracle bounds

We establish oracle bounds on the average prediction error of the aggregated predictors. These bounds ensure that the error is equal to that associated with the best convex combination of the experts or with the best expert (depending on the aggregation strategy), up to two remainder terms. One remainder term depends on the number N of predictors to aggregate and the other one on the *variability* of the original process. The learning rate η can then be used to have a trade-off between these two terms.

The second remainder term indirectly depends on the variability of the experts. We control below this variability in terms of the variability of the original process by using the following Lipschitz property.

Definition 1. *Let $L = (L_s)_{s \geq 1}$ be a sequence of non-negative numbers. A predictor $\hat{x}_t$ of x_t from $(x_s)_{s \leq t-1}$ is said to be L-Lipschitz if*

$$|\hat{x}_t| \leq \sum_{s \geq 1} L_s \, |x_{t-s}| \, .$$

We more specifically consider a sequence L satisfying the following assumption.

(**L**-1) The sequence $L = (L_s)_{s \geq 1}$ satisfies

$$L_* = \sum_{j \geq 1} L_j < \infty \, . \tag{2.6}$$

We now state two upper-bounds on the mean quadratic prediction error of the aggregated predictors defined in the previous section, when the process X fulfills the sublinear property (**M**-1).

Theorem 2.1. *Assume that Assumption (**M**-1) holds. Let* $\left\{(\widehat{X}_t^{(j)})_{1\le t\le T},\ 1\le i\le N\right\}$ *be a collection of sequences of L-Lipschitz predictors with L satisfying (**L**-1).*

*(i) Assume that the noise Z fulfills (**N**-1) with* $p=4$ *and let* $\widehat{X}=(\widehat{X}_t)_{1\le t\le T}$ *denote the aggregated predictor obtained using the weights (2.4) with any* $\eta>0$*. Then, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t-X_t\right)^2\right]\le \inf_{\nu\in\mathcal{S}_N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{[\nu]}-X_t\right)^2\right] + \frac{\log N}{T\eta}+2\eta\,(1+L_*)^4\,A_*^4 m_4\,. \quad (2.7)$$

*(ii) Assume that the noise Z satisfies (**N**-1) with a given* $p\ge 2$ *and let* $\widehat{X}=(\widehat{X}_t)_{1\le t\le T}$ *denote the aggregated predictor obtained using the weights (2.5) with any* $\eta>0$*. Then, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t-X_t\right)^2\right]\le \min_{1\le i\le N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{(i)}-X_t\right)^2\right] + \frac{\log N}{T\eta}+T\,(8\eta)^{(p-2)/2}A_*^p(1+L_*)^p m_p\,. \quad (2.8)$$

*(iii) Assume that the noise Z fulfills (**N**-2) for some positive* ζ *and let* $\widehat{X}=(\widehat{X}_t)_{1\le t\le T}$ *denote the aggregated predictor obtained using the weights (2.5) with*

$$0<\eta\le\frac{1}{32}\left(\frac{\zeta}{a^*(L_*+1)}\right)^2\,, \quad (2.9)$$

where

$$a^*:=\sup_{j\in\mathbb{Z}}\sup_{t\in\mathbb{Z}}A_t(j)\ \le\ A_*\,. \quad (2.10)$$

Then, for any λ *such that* $(32\eta)^{1/2}\le\lambda\le\zeta/(a^*(L_*+1))$*, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t-X_t\right)^2\right]\le \min_{1\le i\le N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{(i)}-X_t\right)^2\right] + \frac{\log N}{T\eta}+\frac{T\,\mathrm{e}^{-\lambda/(8\eta)^{1/2}}}{8\eta}\,(\phi(\zeta))^{\lambda A_*(L_*+1)/\zeta}\,. \quad (2.11)$$

The proof can be found in Section 4.2.

Remark 1. The bound (2.7) (resp. (2.8) and (2.11)) is explicit in the sense that all the constants appearing in them are directly derived from those appearing in Assumptions (**M**-1), (**L**-1) and (**N**-1) (resp. (**N**-1) and (**N**-2)).

Remark 2. To minimize the sum of the two terms appearing in the second line of (2.7), the optimal η is

$$\eta = \frac{1}{(2m_4)^{1/2}\ (1+L_*)^2\, A_*^2} \left(\frac{\log N}{T}\right)^{1/2} , \tag{2.12}$$

which gives

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \inf_{\nu\in\mathcal{S}_N} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t^{[\nu]} - X_t\right)^2\right] + C_1 \left(\frac{\log N}{T}\right)^{1/2} , \tag{2.13}$$

with $C_1 = 2\,(2m_4)^{1/2}\ (1+L_*)^2\, A_*^2$.

Remark 3. To minimize the sum of the two terms appearing in the second line of (2.8), the optimal η is

$$\eta = \frac{1}{8^{(p-2)/p}\,(1+L_*)^2\, A_*^2 m_p^{2/p}} \left(\frac{\log N}{T^2}\right)^{2/p} , \tag{2.14}$$

which gives

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \inf_{\nu\in\mathcal{S}_N} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t^{[\nu]} - X_t\right)^2\right] + C_2 \left(\frac{\log^{p-2} N}{T^{p-4}}\right)^{1/p} , \tag{2.15}$$

with $C_2 = 2\,8^{(p-2)/p}\,(1+L_*)^2\, A_*^2 m_p^{2/p}$. We observe that if $p > 8$, the bound (2.15) improves that in (2.13) by replacing $((\log N)/T)^{1/2}$ by $(\log^{p-2} N/T^{p-4})^{1/p}$.

Remark 4. Minimizing the sum of the two terms appearing in the second line of (2.11) is a bit more involved, since it depends both on η and λ. The constraint (2.9) bounds η away of infinity. If η remains bounded away from zero, then $\lambda \ge (32\eta)^{1/2}$ is bounded away from zero and infinity, and the second line of (2.11) is of order at least $O\left(T^{-1}\ \log N + T\right)$, which is always worst than the bound obtained in (2.13) under much weaker assumptions. The conclusion of this reasoning is that we should let η be small enough to improve this aggregation bound. Now for η small enough, the optimal λ is the largest allowed one, that is, $\lambda = \zeta/(a^*(L_*+1))$. To have a simpler expression, let us take the smaller

$$\lambda = \zeta/(A^*(L_*+1)) , \tag{2.16}$$

in which case (2.11) holds for any $0 < \eta \le \lambda^2/32$ and the second line of (2.11) simplifies into

$$\frac{\log N}{T\eta} + \phi(\zeta)\,\frac{T\,\mathrm{e}^{-\lambda/(8\eta)^{1/2}}}{8\eta} . \tag{2.17}$$

The sum (2.17) is still difficult to minimize in η exactly but a satisfying bound is obtained by equaling the two terms of the sum. Yet, we must also take into account the constraint $0 < \eta \le \lambda^2/32$, so we set

$$\eta = \frac{\lambda^2}{8}\ \left(\max\left\{2, \log\left(\frac{T^2\phi(\zeta)}{8\log N}\right)\right\}\right)^{-2} . \tag{2.18}$$

With our choices (2.16) and (2.18) for λ and η, the bound (2.11) finally ensures

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \leq \min_{1\leq i\leq N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{(i)} - X_t\right)^2\right] + C_3\frac{\log N}{T}\left(\max\left\{2, \log\left(\frac{T^2\phi(\zeta)}{8\log N}\right)\right\}\right)^2 \quad (2.19)$$

with $C_3 = 16A_*^2(1+L_*)^2\zeta^{-2}$. We note that the bound (2.19) improves that in (2.13) by replacing $((\log N)/T)^{1/2}$ by its square, up to a logarithmic factor at most of order $(\log T)^2$. The bound (2.19) also improves that in (2.15) for any $p \geq 2$.

3 Time-varying autoregressive (TVAR) model

3.1 Non-parametric TVAR model

3.1.1 Vector norms and Hölder smoothness norms

We introduce some preliminary notation before defining the model. Throughout this article, vectors are denoted using boldface symbols and $|\mathbf{x}|$ denotes the Euclidean norm of $\mathbf{x}$, $|\mathbf{x}| = (\sum_i |x_i|^2)^{1/2}$. We will also use the ℓ^1-norm $|\mathbf{x}|_1 = \sum_i |x_i|$.
For $\beta \in (0,1]$ and an interval $I \subseteq \mathbb{R}$, the $\beta-$Hölder semi-norm of a function $\mathbf{f} : I \to \mathbb{R}^d$ is defined by

$$|\mathbf{f}|_{\Lambda,\beta} = \sup_{0<|s-s'|<1}\frac{|\mathbf{f}(s) - \mathbf{f}(s')|}{|s-s'|^\beta} .$$

This semi-norm is extended to any $\beta > 0$ as follows. Let $k \in \mathbb{N}$ and $\alpha \in (0,1]$ be such that $\beta = k + \alpha$. If $\mathbf{f}$ is k times differentiable on I, we define

$$|\mathbf{f}|_{\Lambda,\beta} = |\mathbf{f}^{(k)}|_{\Lambda,\alpha} ,$$

and $|\mathbf{f}|_{\Lambda,\beta} = \infty$ otherwise. We consider the case $I = (-\infty, 1]$. For $R > 0$ and $\beta > 0$, the $(\beta, R)-$ Hölder ball is denoted by

$$\Lambda_d(\beta, R) = \left\{\mathbf{f} : (-\infty,1] \to \mathbb{R}^d, \text{ such that } |\mathbf{f}|_{\Lambda,\beta} \leq R\right\} .$$

3.1.2 TVAR parameters in rescaled time

The idea of using a rescaled time with the sample size T for the TVAR parameters goes back from [5]. Since then, it has always been a central example of locally stationary linear processes. In this setting, the time varying autoregressive coefficients and variance which generate the observations $X_{t,T}$ for $1 \leq t \leq T$ are represented by functions from $[0,1]$ to $\mathbb{R}^d$ and form $[0,1]$ to $\mathbb{R}_+$ respectively. The definition sets of these functions are extended to $(-\infty, 1]$ in the following definition.

Definition 2 (TVAR model)**.** *Let $d \geq 1$. Let $\theta_1, \ldots, \theta_d$ and σ be functions defined on $(-\infty, 1]$ and $(\xi_t)_{t\in\mathbb{Z}}$ be a sequence of i.i.d. random variables with zero mean and unit*

variance. For any $T \geq 1$, we say that $(X_{t,T})_{t\leq T}$ is a TVAR process with time varying parameters $\theta_1, \dots, \theta_d, \sigma^2$ sampled at frequency T^{-1} and normalized innovations (ξ_t) if the two following assertions hold.

(i) The process X fulfills the time varying autoregressive equation

$$X_{t,T} = \sum_{j=1}^{d} \theta_j \left(\frac{t-1}{T}\right) X_{t-j,T} + \sigma\left(\frac{t}{T}\right)\xi_t \quad \text{for } -\infty < t \leq T\,. \tag{3.1}$$

(ii) The sequence $(X_{t,T})_{t\leq T}$ is bounded in probability,

$$\lim_{M\to\infty} \sup_{-\infty<t\leq T} \mathbb{P}(|X_{t,T}| > M) = 0\,.$$

This definition extends the usual definition of TVAR processes, where the time-varying parameters $\theta_1, \dots, \theta_d$ and σ^2 are assumed to be constant on $\mathbb{R}_-$, see e.g. [5, Page 144]. The TVAR model is generally used for the sample $(X_{t,T})_{1\leq t\leq T}$. The definition of the process for negative times t can be seen as a way to define initial conditions for $X_{1-d,T}, \dots, X_{0,T}$, which are then sufficient to compute $(X_{t,T})_{1\leq t\leq T}$ by iterating (3.1). However, in the context of prediction, it can be useful to consider predictors $\widehat{X}_{t,T}$ which may rely on historical data $X_{s,T}$ arbitrarily far away in the past, that is, with s tending to $-\infty$. To cope with this situation, our definition of the TVAR process $(X_{t,T})$ holds for all time indices $-\infty < t \leq T$ and we use the following definition for predictors.

