Online tracking of a drifting parameter of a time series

Online Tracking of a Drifting Parameter of a Time Series

Eduard Belitser and Paulo Serra

Department of Mathematics, Eindhoven University of Technology

June 4, 2013

Abstract

We propose an online algorithm for tracking a multivariate time-varying parameter of a time series. The algorithm is driven by a gain function. Under assumptions on the gain function, we derive uniform error bounds on the tracking algorithm in terms of chosen step size for the algorithm and on the variation of the parameter of interest. We give examples of a number of different variational setups for the parameter where our result can be applied, and we also outline how appropriate gain functions can be constructed. We treat in some detail the tracking of time varying parameters of an AR(d) model as a particular application of our method.

Keywords: on-line tracking; recursive algorithm; stochastic approximation procedure; time series; time-varying parameter.

1 Introduction

When one analyzes data that arrive sequentially over time, it is important to detect secular changes in the underlying model which can then be adjusted accordingly. Estimation or tracking of time-varying parameters in stochastic systems is therefore of fundamental interest in sequential analysis. Furthermore, it arises in many engineering, econometric and biomedical applications and has an extensive literature widely scattered in these fields. Motivated by many applications in signal processing, speech recognition, communication systems, neural physiology, environmental and economic modeling, we consider recursive (online) estimation a the multivariate time-varying parameter of a time series.

Consider then an $\mathcal{X}$-valued time series $\{X_k,\, k \in \mathbb{N}_0\}$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $\mathcal{X} \subseteq \mathbb{R}^l$, such that at time moment $t = 0$ the first observation $X_0 \sim P_{\theta_0}$ and subsequently at each time moment $k \in \mathbb{N}$ a new datum X_k arrives according to the model $X_k|\boldsymbol{X}_{k-1} \sim P_{\theta_k}(\cdot|\boldsymbol{X}_{k-1})$ with transition law depending on some multivariate parameter $\theta_k \in \Theta \subseteq \mathbb{R}^d$ and where $\boldsymbol{X}_{k-1} = (X_0, X_1, \ldots, X_{k-1})$. Thus, the growing statistical model is, at time $t = n$, $\mathcal{P}^{(n)} = \mathcal{P}^{(n)}(\Theta^{n+1}) = \{\prod_{k=0}^{n} P_{\theta_k}(x_k|\boldsymbol{x}_{k-1}) : (\theta_0, \ldots, \theta_n) \in \Theta^{n+1},\, \boldsymbol{x}_n \in \mathcal{X}^{n+1}\}$ with the convention that $P_{\theta_0}(y_0|\boldsymbol{x}_{-1}) = P_{\theta_0}(y_0)$. This time series formulation represents the most general sequential setting, sequences of independent observations and Markov chains of arbitrary order are typical examples of models that fit into this framework.

The multivariate parameter $\theta_k \in \Theta \subseteq \mathbb{R}^d$, $k \in \mathbb{N}$, is time-varying and the goal is to estimate (or to track) its value based on the data $\boldsymbol{X}_k$ (and prior information) available by that time moment. Since the data arrives in a successive manner, conventional methods based on samples of fixed size are not easy to use. A more appropriate approach is based on sequential methods, stochastic recursive algorithms, which allow fast updating of parameter or state estimates at each instant as new data arrive and therefore can be used to produce an "online" inference, that is, during the operation of the system. Stochastic recursive algorithms, also known as stochastic approximation, take many forms and have numerous applications in the biomedical, socio-economic and engineering sciences, which highlights the interdisciplinary nature of the subject.

There is a vast literature on stochastic approximation beginning with the seminal papers of Robbins and Monro [1951] and Kiefer and Wolfowitz [1952]. There is a big variety of techniques in the area of stochastic approximation which have been developed and inspired by the applications from other fields. We mention here the books of Wasan [1969], Tsypkin [1971], Nevelson and Khasminskii [1976], Kushner and Clark [1978], Ljung and Söderström [1983], Benveniste et al. [1990] and Kushner and Yin [2003].

A classical topic in adaptive control concerns the problem of tracking drifting parameters of a linear regression model, or somewhat equivalently, tracking the best linear fit when the parameters change slowly. This problem also occurs in communication theory for adaptive equalizers and noise cancellation, etc., where the signal, noise, and channel properties change with time. Successful stochastic approximation schemes for tracking in the time-varying case were given by Brossier [1992], Delyon and Juditsky [1995], Kushner and Yang [1995], Kushner and Yin [2003] (see further references therein).

In Kushner and Yang [1995] (see also Benveniste et al. [1990] and Brossier [1992]) discuss the very important problem of the choice of the step sizes in the tracking algorithm. In general, the step size of the tracking algorithm is not necessarily decreasing to zero because of considerations concerning robustness of the actual physical model in practical online applications and to allow some tracking of the desired parameter as the system changes over time. In signal processing applications, it is usual to keep the step size bounded away from zero.

Coming back to our model $\mathcal{P}^{(n)}$ with time-varying parameter $\{\theta_k \in \Theta,\, k \in \mathbb{N}_0\}$, the problem of tracking a signal θ_k is clearly unfeasible, especially in such general formulation, without some conditions on the model $\mathcal{P}^{(n)}$. In general, some knowledge about the structure of underlying time seres and some control over the variability of the parameter θ_k over time are needed. Interestingly, in this seemingly very general time series framework, we actually do not require the knowledge of the model $\mathcal{P}^{(n)}$. Instead, all we do need is to be able to compute a so called *gain vector* at each time moment $k \in \mathbb{N}$, which is a certain (vector) function of the previous estimate of the parameter θ_k, new observation X_k and prehistory $\boldsymbol{X}_{k-1}$. The essential property of such gain vector is that it, roughly speaking, "pushes" in the right direction of the current value of true parameter to track. Although the assumption about the existence of that gain vector seems to be rather strong, we demonstrate on a number of interesting examples when such an assumption indeed holds. Basically, in case of Markov chain observations, if the form of transition density is known as function of the underlying parameter and it satisfies certain regularity assumptions, then the gain vector can always be constructed, for example, as a score function corresponding to the conditional maximum likelihood method. Under appropriate regularity conditions (the existence of the conditional Fisher information and L_2-differentiability of the conditional log likelihood), such a score function has always the property of gain vector at least locally.