Definition 3 (Predictor)**.** *For all $1 \leq t \leq T$, we say that $\widehat{X}_{t,T}$ is a predictor of $X_{t,T}$ if it is $\mathcal{F}_{t-1,T}$-measurable, where*

$$\mathcal{F}_{t,T} = \sigma\left(X_{s,T},\ s = t, t-1, t-2, \dots\right) \tag{3.2}$$

is the σ-field generated by $(X_{s,T})_{s\leq t}$. For any $T \geq 1$, we denote by $\mathcal{P}_T$ the set of sequences $\widehat{X}_T = (\widehat{X}_{t,T})_{1\leq t\leq T}$ of predictors for $(X_{t,T})_{1\leq t\leq T}$, that is, the set of all processes $\widehat{X}_T = (\widehat{X}_{t,T})_{1\leq t\leq T}$ adapted to the filtration $(\mathcal{F}_{t-1,T})_{1\leq t\leq T}$.

In practice, this general framework allows to use data with possibly long available history, although the prediction is only considered on time indices $t = 1, \dots, T$. Of course, this definition also includes the case where the predictor $\widehat{X}_{t,T}$ only depend on $(X_{s,T})_{1\leq s\leq t-1}$. Having both situations in the same framework may appear to be confusing at first. It is important to note that, in contrast with the usual stationary situation, having observed the process $X_{s,T}$ for infinitely many s's in the past (for all $s \leq t-1$) is not determining for deriving a predictor of $X_{t,T}$, since observations far away in the past may have a completely different statistical behavior.

3.1.3 Stability conditions

Next proposition proves that under standard stability conditions on the time-varying parameters $\theta_1, \dots \theta_d$ and σ^2. Condition (ii) in Definition 2 ensures the existence and

uniqueness of the solution of Eq. (3.1) for $t \leq 0$ (and thus for all $t \leq T$). We define the time-varying autoregressive polynomial by

$$\boldsymbol{\theta}(z;u) \quad = \quad 1 - \sum_{j=1}^{d} \theta_j(u) z^j \ .$$

Let us denote, for any $\delta > 0$,

$$s_d(\delta) = \left\{ \boldsymbol{\theta} : (-\infty, 1] \to \mathbb{R}^d, \boldsymbol{\theta}(z;u) \neq 0, \forall |z| < \delta^{-1}, u \in [0,1] \right\} \ . \tag{3.3}$$

Define, for $\beta > 0$, $R > 0$, $\delta \in (0,1)$, $\rho \in [0,1]$ and $\sigma_+ > 0$, the class of parameters

$$C(\beta, R, \delta, \rho, \sigma_+) = \left\{ (\boldsymbol{\theta}, \sigma) : (-\infty, 1] \to \mathbb{R}^d \times [\rho\sigma_+, \sigma_+] \ : \ \boldsymbol{\theta} \in \Lambda_d(\beta, R) \cap s_d(\delta) \right\} \ .$$

We have the following stability result.

Proposition 1. *Assume that the time varying AR coefficients $\theta_1, \ldots, \theta_d$ are uniformly continuous on $(-\infty, 1]$ and the time varying variance σ^2 is bounded on $(-\infty, 1]$. Assume moreover that there exists $\delta \in (0,1)$ such that $\boldsymbol{\theta} \in s_d(\delta)$. Then, there exists $T_0 \geq 1$ such that, for all $T \geq T_0$, there exists a unique process $(X_{t,T})_{t \leq T}$ which satisfies (i) and (ii) in Definition 2. This solution admits the linear representation*

$$X_{t,T} = \sum_{j=0}^{\infty} a_{t,T}(j)\, \sigma\left(\frac{t-j}{T}\right) \xi_{t-j}, \quad -\infty < t \leq T \ , \tag{3.4}$$

where the coefficients $(a_{t,T}(j))_{t \leq T, j \geq 0}$ satisfy that for any $\delta_1 \in (\delta, 1)$,

$$\bar{K} = \sup_{T \geq T_0} \sup_{-\infty < t \leq T} \sup_{j \geq 0} \delta_1^{-j} |a_{t,T}(j)| < \infty \ .$$

Moreover, if $(\boldsymbol{\theta}, \sigma) \in C(\beta, R, \delta, 0, \sigma_+)$ for some positive constants β, R and σ_+, then the constants T_0 and $\bar{K}$ can be chosen only depending on δ_1, δ, β, and R.

A proof of Proposition 1 is provided in Appendix B. This kind of result is classical under various smoothness assumptions on the parameters and initial conditions for $X_{1-k,T}$, $k = 1, \ldots, d$. For instance, in [8], bounded variations and a constant $\boldsymbol{\theta}$ for negative times are used for the smoothness assumption on $\boldsymbol{\theta}$ and for defining the initial conditions. The linear representation (3.4), in particular was exhibited in the seminal papers [11, 5]. We note that an important consequence of Proposition 1 is that for any $T \geq T_0$, the process $(X_{t,T})_{t \leq T}$ satisfies Assumption (**M**-1) with $Z_t = |\xi_t|$ and $A_t(j) = \left| a_{t,T}(j)\, \sigma\left((t-j)/T\right) \right|$ for $j \geq 0$. Moreover, the constant A_* in (2.2) is bounded independently of T, and we have, for all $(\boldsymbol{\theta}, \sigma) \in C(\beta, R, \delta, 0, \sigma_+)$,

$$A_* \leq \frac{\bar{K}\sigma_+}{1 - \delta_1} \ , \tag{3.5}$$

where $\bar{K} > 0$ and $\delta_1 \in (0,1)$ can be chosen only depending on δ, β, and R.

3.1.4 Main assumptions

Following Proposition 1, the TVAR model is embedded in the following assumption given an i.i.d. sequence $(\xi_t)_{t\in\mathbb{Z}}$ and constants $\delta \in (0,1)$, $\rho \in [0,1]$ $\sigma_+ > 0$, $\beta > 0$ and $R > 0$.

(**M**-2) The sequence $(X_{t,T})_{t\leq T}$ is a TVAR process with time varying standard deviation σ, time varying AR coefficients $\theta_1, \ldots, \theta_d$ and innovations $(\xi_t)_{t\in\mathbb{Z}}$, and $(\boldsymbol{\theta}, \sigma) \in C(\beta, R, \delta, \rho, \sigma_+)$.

Let ξ denote a generic random variable with the same distribution as the ξ_ts. Under Assumption (**M**-2), the distribution of $(X_{t,T})_{1-d\leq t\leq T}$ only depends on that of ξ and on the functions $\boldsymbol{\theta}$ and σ. For a given distribution ψ on $\mathbb{R}$ for ξ, we denote by $\mathbb{P}^{\psi}_{(\boldsymbol{\theta},\sigma)}$ the probability distribution of the whole sequence $(X_{t,T})_{t\leq T}$ and by $\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}$ its corresponding expectation.
The next two assumptions on the innovations are useful to prove upper bounds of the prediction error.

(**I**-1) The innovations $(\xi_t)_{t\in\mathbb{Z}}$ satisfy $m_p := \mathbb{E}\left[|\xi|^p\right] < \infty$.

(**I**-2) The innovations $(\xi_t)_{t\in\mathbb{Z}}$ satisfy $\phi(\zeta) := \mathbb{E}\left[\mathrm{e}^{\zeta|\xi|}\right] < \infty$.

The following one will be used to obtain a lower bound.

(**I**-3) The innovations $(\xi_t)_{t\in\mathbb{Z}}$ admit a density f such that

$$\kappa = \sup_{v\neq 0} v^{-2} \int f(u) \log \frac{f(u)}{f(u+v)}\, \mathrm{d}u < \infty\ .$$

Assumption (**I**-3) is standard for proving lower bounds in non-parametric regression estimation, see [17, Chapter 2]. It is satisfied by the Gaussian density with $\kappa = 1$.

3.1.5 Non-parametric setting

The setting of Definition 2 and of Assumptions derived thereafter is essentially non-parametric, since for given initial distribution ψ, the distribution of the observations $X_{1,T}, \ldots, X_{T,T}$ are determined by the unknown parameter function $(\boldsymbol{\theta}, \sigma)$. The doubly indexed $X_{t,T}$ refers to the fact that this distribution cannot be seen as a distribution on $\mathbb{R}^{\mathbb{Z}}$ marginalized on $\mathbb{R}^T$ as the usual time series setting but rather as a sequence of distributions on $\mathbb{R}^T$ indexed by T. It corresponds to the usual non-parametric approach for studying statistical inference based on this model. In this contribution, we focus on the prediction problem, which is to answer the question: for given smoothness conditions on $(\boldsymbol{\theta}, \sigma)$, what is the mean prediction error for predicting $X_{t,T}$ from its past ? The standard non-parametric approach is to answer this question in a minimax sense by determining, for a given sequence of predictors $\widehat{X}_T = (\widehat{X}_{t,T})_{1\leq t\leq T}$, the maximal risk

$$S_T(\widehat{X}_T; \psi, \beta, R, \delta, \rho, \sigma_+) = \sup_{(\boldsymbol{\theta},\sigma)} \frac{1}{T} \sum_{t=1}^{T} \left(\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)} \left[\left(\widehat{X}_{t,T} - X_{t,T} \right)^2 \right] - \sigma^2 \left(\frac{t}{T} \right) \right)\ , \tag{3.6}$$

where

(a) $\widehat{X}_T$ is assumed to belong to $\mathcal{P}_T$ as in Definition 3,

(b) the sup is taken for $(\boldsymbol{\theta}, \sigma) \in C(\beta, R, \delta, \rho, \sigma_+)$ within a smoothness class of functions,

(c) the expectation $\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}$ is the one associated to Assumption (**M**-2).

The rational for subtracting the average $\sigma^2(t/T)$ over all $1 \leq t \leq T$ in this prediction risk is that it corresponds to the best prediction risk, would the parameters $(\boldsymbol{\theta}, \sigma)$ be exactly known. We observe that dividing $X_{t,T}$ by the class parameter σ_+ amounts to take $\sigma_+ = 1$. In addition, we have

$$S_T(\widehat{X}_T; \psi, \beta, R, \delta, \rho, \sigma_+) = \sigma_+^2 \, S_T(\widehat{X}_T/\sigma_+; \psi, \beta, R, \delta, \rho, 1)\,,$$

so the prediction problem in the class $C(\beta, R, \delta, \rho, \sigma_+)$ can be reduced to the the prediction problem in the class $C(\beta, R, \delta, \rho, 1)$. Accordingly, we define the reduced minimax risk by

$$\begin{aligned} \overline{M}_T(\psi, \beta, R, \delta, \rho) &= \inf_{\widehat{X}_T \in \mathcal{P}_T} S_T(\widehat{X}_T; \psi, \beta, R, \delta, \rho, 1) \\ &= \inf_{\widehat{X}_T \in \mathcal{P}_T} \sigma_+^{-2} \, S_T(\widehat{X}_T; \psi, \beta, R, \delta, \rho, \sigma_+) \quad \text{for all } \sigma_+ > 0\,. \end{aligned} \tag{3.7}$$

In Section 3.2, we provide a lower bound of the minimax rate in the case where the smoothness class is of the form $C(\beta, R, \delta, \rho, \sigma_+)$. Then, in Section 3.3, relying on the aggregation oracle bounds of Section 2.3, we derive an upper bound with the same rate as the lower bound using the same smoothness class of the parameters. Moreover, we exhibit a predictor which do not require any knowledge about the smoothness class and which is thus minimax adaptive. In other words, it is able to adapt to the unknown smoothness of the parameters from the data. To our knowledge, such theoretical results are new for locally stationary models.

3.2 Lower bound

A lower bound on the minimax rate for the estimation error of $\boldsymbol{\theta}$ is given by [14, Theorem 4]. Clearly, a predictor

$$\widehat{X}_{t,T} = \sum_{k=1}^{d} \widehat{\boldsymbol{\theta}}_{t,T}(k) X_{t-k,T}$$

can be defined from an estimator $\widehat{\boldsymbol{\theta}}_{t,T}$, and the resulting prediction rate can be controlled using the estimation rate, see the Appendix A.1 for the details. The next theorem provides a lower bound of the minimax rate of the risk of *any* predictor of the process $\{X_{t,T}\}_{1 \leq t \leq T}$. Combining this result with Lemma 7 in the Appendix A.1 shows that a predictor obtained by (A.1) from a minimax rate estimator of $\boldsymbol{\theta}$ automatically achieves the minimax prediction rate.

Theorem 3.1. *Let $\delta \in (0,1)$, $\beta > 0$, $R > 0$ and $\rho \in [0,1]$. Suppose that Assumption (M-2) holds and assume (I-3) on the distribution ψ of the innovations. Then, we have*

$$\liminf_{T\to\infty} T^{2\beta/(1+2\beta)}\, \overline{M}_T(\psi,\beta,R,\delta,\rho) > 0\ , \tag{3.8}$$

where $\overline{M}_T$ is defined in (3.7).

The proof is postponed to Section 5.

3.3 Minimax adaptive forecasting of the TVAR process

Our minimax adaptive predictor is based on the aggregation of sufficiently many predictors, assuming that one at least among them is minimax rate. The oracle bounds found in Section 2.3 imply that the aggregated predictor is minimax rate adaptive under appropriate assumptions.

In the TVAR model (M-2), it is natural to consider L-Lipschitz predictors $(\widehat{X}_{t,T})_{1\le t\le T}$ of $(X_{t,T})_{1\le t\le T}$ with a sequence L which has support on $\{1,\dots,d\}$. Then L^* in (2.6) corresponds to the maximal ℓ^1-norm of the TVAR parameters. Since for the process itself to be stable, this norm has to be bounded independently of T, (L-1) is a quite natural assumption for the TVAR model, see Appendix A.1 for the details.

A practical advantage of the proposed procedures is that, given a set of experts that behaves well under particular smoothness assumptions, we obtain an aggregated predictor which performs almost as well as or better than the best of these experts, hence which behaves well without any prior knowledge on the smoothness of the unknown parameter. Such an adaptive property can be formally demonstrated by exhibiting an adaptive minimax rate for the aggregated estimator which coincides with the lower bound given in Theorem 3.1.

The first ingredient that we need is the following.

Definition 4 ((ψ,β)-minimax-rate predictor)**.** *Let ψ be a distribution on $\mathbb{R}$ and $\beta > 0$. We say that $\widehat{X} = (\widehat{X}_T)_{T\ge1}$ is a (ψ,β)-minimax-rate sequence of predictors if, for all $T \ge 1$, $\widehat{X}_T \in \mathcal{P}_T$ and, for all $\delta \in (0,1)$, $R > 0$, $\rho \in (0,1]$ and $\sigma_+ > 0$,*

$$\limsup_{T\to\infty} T^{2\beta/(1+2\beta)} S_T(\widehat{X}_T;\psi,\beta,R,\delta,\rho,\sigma_+) < \infty\ , \tag{3.9}$$

where S_T is defined by (3.6).

The term *minimax-rate* in this definition refers to the fact that the maximal rate in (3.9) is equal to the minimax lower bound (3.8) for the class $C(\beta,R,\delta,\rho,\sigma_+)$. To adapt to an unknown smoothness, we rely on a collection of (ψ,β)-minimax-rate predictors with β within $(0,\beta_0)$, where β_0 is the (possibly infinite) maximal smoothness index. We explain in Appendix A how to build such predictors which are moreover L-Lipschitz for some L only depending on d.

Definition 5 (Locally bounded set of ψ-minimax-rate predictors)**.** *Let ψ be a distribution on $\mathbb{R}$. We say that $\{\widehat{X}^{(\beta)}, \beta \in (0,\beta_0)\}$ is a locally bounded set of ψ-minimax-rate*

predictors if for each β, $\widehat{X}^{(\beta)}$ is a (ψ,β)-minimax-rate predictor and if moreover, for all $\delta \in (0,1)$, $R > 0$, $\rho \in (0,1]$, $\sigma_+ > 0$ and for each closed interval $I \subset (0,\beta_0)$,

$$\limsup_{T\to\infty} \sup_{\beta\in I} T^{2\beta/(1+2\beta)} S_T(\widehat{X}_T^{(\beta)};\psi,\beta,R,\delta,\rho,\sigma_+) < \infty \ ,$$

where S_T is defined by (3.6).

The following lemma shows that, given a locally bounded set of minimax-rate predictors, we can always pick a finite subset of at most $N = \lceil (\log T)^2 \rceil$ predictors among which the best one achieves the minimax rate of any unknown smoothness index.

Lemma 1. *Let ψ be a distribution on $\mathbb{R}$. Let $\beta_0 \in (0,\infty]$ and $\{\widehat{X}^{(\beta)}, \beta \in (0,\beta_0)\}$ be a corresponding locally bounded set of ψ-minimax-rate predictors. Set, for any $N \geq 1$,*

$$\beta_i = \begin{cases} (i-1)\beta_0/N & \text{if } \beta_0 < \infty, \\ (i-1)/N^{1/2} & \text{otherwise,} \end{cases} \qquad 1 \leq i \leq N \ . \tag{3.10}$$

Suppose moreover, in the case where $\beta_0 < \infty$, that $N \geq \lceil \log T \rceil$, and, in the case where $\beta_0 = \infty$, that $N \geq \lceil (\log T)^2 \rceil$. Then, we have, for all $\beta \in (0,\beta_0)$, $\delta \in (0,1)$, $R > 0$, $\rho > 0$ and $\sigma_+ > 0$,

$$\limsup_{T\to\infty} T^{2\beta/(1+2\beta)} \min_{i=1,\dots,N} S_T(\widehat{X}_T^{(\beta_i)};\psi,\beta,R,\delta,\rho,\sigma_+) < \infty \ .$$

The proof of this lemma is postponed to Section B.3 in Appendix B. Lemma 1 says that to obtain a minimax-rate predictor which adapts to an unknown smoothness index β, it is in fact sufficient to select it judiciously among $\log T$ or $(\log T)^2$ well chosen non-adaptive minimax-rate predictors. As a consequence of Theorem 2.1 and Lemma 1, we obtain an adaptive predictor by aggregating them (instead of selecting one of them), as stated in the following result.

Theorem 3.2. *Let ψ be a distribution on $\mathbb{R}$. Let $\beta_0 \in (0,\infty]$ and $\{\widehat{X}^{(\beta)}, \beta \in (0,\beta_0)\}$ be a locally bounded set of ψ-minimax-rate and L-Lipschitz predictors with L satisfying (L-1). Define $(\widehat{X}_{t,T})_{1\leq t\leq T}$ as the predictor aggregated from $\{\widehat{X}^{(\beta_i)}, 1 \leq i \leq N\}$ with N defined by*

$$N = \begin{cases} \lceil \log T \rceil & \text{if } \beta_0 < \infty, \\ \lceil (\log T)^2 \rceil & \text{otherwise,} \end{cases} \tag{3.11}$$

β_i defined by (3.10), and with weights defined according to one of the following setting depending on the assumption on ψ and β_0 :

(i) If ψ satisfies (I-1) with $p \geq 4$ and $\beta_0 \leq 1/2$, use the weights (2.4) with $\eta = \sigma_+^{-2}(\log(\lceil \log T \rceil)/T)^{1/2}$,

(ii) If ψ satisfies (I-1) with $p \geq 2$ and $\beta_0 \leq (p-4)/8$, use the weights (2.5) with $\eta = \sigma_+^{-2}(\log(\lceil \log T \rceil)/T^2)^{2/p}$,

(iii) If ψ satisfies (I-2), use the weights (2.5) with $\eta = \sigma_+^{-2}(\log T)^{-3}$.

Then, we have, for any $\beta \in (0, \beta_0)$, $\delta \in (0,1)$, $R > 0$, $\rho \in (0,1]$ *and* $\sigma_+ > 0$,

$$\limsup_{T\to\infty} T^{2\beta/(1+2\beta)} S_T(\widehat{X}_T; \psi, \beta, R, \delta, \rho, \sigma_+) < \infty \,. \tag{3.12}$$

The proof of this theorem is postponed to Section 4.3.

Remark 5. The limitation to $\beta_0 \le 1/2$ in (i) under Assumption (I-1) for ψ follows from the factor $((\log N)/T)^{1/2}$ obtained in the oracle inequality (2.7) of Theorem 2.1 after optimizing in η, see (2.13). If $p > 8$ this restriction is weakened to $\beta_0 \le (p-4)/8$ in (ii) taking into account the factor $((\log^{p-2} N)/T^{p-4})^{1/p}$ obtained in the oracle inequality (2.8) of Theorem 2.1 after optimizing in η, see (2.15). In the last case, the limitation of β_0 drops when applying the oracle inequality (2.11) of the same theorem. However a stronger condition on ψ is then required.

4 Proofs of the upper bounds

4.1 preliminary results

We start with a lemma which gathers useful adaptations of well known inequalities applying to the aggregation of deterministic predicting sequences.

Lemma 2. *Let* $(x_t)_{1\le t\le T}$ *be a real valued sequence and* $\left\{(\hat{x}_t^{(i)})_{1\le t\le T}, 1 \le i \le N\right\}$ *be a collection of predicting sequences. Define* $(\hat{x}_t)_{1\le t\le T}$ *as the sequence of aggregated predictors obtained from this collection with the weights (2.4). Then, for any* $\eta > 0$, *we have*

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{x}_t - x_t)^2 \le \inf_{\nu\in\mathcal{S}_N} \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{i=1}^{N}\nu_i\hat{x}_t^{(i)} - x_t\right)^2 + \frac{\log N}{T\eta} + 2\frac{\eta}{T}\sum_{t=1}^{T}\max_{1\le i\le N}\left|\hat{x}_t^{(i)}\right|^2\left(\max_{1\le i\le N}\left|\hat{x}_t^{(i)}\right| + |x_t|\right)^2 . \tag{4.1}$$

Define now $(\hat{x}_t)_{1\le t\le T}$ *as the sequence of aggregated predictors obtained with the weights (2.5). Then, for any* $\eta > 0$ *and* $p \ge 2$ *we have*

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{x}_t - x_t)^2 \le \min_{i=1,\dots,N}\frac{1}{T}\sum_{t=1}^{T}\left(\widetilde{x}_t^{(i)} - x_t\right)^2 + \frac{\log N}{T\eta} + (8\eta)^{(p-2)/2} y_T^p \,, \tag{4.2}$$

where

$$y_T = \max_{1\le t\le T}\left(|x_t| + \max_{1\le i\le N}\left|\widetilde{x}_t^{(i)}\right|\right). \tag{4.3}$$

Furthermore, for any positive constants η *and* λ *such that* $\eta \le \lambda^2/32$, *we have*

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{x}_t - x_t)^2 \le \min_{i=1,\dots,N}\frac{1}{T}\sum_{t=1}^{T}\left(\hat{x}_t^{(i)} - x_t\right)^2 + \frac{\log N}{T\eta} + \frac{e^{-\lambda/(8\eta)^{1/2}}}{8\eta}e^{\lambda y_T} \,. \tag{4.4}$$

Proof. With weights defined by (2.4), by slightly adapting [16, Theorem 1.7], we have that

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{x}_t - x_t)^2 - \inf_{\nu\in\mathcal{S}_N}\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{i=1}^{N}\nu_i\hat{x}_t^{(i)} - x_t\right)^2 \quad\leq\quad \frac{\log N}{T\eta} + \frac{\eta}{8T}\,s_T^*\,,$$

where $s_T^* = \sum_{t=1}^{T} s_t^2$ and $s_t = 2\max_{1\leq i\leq N}|2(\sum_{j=1}^{N}\widehat{\alpha}_{j,t}\hat{x}_t^{(j)} - x_t)\hat{x}_t^{(i)}|$. The bound (4.1) follows by using that that $\{\widehat{\alpha}_{i,t}\}_{1\leq i\leq N}$ is in the simplex $\mathcal{S}_N$ defined in (2.3).
We now prove (4.2). Using the same arguments as in [3, Proposition 2.2.1.], the aggregation (2.5) satisfies

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{x}_t - x_t)^2\,1_{\{y_T\leq 1/(8\eta)^{1/2}\}} \leq \min_{i=1,\dots,N}\frac{1}{T}\sum_{t=1}^{T}\left(\hat{x}_t^{(i)} - x_t\right)^2 + \frac{\log N}{T\eta}\,. \tag{4.5}$$

We bound the indicator function of $\left\{y_T > 1/(8\eta)^{1/2}\right\}$ by $(y_T(8\eta)^{1/2})^{p-2}$ and thus, for all $t = 1,\dots,T$,

$$(\hat{x}_t - x_t)^2\,1_{\{y_T>1/(8\eta)^{1/2}\}} \leq y_T^p(8\eta)^{(p-2)/2}\,.$$

Taking the average over $t = 1,\dots,T$ and summing with (4.5), we get the bound (4.2).
The bound (4.4) is obtained by following a similar idea. For all $t = 1,\dots,T$, we have for $\eta > 0$

$$(\hat{x}_t - x_t)^2 \leq y_T^2 \leq \frac{1}{\mathrm{e}^2 8\eta}\,\mathrm{e}^{2(8\eta)^{1/2}\,y_T}\,.$$

Bounding the indicator function of $\left\{y_T > 1/(8\eta)^{1/2}\right\}$ by $\mathrm{e}^{\gamma y_T}\mathrm{e}^{-\gamma/(8\eta)^{1/2}}$, with $\gamma = \lambda - 2(8\eta)^{1/2} \geq 0$ we get

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{x}_t - x_t)^2\,1_{\{y_T>1/(8\eta)^{1/2}\}} \leq \frac{1}{8\eta}\,\mathrm{e}^{\lambda y_T}\mathrm{e}^{-\lambda/(8\eta)^{1/2}}\,.$$

Summing with (4.5), we get the bound (4.4). □

4.2 Proof of Theorem 2.1

Case (i). Applying (4.1) in Lemma 2 with $\mathbb{E}[\inf\dots] \leq \inf\mathbb{E}[\dots]$, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \leq \inf_{\nu\in\mathcal{S}_N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\sum_{i=1}^{N}\nu_i\widehat{X}_t^{(i)} - X_t\right)^2\right]\\ + \frac{\log N}{T\eta} + 2\frac{\eta}{T}\sum_{t=1}^{T}\mathbb{E}\left[\max_{1\leq i\leq N}\left|\widehat{X}_t^{(i)}\right|^2\left(\max_{1\leq i\leq N}\left|\widehat{X}_t^{(i)}\right| + |X_t|\right)^2\right]\,.\end{aligned} \tag{4.6}$$

Using that the predictors are L-Lipschitz and the process $(X_t)_{t\in\mathbb{Z}}$ satisfies (**M**-1), we have, for all $1 \leq t \leq T$,

$$\begin{aligned}|X_t| + \max_{1\leq i\leq N}\left|\widehat{X}_t^{(i)}\right| &\leq \sum_{j\in\mathbb{Z}}A_t(j)\,Z_{t-j} + \sum_{s\geq 1}\sum_{j\in\mathbb{Z}}L_s\,A_{t-s}(j)\,Z_{t-s-j}\\ &\leq \sum_{j\in\mathbb{Z}}B_t(j)Z_{t-j},\end{aligned} \tag{4.7}$$

where

$$B_t(j) = A_t(j) + \sum_{s\geq 1} L_s\, A_{t-s}(j-s)\ .$$

Applying the Minkowski inequality together with (4.7), (2.2) and (2.6), we obtain, for all $1 \leq t \leq T$,

$$\mathbb{E}\left[\max_{1\leq i\leq N}\left|\widehat{X}_t^{(i)}\right|^2\left(\max_{1\leq i\leq N}\left|\widehat{X}_t^{(i)}\right|+|X_t|\right)^2\right] \leq \mathbb{E}\left[\left(\sum_{j\in\mathbb{Z}} B_t(j)Z_{t-j}\right)^4\right] \leq A_*^4(1+L_*)^4\ \sup_{t\in\mathbb{Z}}\mathbb{E}[Z_t^4]\ .$$

Since the process Z fulfills (**N**-1) with $p = 4$, plugging this bound in (4.6) we obtain (2.7).

Case (ii). We use (4.2) in Lemma 2 and since it is assumed that $p \geq 2$, we get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_{t,T}-X_{t,T}\right)^2\right] \leq \min_{i=1,\dots,N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_{t,T}^{(i)}-X_{t,T}\right)^2\right]+\frac{\log N}{T\eta} + (8\eta)^{(p-2)/2}\mathbb{E}\left[Y_T^p\right]\ , \quad (4.8)$$

where $Y_T = \max_{1\leq t\leq T}\left(|X_t| + \max_{1\leq i\leq N}\left|\widehat{X}_t^{(i)}\right|\right)$. Observe that

$$\mathbb{E}\left[Y_T^p\right] \leq \sum_{t=1}^{T}\mathbb{E}\left[\left(|X_t|+\max_{1\leq i\leq N}\left|\widehat{X}_t^{(i)}\right|\right)^p\right]\ .$$

Using the Minkowski inequality, (4.7) and Assumption (**N**-2)

$$\mathbb{E}\left[Y_T^p\right] \leq \sum_{t=1}^{T}\left(\sum_{j\in Z} B_t(j)\left(\mathbb{E}\left[Z_{t-j}^p\right]\right)^{1/p}\right)^p \leq A_*^p(1+L_*)^p\, T\, \sup_{t\in Z}\mathbb{E}[Z_t^p]\ .$$

Using thins bound with (**N**-1) and (4.8), we obtain (2.8).
Case (iii). To obtain (2.11), we now use (4.4) in Lemma 2 and get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_{t,T}-X_{t,T}\right)^2\right] \leq \min_{i=1,\dots,N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_{t,T}^{(i)}-X_{t,T}\right)^2\right]+\frac{\log N}{T\eta} + \frac{\mathrm{e}^{-\lambda/(8\eta)^{1/2}}}{8\eta}\mathbb{E}\left[\mathrm{e}^{\lambda Y_T}\right]\ . \quad (4.9)$$

We now use Assumption (**N**-2). Since $B_t(j) \leq a^*(1+L_*)$ for all $j, t \in \mathbb{Z}$ and

$$\sum_{j\in\mathbb{Z}} B_t(j) \leq A_*(1+L_*),$$

Jensen's inequality and (4.7) gives that, for any $\lambda \le \zeta/(a^*(1+L_*))$,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda Y_T}\right] &\le \sum_{t=1}^{T} \mathbb{E}\left[e^{\lambda\left(|X_t|+\max_{1\le i\le N}|\widehat{X}_t^{(i)}|\right)}\right] \\ &\le \sum_{t=1}^{T}\prod_{j\in\mathbb{Z}} \mathbb{E}\left[e^{\lambda B_t(j) Z_{t-j}}\right] \\ &\le \sum_{t=1}^{T}\prod_{j\in\mathbb{Z}} (\phi(\zeta))^{\lambda B_t(j)/\zeta} \le T\,(\phi(\zeta))^{\lambda A_*(1+L_*)/\zeta}\,.\end{aligned}$$

Combining this bound with (4.9) gives (2.11). The proof of Theorem 2.1 is complete.

4.3 Application to the TVAR process: proof of Theorem 3.2

Theorem 3.2 is an application of Theorem 2.1 to the aggregation of minimax predictors for the TVAR model (**M**-2).

We first note that Proposition 1 shows that, for T large enough the TVAR model (**M**-2) satisfies (**M**-1) with A_* bounded independently of T as in (3.5) and $Z_t = |\xi_t|$ for all $t \in \mathbb{Z}$. Hence Assumptions (**I**-1) and (**I**-2) respectively imply (**N**-1) and (**N**-2).

This shows that Theorem 2.1 applies under the assumptions of Theorem 3.2 and that the constants A_* and a^* appearing in (2.7), (2.9) and (2.11) can be replaced by $\bar{K}\sigma_+/(1-\delta_1)$ and $\bar{K}\sigma_+$, respectively, where $\bar{K} > 0$ and $\delta_1 \in (0,1)$ can be chosen only depending on δ, β, and R.

On the other hand, Lemma 1 shows that, under the given assumptions on the predictors and with the given choices of N, the smallest prediction risk among the selected predictors, achieves a rate $T^{-2\beta/(1+2\beta)}$ for some positive constant C only depending on β, δ, $R > 0$, ρ and ψ. Hence, we get with Theorem 2.1 that

$$\limsup_{T\to\infty} T^{2\beta/(1+2\beta)} S_T(\widehat{X}_T;\psi,\beta,R,\delta,\rho,\sigma_+) \le C + \limsup_{T\to\infty} T^{2\beta/(1+2\beta)}\mathcal{R}(N,T)\,, \tag{4.10}$$

where C is a positive constant and $\mathcal{R}(N,T)$ is a remainder term which, in the setting (i) in Theorem 3.2, is given by

$$\mathcal{R}(N,T) = \frac{\log N}{T\eta} + 2\eta\,(1+L_*)^4\, m_4\, \frac{\bar{K}^4\sigma_+^4}{(1-\delta_1)^4}\,, \tag{4.11}$$

in the setting (ii), is given by

$$\mathcal{R}(N,T) = \frac{\log N}{T\eta} + T\,(8\eta)^{(p-2)/2}(1+L_*)^p m_p\, \frac{\bar{K}^p\sigma_+^p}{(1-\delta_1)^p}\,, \tag{4.12}$$

and, in the setting (iii), is given by

$$\mathcal{R}(N,T) = \frac{\log N}{T\eta} + \frac{T\,e^{-\lambda/(8\eta)^{1/2}}}{8\eta}\,(\phi(\zeta))^{\lambda\,\bar{K}\sigma_+(L_*+1)/(\zeta(1-\delta_1))}\,, \tag{4.13}$$

provided that η and λ satisfy

$$0 < \eta \leq \frac{1}{32}\left(\frac{\zeta}{\bar{K}\sigma_+(L_*+1)}\right)^2 , \quad \text{and} \quad (32\eta)^{1/2} \leq \lambda \leq \zeta/(\bar{K}\sigma_+(L_*+1)) . \tag{4.14}$$

Replacing η and N in (4.11) as given by (i) and (3.11), we get

$$\sigma_+^{-2}\,\mathcal{R}(N,T) \leq \left(\frac{\log\lceil\log T\rceil}{T}\right)^{1/2}\left(1+2\,(1+L_*)^4\,m_4\,\frac{\bar{K}^4}{(1-\delta_1)^4}\right) .$$

Hence, using that $\beta < \beta_0 \leq 1/2$, this upper bound is negligible with respect to $T^{-2\beta/(2\beta+1)}$ and, with (4.10), we get (3.12).
Analogously, we replace η and N in (4.12) as given by (ii) and (3.11), we get

$$\sigma_+^{-2}\,\mathcal{R}(N,T) \leq \frac{(\log\lceil\log T\rceil)^{(p-2)/2}}{T^{(p-4)/p}}\left(1+8^{(p-2)/2}\,(1+L_*)^p\,m_p\,\frac{\bar{K}^p}{(1-\delta_1)^p}\right) .$$

Since $\beta < \beta_0 \leq (p-4)/8$, this upper bound is negligible with respect to $T^{-2\beta/(2\beta+1)}$ and, with (4.10), we get (3.12).
Using the specific form of η in (iii) and choosing λ equal to the upper bound of the given condition (4.14), we get that, in the setting (iii),

$$\mathcal{R}(N,T) = \frac{\log N}{T\eta} + \frac{T\,\mathrm{e}^{-\zeta/(\bar{K}\sigma_+(L_*+1)(8\eta)^{1/2})}}{8\eta}\,(\phi(\zeta))^{1/(1-\delta_1)} . \tag{4.15}$$

Now, using η and N in (4.15) as given by (iii) and (3.11) and provided that

$$\log T \geq 2\left(2\frac{\bar{K}(L_*+1)}{\zeta}\right)^{2/3} ,$$

holds then we get

$$\sigma_+^{-2}\,\mathcal{R}(N,T) \leq \frac{(\log T)^3}{T}\left(\log\left(\lceil\log T\rceil^2\right) + \frac{(\phi(\zeta))^{1/(1-\delta_1)}}{8\,T^{\zeta\,(8^{1/2}\bar{K}(L_*+1))^{-1}(\log T)^{1/2}-2}}\right) .$$

For any $\beta > 0$, this upper bound is negligible with respect to $T^{-2\beta/(2\beta+1)}$ and, with (4.10) we get (3.12).

5 Proof of the lower bound

We now provide a proof of Theorem 3.1. We consider an autoregressive equation of order one

$$X_{t,T} = \theta((t-1)/T)X_{t-1,T} + \xi_t, \tag{5.1}$$

where $(\xi_t)_{t\in\mathbb{Z}}$ is i.i.d. with density f as in (I-3). In this case, provided that $\sup_{u\leq 1}|\theta(u)| < 1$, the representation (3.4) of the stationary solution reads, for all $t \leq T$ as

$$X_{t,T} = \sum_{j=0}^{\infty}\prod_{s=1}^{j}\theta((t-s)/T)\,\xi_{t-j} , \tag{5.2}$$

with the convention $\prod_{s=1}^{0} \theta((t-s)/T) = 1$. The class of models so defined with $\theta \in \Lambda_1(\beta, R) \cap s_1(\delta)$ corresponds to Assumption (**M**-2) with $(\boldsymbol{\theta}, \sigma)$ in $C(\beta, R, \delta, \rho, 1)$ such that only the first component of $\boldsymbol{\theta}$ is nonzero and σ is constant and equal to one.
We write henceforth in this proof section $\mathbb{P}_\theta$ for the law of the process $X = (X_{t,T})_{t \le T, T \ge 1}$ and $\mathbb{E}_\theta$ for the corresponding expectation.
Let $\widehat{X} = (\widehat{X}_{t,T})_{1 \le t \le T}$ be any predictor of $(X_{t,T})_{1 \le t \le T}$ in the sense of Definition 3. Define $\widehat{\theta} = (\widehat{\theta}_{t,T})_{0 \le t \le T-1} \in \mathbb{R}^T$ by

$$\widehat{\theta}_{t,T} = \begin{cases} \widehat{X}_{t+1,T} / X_{t,T} & \text{if } X_{t,T} \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

For any vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^T$, we define

$$d_X(\mathbf{u}, \mathbf{v}) = \left(\frac{1}{T} \sum_{t=0}^{T-1} X_{t,T}^2 \, (u_t - v_t)^2 \right)^{1/2}. \tag{5.3}$$

By (5.1), since $X_{t,T}$ and $\widehat{\theta}_{t,T}$ are $\mathcal{F}_{t,T}$-measurable, they are independent of ξ_{t+1} and we have

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_\theta \left[\left(\widehat{X}_{t,T} - X_{t,T} \right)^2 \right] - 1 = \mathbb{E}_\theta \left[d_X^2(\widehat{\theta}, \mathrm{v}_T\{\theta\}) \right],$$

where, for any $\theta : (-\infty, 1] \to \mathbb{R}$, $\mathrm{v}_T\{\theta\} \in \mathbb{R}^T$ denotes the T-sample of θ on the regular grid $0, 1/T, \ldots, (T-1)/T$,

$$\mathrm{v}_T\{\theta\} = (\theta(t/T))_{0 \le t \le T-1} \, .$$

Hence to prove the lower bound of Theorem 3.1, it is sufficient to show that there exist $\theta_0, \ldots, \theta_M \in \Lambda_1(\beta, R) \cap s_1(\delta)$, $c > 0$ and $T_0 \ge 1$ both depending only on δ, β, R and the density f, such that for any $\widehat{\theta} = (\widehat{\theta}_{t,T})_{0 \le t \le T-1}$ adapted to $(\mathcal{F}_{t,T})_{0 \le t \le T-1}$ and $T \ge T_0$, we have

$$\max_{j=0,\ldots,M} \mathbb{E}_{\theta_j} \left[d_X^2(\widehat{\theta}, \mathrm{v}_T\{\theta_j\}) \right] \ge c \, T^{-2\beta/(2\beta+1)}. \tag{5.4}$$

We now face the more standard problem of providing a lower bound for the minimax rate of an estimation error, since $\widehat{\theta}$ is an estimator of $\mathrm{v}_T\{\theta\}$. The path for deriving such a lower bound is explained in [17, Chapter 2]. However we have to deal with a loss function d_X which depends on the observed process X. Not only the loss function is random, but it is also not independent of the estimator $\widehat{\theta}$. The proof of the lower bound (5.4) thus requires nontrivial adaptations. It relies on some intermediate lemmas.

Lemma 3. *We write $\mathcal{K}(\mathbb{P}, \mathbb{P}')$ for the Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{P}'$. For any functions $\theta_0, \ldots, \theta_M$ from $[0,1]$ to $\mathbb{R}$ such that*

$$\max_{j=0,\ldots,M} \mathcal{K}(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0}) \le \frac{2\mathrm{e}}{2\mathrm{e}+1} \log(1+M) \tag{5.5}$$

and any $r > 0$ we have

$$\max_{j=0,\ldots,M} \mathbb{E}_{\theta_j} \left[d_X^2 \left(\widehat{\theta}, \mathrm{v}_T\{\theta_j\} \right) \right] \ge \frac{r^2}{4} \left(\frac{1}{2\mathrm{e}+1} - \max_{j=0,\ldots,M} \mathbb{P}_{\theta_j} \left(\min_{i : i \neq j} d_{X,T}(\theta_i, \theta_j) \le r \right) \right),$$

where we denote, for any two functions θ, θ' from $(-\infty, 1]$ to $\mathbb{R}$,

$$d_{X,T}(\theta, \theta') = d_X\left(\mathrm{v}_T\{\theta\}, \mathrm{v}_T\{\theta'\}\right) .$$

Proof. We define $\hat{\jmath}$ as the (random) smallest index which minimizes $d_X(\widehat{\theta}, \mathrm{v}_T\{\theta_j\})$ over $j \in \{0, \dots, M\}$ so that $d_X(\widehat{\theta}, \mathrm{v}_T\{\theta_{\hat{\jmath}}\}) = \min_{\theta \in \{\theta_0, \dots, \theta_M\}} d_X(\widehat{\theta}, \mathrm{v}_T\{\theta\})$. Note that $d_{X,T}(\theta_{\hat{\jmath}}, \theta_j) \le d_X(\mathrm{v}_T\{\theta_{\hat{\jmath}}\}, \widehat{\theta}) + d_X(\widehat{\theta}, \mathrm{v}_T\{\theta_j\}) \le 2 d_X(\widehat{\theta}, \mathrm{v}_T\{\theta_j\})$. Hence

$$\begin{aligned}
\max_{j=0,\dots,M} \mathbb{E}_{\theta_j}\left[d_X^2(\widehat{\theta}, \mathrm{v}_T\{\theta_j\})\right] &\ge \frac{1}{4} \max_{j=0,\dots,M} \mathbb{E}_{\theta_j}\left[d_{X,T}^2(\theta_{\hat{\jmath}}, \theta_j)\right] \\
&\ge \frac{r^2}{4} \max_{j=0,\dots,M} \mathbb{P}_{\theta_j}\left(\{\hat{\jmath} \neq j\} \cap \left\{\min_{i:i\neq j} d_{X,T}(\theta_i, \theta_j) > r\right\}\right) \\
&\ge \frac{r^2}{4}\left(1 - \min_{j=0,\dots,M} \mathbb{P}_{\theta_j}(\hat{\jmath} = j) - \max_{j=0,\dots,M} \mathbb{P}_{\theta_j}\left(\min_{i:i\neq j} d_{X,T}(\theta_i, \theta_j) \le r\right)\right).
\end{aligned}$$

Birgé's lemma ([13, Corollary 2.18]) implies that

$$\min_{j=0,\dots,M} \mathbb{P}_{\theta_j}(\hat{\jmath} = j) \le \max\left\{\left(\frac{2\mathrm{e}}{2\mathrm{e}+1}\right), \left(\frac{\max_{j=0,\dots,M} \mathcal{K}(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0})}{\log(1+M)}\right)\right\},$$

so the lemma follows from Condition (5.5). □

We next construct some functions $\theta_0, \dots, \theta_M \in \Lambda_1(\beta, R) \cap s_1(\delta)$ fulfilling (5.5) and well spread in terms of the pseudo-distance $d_{X,T}$. Consider the infinitely differentiable kernel K defined by

$$K(u) = \exp\left(-\frac{1}{1-4u^2}\right) 1_{|u|<1/2} .$$

Given any $m \ge 8$, Vershamov-Gilbert's lemma ([17, Lemma 2.9]) ensures the existence of $M+1$ points $w^{(0)}, \dots, w^{(M)}$ in the hypercube $\{0,1\}^m$ such that

$$M \ge 2^{m/8}, \ w^{(0)} = 0 \quad \text{and} \quad \mathrm{card}\left\{k : w_l^{(j)} \neq w_l^{(i)}\right\} \ge m/8 \quad \text{for all } j \neq i. \tag{5.6}$$

We then define $\theta_0, \dots, \theta_M$ by setting, for all $x \le 1$,

$$\theta_j(x) = \frac{R_0}{m^\beta} \sum_{l=1}^{m} w_l^{(j)} K(mx - l + 1/2) \quad \text{for} \quad j = 0, \dots, M , \tag{5.7}$$

where

$$R_0 = \min\left(\delta, R/\left(2\,|K|_{\Lambda,\beta}\right)\right) . \tag{5.8}$$

Since $K = 0$ out of $(-1/2, 1/2)$, we observe that $\theta_j(x) = 0$ for all $x \le 0$ and

$$\theta_j(x) = \frac{R_0}{m^\beta}\, w_{\lfloor mx \rfloor + 1}^{(j)}\, K(\{mx\} - 1/2), \quad \text{for all } x \in [0,1], \tag{5.9}$$

where $\{mx\} = mx - \lfloor mx \rfloor$ denotes the fractional part of mx. Thus we have

$$\theta^* := \max_{0\le j\le M} \sup_{x\in[0,1]} |\theta_j(x)| \le \frac{R_0 \mathrm{e}^{-1}}{m^\beta} \le \delta < 1 . \tag{5.10}$$

We first check that the definition of R_0 ensures that the θ_j's are in the expected set of parameters.

Lemma 4. *For all $j = 0, \dots, M$, we have $\theta_j \in \Lambda_1(\beta, R) \cap s_1(\delta)$.*

Proof. By (5.10), we have $\theta_j \in s_1(\delta)$ for all $j = 0, \dots, M$. Decompose the Hölder-exponent $\beta = k + \alpha$ where k is an integer and $\alpha \in (0, 1]$. Differentiating (5.7) k times, we have, as in (5.9),

$$\theta_j^{(k)}(x) = \frac{R_0}{m^\alpha}\, w^{(j)}_{\lfloor mx \rfloor + 1}\, K^{(k)}(\{mx\} - 1/2), \quad \text{for all } x \in [0, 1].$$

Thus, for s, s' in the same interval $[l/m, (l+1)/m]$ with $l = 0, \dots, m-1$, we get

$$\begin{aligned}\left|\theta_j^{(k)}(s) - \theta_j^{(k)}(s')\right| &\le \frac{R_0}{m^\alpha}\, \left|K^{(k)}(m\, s - l - 1/2) - K^{(k)}(m\, s' - l - 1/2)\right| \\ &\le R_0\ |K|_{\Lambda,\beta}\ |s - s'|^\alpha\end{aligned}$$

The same inequality then follows with R_0 replaced by $2R_0$ for s, s' in two such consecutive intervals. Now, if s, s' are separated by at least one such interval, we have $|s - s'| \ge m^{-1}$ and, using that K has support in $(-1/2, 1/2)$, we have that $\left|K^{(k)}(x)\right|$ is bounded by $|K|_{\Lambda,\beta}$. We thus get in this case that

$$\left|\theta_j^{(k)}(s) - \theta_j^{(k)}(s')\right| \le \frac{2R_0}{m^\alpha} \sup_{-1/2 \le x \le 1/2} \left|K^{(k)}(x)\right| \le 2R_0\ |K|_{\Lambda,\beta}\ |s - s'|^\alpha\ .$$

The last two displays and (5.8) then yields $\theta_j \in \Lambda_1(\beta, R)$. □

Next we provide a bound to check the required condition (5.5) on the chosen θ_j's.

Lemma 5. *For all $j = 0, \dots, M$, we have*

$$\mathcal{K}(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0}) \le \frac{8\, \mathrm{e}^{-2}\, \kappa\, R_0^2}{(1 - \delta^2) \log 2}\ \frac{T}{m^{1+2\beta}} \log(1 + M)\ ,$$

where κ is the constant appearing in (I-3).

Proof. We note that under (I-3), the likelihood ratio $\mathrm{d}\mathbb{P}_{\theta_j}/\mathrm{d}\mathbb{P}_{\theta_0}$ of $(X_{s,T})_{s \le T}$ reads

$$\frac{\mathrm{d}\mathbb{P}_{\theta_j}}{\mathrm{d}\mathbb{P}_{\theta_0}} = \prod_{t=1}^{T} \frac{f\left(X_{t,T} - \theta_j((t-1)/T) X_{t-1,T}\right)}{f\left(X_{t,T} - \theta_0((t-1)/T) X_{t-1,T}\right)}\ .$$

Using that $\theta_0 \equiv 0$ by (5.6) and that, under $\mathbb{P}_{\theta_j}$, we have $X_{t,T} = \theta_j((t-1)/T) X_{t-1,T} + \xi_t$, we get

$$\begin{aligned}\mathcal{K}\left(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0}\right) &= \mathbb{E}_{\theta_j}\left[\log \frac{\mathrm{d}\mathbb{P}_{\theta_j}}{\mathrm{d}\mathbb{P}_{\theta_0}}\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{\theta_j}\left[\log \frac{f(\xi_t)}{f(\theta_j((t-1)/T) X_{t-1,T} + \xi_t)}\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{\theta_j} \int \log\left(\frac{f(u)}{f(\theta_j((t-1)/T) X_{t-1,T} + u)}\right) f(u)\, \mathrm{d}u\end{aligned}$$

Using Assumption (I-3) yields

$$\mathcal{K}\left(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0}\right) \leq \sum_{t=1}^{T} \mathbb{E}_{\theta_j}\left[\kappa\theta_j^2\left(\frac{t-1}{T}\right)X_{t-1,T}^2\right] \leq \kappa\theta^{*2}\sum_{t=1}^{T}\mathbb{E}_{\theta_j}\left[X_{t-1,T}^2\right]. \tag{5.11}$$

The series representation (5.2), the fact that ξ is centered with unit variance and (5.10) imply that for all $t = 0, \ldots, T$

$$\mathbb{E}_{\theta_j}\left[X_{t,T}^2\right] \leq (1-\theta^{*2})^{-1} .$$

Using this bound and (5.10) in (5.11), we obtain

$$\mathcal{K}\left(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0}\right) \leq \frac{R_0^2\,\mathrm{e}^{-2}\,\kappa\,T}{(1-\delta^2)\,m^{2\beta}} .$$

The proof of Lemma 5 now follows by applying the first bound in (5.6). □

Finally we need a control on the distances $d_{X,T}^2(\theta_i, \theta_j)$.

Lemma 6. *For any $\varepsilon > 0$, there exists a constant A depending only on ε and the density f of ξ such that for all $m \geq 16$, $T \geq 4m$ and $j = 0, \ldots, M$,*

$$\mathbb{P}_{\theta_j}\left(\min_{i:i\neq j} d_{X,T}^2(\theta_i,\theta_j) \leq A\frac{R_0^2}{m^{2\beta}}\right) \leq \varepsilon + \frac{2R_0\mathrm{e}^{-3}}{A\,(1-\delta)m^{\beta}} . \tag{5.12}$$

Proof. The proof relies on an upper bound of $d_{X,T}^2(\theta_i, \theta_j)$ involving the noise (ξ_t). By the expression of θ_j in (5.9), we have

$$d_{X,T}^2(\theta_i,\theta_j) = \frac{R_0^2}{Tm^{2\beta}}\sum_{t=0}^{T-1} X_{t,T}^2\left(w_{k(t)}^{(i)} - w_{k(t)}^{(j)}\right)^2 K^2\left(\varphi\left(t\right)\right) , \tag{5.13}$$

where we denoted $\varphi(t) = \{mt/T\} - 1/2$ and $k(t) = \lfloor mt/T \rfloor + 1$. Using (5.2) and (5.10), we have, for all $0 \leq t \leq T-1$,

$$\left|X_{t,T}\right| \geq |\xi_t| - \sum_{j=1}^{\infty}\theta^{*j}\left|\xi_{t-j}\right| ,$$

which implies

$$X_{t,T}^2 \geq \xi_t^2 - 2\,|\xi_t|\sum_{j=1}^{\infty}\theta^{*j}\left|\xi_{t-j}\right| .$$

Inserting this bound in (5.13), we get

$$\frac{m^{2\beta}}{R_0^2}\,d_{X,T}^2(\theta_i,\theta_j) \geq \frac{1}{T}\sum_{t=0}^{T-1}\xi_t^2\left(w_{k(t)}^{(i)} - w_{k(t)}^{(j)}\right)^2 K^2\left(\varphi\left(t\right)\right) - \mathcal{R}_T , \tag{5.14}$$

where

$$\mathcal{R}_T = \frac{2\mathrm{e}^{-2}}{T}\sum_{t=0}^{T-1}\sum_{j=1}^{\infty}\theta^{*j}\,|\xi_t|\,\left|\xi_{t-j}\right|$$

Thus, with (5.14), the left-hand side of Inequality (5.12) is upper bounded by

$$\mathbb{P}_{\theta_j}\left(\min_{i:i\neq j}\frac{1}{T}\sum_{t=0}^{T-1}\xi_t^2\left(w_{k(t)}^{(i)}-w_{k(t)}^{(j)}\right)^2 K^2\left(\varphi\left(t\right)\right)<2A\right)+\mathbb{P}(\mathcal{R}_T>A)\ .$$

Using that ξ is centered with unit variance and then (5.10), we easily get that

$$\mathbb{E}_{\theta_j}\left[\mathcal{R}_T\right]\leq\frac{2\mathrm{e}^{-2}}{T}\sum_{t=0}^{T-1}\sum_{j=1}^{\infty}\theta^{*j}\leq\frac{2\mathrm{e}^{-2}\theta^*}{1-\theta^*}\leq\frac{2R_0\mathrm{e}^{-3}}{(1-\delta)m^\beta}\ .$$

Hence, By Markov Inequality, to conclude the proof, it now suffices to show that, for A well chosen,

$$\mathbb{P}_{\theta_j}\left(\min_{i:i\neq j}\frac{1}{T}\sum_{t=0}^{T-1}\xi_t^2\left(w_{k(t)}^{(i)}-w_{k(t)}^{(j)}\right)^2 K^2\left(\varphi\left(t\right)\right)<2A\right)\leq\varepsilon\ . \tag{5.15}$$

For $k\in\{1,\ldots,m\}$ we define $J_k=\{\lfloor(k-1)T/m\rfloor+i\ :\ \lceil T/(4m)\rceil+1\leq i\leq\lfloor 3T/(4m)\rfloor\}$. We observe that the cardinality of J_k is

$$\Gamma(T/m)=\lfloor 3T/(4m)\rfloor-\lceil T/(4m)\rceil\geq 1\ ,$$

where the lower bound is a consequence of the assumption $T\geq 4m$ in the lemma. Moreover, it is easy to check that we have $|\varphi(t)|\leq 1/4$ for all index $t\in J_k$ and that, for each $1\leq k\leq m$, the set J_k is included in the set $\{1\leq t\leq T-1 : k(t)=k\}$ (so that, in particular, $J_k\cap J_{k'}=\emptyset$ for $k<k'$). It follows that random variables

$$S_k=\frac{1}{\Gamma(T/m)}\sum_{t\in J_k}\xi_{t-1}^2,\quad\text{for } k=1,\ldots,m$$

are i.i.d. By the monotonicity of K in $\mathbb{R}_-$ and its symmetry we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\xi_t^2\left(w_{k(t)}^{(i)}-w_{k(t)}^{(j)}\right)^2 K^2\left(\varphi\left(t\right)\right)&\geq\frac{1}{T}\sum_{k=1}^{m}\left(w_k^{(i)}-w_k^{(j)}\right)^2\sum_{t\in J_k}\xi_t^2K^2\left(\varphi\left(t\right)\right)\\&\geq\frac{K^2(1/4)\Gamma(T/m)}{T}\sum_{k=1}^{m}\left(w_k^{(i)}-w_k^{(j)}\right)^2 S_k.\end{aligned}$$

From (5.6), for any $i,j\in\{1,\ldots,M\}$ there exist at least $\lceil m/8\rceil$ values of k for which $(w_k^{(i)}-w_k^{(j)})^2$ equals one in the above sum. Hence using the order statistics $S_{(1,m)}\leq\ldots\leq S_{(m,m)}$, we thus obtain that

$$\begin{aligned}\min_{i:i\neq j}\frac{1}{T}\sum_{t=0}^{T-1}\xi_t^2\left(w_{k(t)}^{(i)}-w_{k(t)}^{(j)}\right)^2 K^2\left(\varphi\left(t\right)\right)&\geq\frac{K^2(1/4)\Gamma(T/m)}{T}\sum_{k=1}^{\lceil m/8\rceil}S_{(k,m)}\\&\geq\frac{K^2(1/4)\,m\,\Gamma(T/m)}{16\,T}S_{(\lfloor m/16\rfloor,m)}\\&\geq\frac{K^2(1/4)}{128}S_{(\lfloor m/16\rfloor,m)},\end{aligned}$$

where we used $\Gamma(T/m) \geq T/(8m)$ for $T/m \geq 4$ in the last inequality. Let us denote by F the cumulative distribution function of S_1, which only depends on $\Gamma(T/m)$ and on the distribution of ξ_0. For $x > 0$, we have

$$\begin{aligned}\mathbb{P}(S_{(\lfloor m/16\rfloor,m)} \leq x) &= \mathbb{P}(\mathrm{Bin}(m, F(x)) \geq \lfloor m/16\rfloor) \\ &\leq \frac{m}{\lfloor m/16\rfloor} F(x) \leq 32F(x).\end{aligned}$$

Gathering the last two bounds, we get that

$$\begin{aligned}\mathbb{P}_{\theta_j}\left(\min_{i:i\neq j} \frac{1}{T}\sum_{t=1}^{T-1} \xi_t^2 \left(w_{k(t)}^{(i)} - w_{k(t)}^{(j)}\right)^2 K^2\left(\varphi(t)\right) \leq 2A\right) &\leq \mathbb{P}\left((S_{(\lfloor m/16\rfloor,m)} \leq \frac{256\,A}{K^2(1/4)}\right) \\ &\leq 32\; F\left(\frac{256\,A}{K^2(1/4)}\right).\end{aligned}$$

Recall that $\Gamma(T/m) \geq 1$ and note that S_1 admits a density, since ξ does. By the strong law of large numbers, we further have that the random variable S_1 converges to 1 almost surely when $\Gamma(T/m)$ goes to infinity, so there exists $x_0 > 0$ depending only on the density of ξ such that $F(x_0) \leq \varepsilon/32$ whatever the value of $\Gamma(T/m) \geq 1$. Therefore, there exists some $A > 0$, depending only on the distribution of ξ, such that (5.15) holds, which achieves the proof. □

We can now conclude the proof of Theorem 3.1.

Proof of Theorem 3.1. Recall that $\theta_0, \ldots, \theta_M$ in (5.7) are some parameters only depending on β and δ and some integer $m \geq 8$ and that, whatever the value of m, Lemma 4 insures that $\theta_0, \ldots, \theta_M$ belongs to $\Lambda_1(\beta, R) \cap s_1(\delta)$.
Hence it is now sufficient to show that (5.4) holds for a correct choice of m, relying on Lemmas 3, 5 and 6. Let us set

$$m = \max\left\{\left\lceil c_0 T^{1/(2\beta+1)}\right\rceil, 16\right\}, \tag{5.16}$$

where c_0 is a constant to be chosen. Then $Tm^{-1-2\beta} \leq c_0^{-1-2\beta}$ and, by Lemma 5, we can choose c_0 only depending on β, R, κ and δ so that Condition (5.5) of Lemma 3 is met. We thus get that, for any $r > 0$,

$$\max_{j=0,\ldots,M} \mathbb{E}_{\theta_j}\left[d_X^2\left(\widehat{\theta}, \mathrm{v}_T\{\theta_j\}\right)\right] \geq \frac{r^2}{4}\left(\frac{1}{2\mathrm{e}+1} - \max_{j=0,\ldots,M} \mathbb{P}_{\theta_j}\left(\min_{i:i\neq j} d_{X,T}(\theta_i, \theta_j) \leq r\right)\right),$$

Applying Lemma 6 with $\varepsilon = 1/(4\mathrm{e}+2)$ and the previous bound with $r^2 = A\,R_0^2\,m^{-2\beta}$, we get, as soon as $T \geq 4m$,

$$\max_{j=0,\ldots,M} \mathbb{E}_{\theta_j}\left[d_X^2\left(\widehat{\theta}, \mathrm{v}_T\{\theta_j\}\right)\right] \geq \frac{r^2}{4}\left(\frac{1}{4\mathrm{e}+2} - \frac{2R_0\mathrm{e}^{-1}}{A\,(1-\delta)m^\beta}\right).$$

The proof is concluded by observing that, as a consequence of (5.16), we can choose a constant T_0 only depending on β, R, κ and δ such that $T \geq T_0$ implies that $T \geq 4m$ and that the term between parentheses is bounded by $1/(8\mathrm{e}+4)$ from below. □

A Application to online minimax adaptive prediction

A.1 From estimation to prediction

We define a sequence $(L_k)_{k\geq 1}$ by

$$L_k = \begin{cases} \binom{d}{k} & \text{if } 1 \leq k \leq d \\ 0 & \text{otherwise,} \end{cases}$$

which fulfills (L-1) with $L_* = \sum_{k=1}^{d} \binom{d}{k} = 2^d - 1$.
Given an estimator $\widehat{\boldsymbol{\theta}}_{t,T} = \left[\widehat{\theta}_{t,T}(1) \quad \dots \quad \widehat{\theta}_{t,T}(d)\right]'$, we define a predictor $\widehat{X}_{t,T}$ which is L-Lipschitz by setting

$$\widehat{X}_{t,T} = \sum_{k=1}^{d} \left(\min\left\{\max\left\{-L_k, \widehat{\theta}_{t,T}(k)\right\}, L_k\right\}\right) X_{t-k,T} \,. \tag{A.1}$$

The predictor $\widehat{X}_{t,T}$ is the natural linear predictor $\widehat{\boldsymbol{\theta}}'_{t-1,T}\mathbf{X}_{t-1,T}$ normalized to be at most L-Lipschitz. The normalization step amounts to project $\widehat{\boldsymbol{\theta}}_{t,T}$ on a rectangle $[-L_1, L_1] \times \cdots \times [-L_d, L_d]$ before deriving the linear predictor. This can only improve the quality of estimation for a stable TVAR model, since $\boldsymbol{\theta}$ takes values in the maximal set of stability $s_d(1)$, which implies that it is included in this rectangle at every point, see [14, Eq. 12]. We get the following result.

Lemma 7. *Assume that Assumption (M-2) holds. Consider, for some $1 \leq t \leq T$, an estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_{t,T})_{0\leq t\leq T-1}$ adapted to the filtration $(\mathcal{F}_{t,T})_{0\leq t\leq T-1}$. Define a predictor $\widehat{X} = (\widehat{X}_{t,T})_{1\leq t\leq T}$ as in (A.1). Then, for any $q > 1$ and for all and $1 \leq t \leq T$,*

$$\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}\left[\left(\widehat{X}_{t,T} - X_{t,T}\right)^2\right] - \sigma^2(t/T) \leq C_T \left(\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}\left[\left|\widehat{\boldsymbol{\theta}}_{t-1,T} - \boldsymbol{\theta}_{t-1,T}\right|^{2q}\right]\right)^{1/q} , \tag{A.2}$$

where

$$C_T(q) = \max_{1\leq t\leq T} \left(\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}\left[\left|\mathbf{X}_{t-1,T}\right|^{2q'}\right]\right)^{1/q'} ,$$

with $1/q' + 1/q = 1$.

Remark 6. Assume that the distribution ψ of the innovations satisfies (I-1) for some $p \geq 2q' > 2$. Then, the Proposition 1 combined with the Minkowski inequality ensure that there exists T_0, $\bar{K}$, δ_1 such that, for any $(\boldsymbol{\theta}, \sigma) \in C(\beta, R, \delta, 0, \sigma_+)$,

$$C_T(q) \leq d\left(\frac{\bar{K}\sigma_+}{1-\delta_1}\right)^2 m_{2q'}^{1/q'}, \quad \text{for all } T \geq T_0 \,.$$

Proof. Denote by $\widetilde{\boldsymbol{\theta}}_{t,T}$ the projection of $\widehat{\boldsymbol{\theta}}_{t,T}$ onto the rectangle $[-L_1, L_1]\times\cdots\times[-L_d, L_d]$, that is, $\widetilde{\theta}_{t,T}(k) = \min\left\{\max\left\{-L_k, \widehat{\theta}_{t,T}(k)\right\}, L_k\right\}$. By [14, Eq. 12], $\boldsymbol{\theta}_{t,T}$ lies in this rectangle and thus

$$\left|\widetilde{\boldsymbol{\theta}}_{t,T} - \boldsymbol{\theta}_{t,T}\right| \leq \left|\widehat{\boldsymbol{\theta}}_{t,T} - \boldsymbol{\theta}_{t,T}\right| \,. \tag{A.3}$$

Using (B.2) and that $\widehat{\boldsymbol{\theta}}_{t-1,T}$ is a $\mathcal{F}_{t-1,T}$-measurable, we have, for all $t = 1, \dots, T$,

$$\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}\left[\left(\widehat{X}_{t,T} - X_{t,T}\right)^2\right] = \mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}\left[\left(\left(\widetilde{\boldsymbol{\theta}}_{t-1,T} - \boldsymbol{\theta}_{t-1,T}\right)' \mathbf{X}_{t-1,T}\right)^2\right] + \sigma^2(t/T) \; .$$

Define q' by the relation $1/q' + 1/q = 1$. Thus, with (A.3) and the Hölder inequality, we get that the left-hand side of (A.2) is bounded from above by

$$\left(\mathbb{E}_{(\boldsymbol{\theta},\sigma)}\left[\left|\widehat{\boldsymbol{\theta}}_{t-1,T} - \boldsymbol{\theta}_{t-1,T}\right|^{2q}\right]\right)^{1/q} \left(\mathbb{E}_{(\boldsymbol{\theta},\sigma)}\left[\left|\mathbf{X}_{t-1,T}\right|^{2q'}\right]\right)^{1/q'}$$

which concludes the proof of Lemma 7. □

By Lemma 7, to exhibit (ψ,β)-minimax-rate predictors in the sense of Definition 4, it suffices to have (ψ,β)-minimax-rate estimators of $\boldsymbol{\theta}$ in the sense of L^q-norm. Parameter estimation for TVAR models, or, more generally for locally stationary processes has been intensively studied in the past two decades, see [6] for a recent overview on this problem. To our knowledge, minimax-rate estimation results are sparse. The more widely spread approach for studying the behaviour of such estimators consists in establishing a central limit theorem under differentiablity conditions. Moment upper bound are provided in [7] and could be used to obtain minimax rate results. However the estimator, which is based on a localized Yule-Walker estimation method is not naturally adapted to the filtration $(\mathcal{F}_{t,T})_{0\le t\le T-1}$ as required for $(\widetilde{\boldsymbol{\theta}}_{t,T})_{0\le t\le T-1}$ above. Such a constraint could be clearly be met with some adaptation of the Yule-Walker approach. On the other hand it is directly satisfied by the estimators studied in [14]. There, an online estimator is proposed, the normalized least mean squares estimator $\widehat{\boldsymbol{\theta}}_{t,T}(\mu)$, depending on a gradient step size μ. For any $\beta \in (0,1]$, provided that the gradient step μ is well chosen the NLMS estimator is (ψ,β)-minimax-rate, see [14, Corollary 3]. More precisely, assume (**M**-2) with ψ satisfying (**I**-1) for some $p \ge 4$. Then, for any $\varepsilon > 0$, $R > 0$, $\delta \in (0,1)$, $\rho \in [0,1]$ and $q \in [1, p/6)$, there exists $M > 0$ such that, for all $(\boldsymbol{\theta},\sigma) \in \mathcal{C}(\beta, R, \delta, \sigma_-, \sigma_+)$ and $\varepsilon > 0$,

$$\sup_{\varepsilon \le t/T \le 1} \left(\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}\left[\left|\widehat{\boldsymbol{\theta}}_{t,T}(T^{-2\beta/(1+2\beta)}) - \boldsymbol{\theta}_{t,T}\right|^{2q}\right]\right)^{1/q} \le M\, T^{-2\beta/(1+2\beta)} \; .$$

Clearly, from [14], the constant M can be bounded uniformly for β in any compact subinterval away from 0, as required in Definition 5. Lemma 7 applies for $q \ge p/(p-2)$ so to meet the condition $q \in [1, p/6)$, we set $q = p/(p-2)$ and impose $p > 8$ and finally obtain that

$$\sup_{\varepsilon \le t/T \le 1} \mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}\left[\left(\widehat{X}_{t,T}(T^{-2\beta/(1+2\beta)}) - X_{t,T}\right)^2\right] - \sigma^2(t/T) \le C'\, \sigma_+^2\, T^{-2\beta/(1+2\beta)} \; ,$$

where $\widehat{X}_{t,T}(\mu)$ is the predictor defined from the estimator $\widehat{\boldsymbol{\theta}}_{t,T}(\mu)$ as in (A.1). This is almost what is required in our Definition 5 except that in (3.9) we have $T^{-1}\sum_{t=1}^{T}(\dots)$ instead of $\sup_{\varepsilon\le t/T\le 1}(\dots)$. In fact one can take $\varepsilon = 0$, provided that a burn-in period of observation is assumed prior to the time origin. It would only require the NLMS estimator to be running from observations $X_{t,T}$ started at times $t \ge -\varepsilon T$ for some positive ε, which seems a reasonable assumption in practice. Finally, let us recall

that, as shown in [14], NLMS estimators are no longer minimax rate for an Hölder smoothness index $\beta > 1$. However, a bias reduction technique can be used to obtain a minimax-rate estimator for $\beta \in (1, 2]$, see [14, Corollary 9].

B Postponed proofs

B.1 A useful lemma

The following lemma provides a uniform bound on the norm of a product of matrices sampled from a continuous function defined on an interval I and valued in a set of $d \times d$ matrices with bounded spectral radius and norm.

Lemma 8. *Let $d \geq 1$ and I an interval of $\mathbb{R}$. Let A be a function defined on I taking values in the set of $d \times d$ matrices with eigenvalues moduli at most equal to δ. Let $|\cdot|$ be any matrix norm. Denote by A^* the corresponding uniform norm of A,*

$$A^* = \sup_{t \in I} |A(t)| \ ,$$

and, for any $h > 0$, $\omega_h(A, I)$ the modulus of continuity of A over I,

$$\omega_h(A; I) = \sup \{|A(t) - A(s)| \ : \ s, t \in I, \, |s - t| \leq h\} \ .$$

Let $\delta_1 > \delta$. Then there exist some positive constants ε, ℓ and K only depending on A^*, δ and δ_1 such that, for any $h \in (0, 1)$ fulfilling $\omega_h(A; I) \leq \varepsilon$, we have, for all $s < t$ in I and all integer $p \geq \ell(t - s)/h$,

$$\left| \underbrace{A(t)A(t - (t - s)/p)A(t - 2(t - s)/p) \ldots A(s)}_{p+1 \text{ terms}} \right| \leq K \, \delta_1^{p+1} \ . \tag{B.1}$$

Proof. Denote by $\Pi(s, t; p)$ the product of matrices appearing in the left-hand side of (B.1). The proof goes along the same lines as [14, Proposition 13] but we use the modulus of continuity instead of the β-Lipschitz norm to control the local oscillation of matrices. Take an arbitrary $\delta_2 \in (\delta, \delta_1)$ (say the middle point). Then, for any integer $\ell \geq 1$, using the decomposition of the product into a product of blocks of same length, we obtain that

$$|\Pi(s, t; p)| \leq \left(K_1 \delta_2^\ell + K_2 \omega_h(A; I)\right)^{\lfloor (p+1)/\ell \rfloor} \ (K_1 \delta_2^r + K_2 \omega_h(A; I)) \ ,$$

where $h = \ell(t-s)/p$, $r = p+1-\ell\lfloor (p+1)/\ell \rfloor$ and K_1 and K_2 are two constants depending only on A^*, δ_2 and δ. We can choose a positive integer ℓ and a positive number ε_0 only depending on δ_2, δ_1 and K_1 such that

$$K_1 \delta_2^\ell \leq \delta_1^\ell - \varepsilon_0 \ .$$

In the following we set $\varepsilon = \varepsilon_0 / K_2$. So the previous bound gives that for any $h \in (0, 1)$ such that $\omega_h(A; I) \leq \varepsilon$ and $\ell(t - s)/p \leq h$,

$$|\Pi(s, t; p)| \leq \delta_1^{\ell \lfloor (p+1)/\ell \rfloor} \ (K_1 \delta_2^r + \varepsilon_0) \leq K_1 \delta_1^{p+1} + \varepsilon_0 \delta_1^{\ell \lfloor (p+1)/\ell \rfloor} \leq (K_1 + \varepsilon_0) \, \delta_1^{p+1} \ .$$

Hence the result. □

B.2 Proof of Proposition 1

We can now provide a proof of Proposition 1.
Eq. (3.1) can be more compactly written as

$$X_{t,T} = \boldsymbol{\theta}'\left(\frac{t-1}{T}\right)\mathbf{X}_{t-1,T} + \sigma\left(\frac{t}{T}\right)\xi_{t,T}\ , \tag{B.2}$$

where A' denotes the transpose of matrix A, and we denote $\boldsymbol{\theta} = [\theta_1\ \ldots\ \theta_d]'$ and $\boldsymbol{X}_{t-1,T} = [X_{t-1,T}\ \ldots\ X_{t-d,T}]'$.
For all $k \geq 0$, iterating this recursive equation k times, we have

$$X_{t,T} = \boldsymbol{e}_1'\left[\prod_{i=1}^{k+1} A\left(\frac{t-i}{T}\right)\right]\boldsymbol{X}_{t-k-1,T} + \sum_{j=0}^{k}\sigma\left(\frac{t-j}{T}\right)\boldsymbol{e}_1'\left[\prod_{i=1}^{j} A\left(\frac{t-i}{T}\right)\right]\boldsymbol{e}_1\xi_{t-j}\ ,$$

where $\boldsymbol{e}_1 = \begin{bmatrix}1 & 0 & \ldots & 0\end{bmatrix}$ and

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

Note that the eigenvalues of $A(u)$ are the reciprocals of the roots of the local time-varying autoregressive polynomial $z \mapsto \boldsymbol{\theta}(z;u)$ and thus are at most $\delta < 1$. Moreover since $\boldsymbol{\theta}$ is bounded by a constant only depending on d and is uniformly continuous on $I = (-\infty, 1]$, so is A as a function defined on I and we can find $h \in (0,1)$ such that $\omega_h(A, I) \leq \varepsilon$ for any positive ε. Thus, Lemma 8 gives that there exist positive constants K_1 and K_2 only depending on δ_1 and h such that, for all $T \geq 1$, $t \leq T$ and $j \geq 1$ such that $j > K_1(j/T)$ (that is, $T > K_1$),

$$\left|\prod_{i=1}^{j} A\left(\frac{t-i}{T}\right)\right| \leq K_2\,\delta_1^j\ .$$

Hence we obtain that

$$X_{t,T} = \sum_{i=1}^{d} b_{t,T}(k,i)X_{t-k-i,T} + \sum_{j=0}^{k} a_{t,T}(j)\,\sigma\left(\frac{t-j}{T}\right)\xi_{t-j,T}, \quad 1 \leq t \leq T\ . \tag{B.3}$$

with, provided that $T > K_1$, for all $t \leq T$, $k, j \geq 1$ and $i = 1, \ldots, d$,

$$\left|b_{t,T}(k,i)\right| \leq K_2\delta_1^k\ ,$$
$$\left|a_{t,T}(j)\right| \leq K_2\delta_1^j\ .$$

The result follows.

B.3 Proof of Lemma 1

We conclude the appendix with the postponed proof of Lemma 1. The idea is to choose a convenient $i_N \in \{1, \dots, N\}$ and use that

$$\min_{1\le i\le N} S_T(\widehat{X}_T^{(\beta_i)}; \psi, \beta, R, \delta, \rho, \sigma_+) \le S_T(\widehat{X}_T^{(\beta_{i_N})}; \psi, \beta, R, \delta, \rho, \sigma_+) \ .$$

The choice of i_N differs depending on the finiteness of β_0.
Let us first consider the case $\beta_0 < \infty$. Let $\beta \in (0, \beta_0)$, $\delta \in (0,1)$, $R > 0$ and $\rho \in [0,1]$. Let $i_N \in \{1, \dots, N\}$ be such that $\beta_{i_N} = (i_N - 1)\beta_0/N < \beta \le i_N\beta_0/N$. Since $C(\beta, R, \delta, \rho, \sigma_+) \subset C(\beta_{i_N}, R, \delta, \rho, \sigma_+)$, we have, for all $\delta \in (0,1)$, $R > 0$, $\rho > 0$ and $\sigma_+ > 0$,

$$\begin{aligned} T^{2\beta/(1+2\beta)} S_T(\widehat{X}_T^{(\beta_{i_N})}; \psi, \beta, R, \delta, \rho, \sigma_+) &\le T^{2\beta/(1+2\beta)} S_T(\widehat{X}_T^{(\beta_{i_N})}; \psi, \beta_{i_N}, R, \delta, \rho, \sigma_+) \\ &\le T^{2\beta_0/N}\ T^{2\beta_{i_N}/(1+2\beta_{i_N})} S_T(\widehat{X}_T^{(\beta_{i_N})}; \psi, \beta_{i_N}, R, \delta, \rho, \sigma_+) \ , \end{aligned}$$

where we used that $\beta_{i_N} < \beta \le \beta_{i_N} + \beta_0/N$. Recall that we assumed $N \ge \lceil \log T \rceil$, so that $T^{2\beta_0/N} \le \mathrm{e}^{2\beta_0}$. Now, since for N large enough β_{i_N} remains in a closed interval of $(0, \beta_0)$ we get by Definition 5 that

$$\limsup_{T\to\infty} T^{2\beta_{i_N}/(1+2\beta_{i_N})} S_T(\widehat{X}_T^{(\beta_{i_N})}; \psi, \beta_{i_N}, R, \delta, \rho, \sigma_+) < \infty \ ,$$

which concludes the proof in the case $\beta_0 < \infty$.
We next consider the case where $\beta_0 = \infty$. In this case we take i_N such that $\beta_{i_N} = (i_N - 1)/N^{1/2} < \beta \le i_N/N^{1/2}$ which defines $i_N \in \{1, \dots, N\}$ uniquely as soon as $N^{1/2} > \beta$. The remainder of the proof is similar to the case $\beta_0 < \infty$ using the bound

$$T^{2\beta/(1+2\beta)} \le T^{2/N^{1/2}}\ T^{2\beta_{i_N}/(1+2\beta_{i_N})} \le \mathrm{e}^2\ T^{2\beta_{i_N}/(1+2\beta_{i_N})} \ ,$$

under the assumption $N \ge \lceil (\log T)^2 \rceil$.

Acknowledgements

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] Pierre Alquier and Olivier Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913, 2012.

[2] Oren Anava, Elad Hazan, Shie Mannor, and Ohad Shamir. Online learning for time series prediction. *arXiv preprint arXiv:1302.6927*, 2013. Preprint arXiv:1302.6927.

[3] 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.

[4] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006.

[5] R. Dahlhaus. On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Process. Appl.*, 62(1):139–168, 1996.

[6] Rainer Dahlhaus. Local inference for locally stationary time series based on the empirical spectral measure. *J. Econometrics*, 151(2):101–112, 2009.

[7] Rainer Dahlhaus and Liudas Giraitis. On the optimal segment length for parameter estimates for locally stationary time series. *J. Time Ser. Anal.*, 19(6):629–655, 1998.

[8] Rainer Dahlhaus and Wolfgang Polonik. Empirical spectral processes for locally stationary time series. *Bernoulli*, 15(1):1–39, 2009.

[9] Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

[10] Sébastien Gerchinovitz. *Prediction of individual sequences and prediction in the statistical framework: some links around sparse regression and aggregation techniques*. PhD thesis, Université Paris Sud-Paris XI, 2011.

[11] Hans Rudolf Künsch. A note on causal solutions for locally stationary ar-processes. 1995.

[12] Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.

[13] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[14] Eric Moulines, Pierre Priouret, and François Roueff. On recursive estimation for time varying autoregressive processes. *Ann. Statist.*, 33(6):2610–2654, 2005.

[15] Philippe Rigollet and Alexandre B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575, 2012.

[16] Gilles Stoltz. Contributions to the sequential prediction of arbitrary sequences: applications to the theory of repeated games and empirical studies of the performance of the aggregation of experts. 2011.

[17] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[18] Volodimir G Vovk. Aggregating strategies. In *Proc. Third Workshop on Computational Learning Theory*, pages 371–383, 1990.

[19] Yuhong Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 74(1):135–161, 2000.