A gain function, together with a step sequence and new observations from the model, can be used to adjust the current approximation of the drifting parameter, resulting in a tracking algorithm. To ease the verification of our assumptions on the gain function, we formulate them in two equivalent forms. Under some assumptions on the gain vectors, we establish a uniform non-asymptotic bound the L_1 error of the resulting tracking algorithm, in terms of the variation of the drifting parameter. Under the extra assumption that the gain function is bounded, we can strengthen this result to a uniform bound on the L_p error (and then an almost sure bound). These error bounds constitute our main result and they also guide us in the choice of the step size for the algorithm. Some extensions are also presented where we allow for approximation terms and approximate gains.

Based on our main result, we specify the appropriate choice for the step sequence in three different variational setups for the drifting parameter. We treat first the simple case of a constant parameter. Although we are mainly concerned with tracking time-varying parameters, our algorithm is still of interest in the constant parameter case since it should result in an algorithm which is both recursive and robust. We also consider a setup where the parameter is stabilizing. This covers both the case where the parameter is converging and where we sample the signal with increasing frequency. The third variational setup covers the important case of tracking smooth signals. This setup is somewhat different in that we make observations with a certain frequency from an underlying continuous time process which is indexed by a parameter changing like a Lipschitz function. Our result can then either be interpreted as a uniform, non-asymptotical result for each fixed sampling frequency or as an asymptotic statement in the observation frequency.

Examples are also given for different possible gain functions. These fall into two categories: general, score based gain functions for tracking multidimensional parameters in regular models and specialized gains for tracking more specific quantities. The latter include gains to track level sets or maxima of drifting functions (extending the classical Robbins-Monro and Kiefer-Wolfowitz algorithms) and gains to track drifting conditional quantiles. We also propose modifications for a given gain function (rescaling, truncation, projection) which can be used to design gains tailored specifically to verify our assumptions.

We illustrate our method by treating some concrete applications of the proposed algorithm but we focus mostly on the problem of tracking drifting parameters in autoregressive models. Results on tracking algorithms for these models already exist in the literature (cf. Belitser [2000], Moulines et al. [2005]) and we can derive similar results by choosing an appropriate gain function. Using our approach, obtaining error bounds on the resulting tracking algorithm reduces to verifying our assumptions for the chosen gain function which considerably simplifies the derivation of results.

This paper is structured as follows. In Section 2 we summarize the notation that will be used thought the paper, as well as our model and two equivalent formulations for our assumptions. Section 3 contains our main result and respective proof as well as some straightforward extensions of the main result. The construction and modification of gain functions for different models and different parameters of interest is explained in Section 4. Section 5 contains three examples of variational setups for the time-varying parameter for which we specify the tracking error implied by our main result. We collect in Section 6 some examples of applications. Section 7 contains the proofs for our lemmas.

2 Preliminaries

First we introduce some notation that we are going to use throughout the paper. All vectors are always column vectors. We use bold uppercase letters to represent sets of vectors. For vectors $x, y \in \mathbb{R}^d$, denote by $\|x\|_2$ and $\langle x, y\rangle = x^T y$ the usual Euclidean norm and the inner product in $\mathbb{R}^d$, respectively, and by $\|x\|_p$ the l_p norm on vectors in $\mathbb{R}^d$. For an event A we will represent the indicator of the event A as $\mathbf{1}_A$. For a symmetric $d \times d$ matrix M, let $\lambda_{(1)}(M)$ and $\lambda_{(d)}(M)$ be the smallest and the largest eigenvalues of M respectively. Denote $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Let also O denote the zero matrix, I the identity matrix and J the exchange matrix whose dimensions will be determined by the context. We will use the convention that $\sum_{i\in\varnothing} A_i = O$ and $\prod_{i\in\varnothing} B_i = I$ for matrices A_i and B_i with such dimensions that these matrix operations (summation and product) are well defined. When applied to matrices, the symbol $\|\cdot\|_p$ will represent the operator norm induced by the l_p vector norm, which is a matrix norm defined as

$$\|A\|_p = \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{x=1} \|Ax\|_p = \max_{x\leq 1} \|Ax\|_p.$$

Assume that by time $n \in \mathbb{N}$, we have observed $\boldsymbol{X}_n = (X_0, X_1, \ldots, X_n)$ according to the following model:

$$X_0 \sim P_{\theta_0}, \qquad X_k|\boldsymbol{X}_{k-1} \sim P_{\theta_k}(\cdot|\boldsymbol{X}_{k-1}), \quad k \in \mathbb{N}. \tag{1}$$

Here the time series $\{X_k,\, k \in \mathbb{N}_0\}$ takes value in some set $\mathcal{X} \subseteq \mathbb{R}^l$, i.e., $P(X_k \in \mathcal{X}) = 1$, $k \in \mathbb{N}_0$. Let $\mathcal{F}_k = \sigma(\boldsymbol{X}_k)$ denote the σ-algebra generated by $\boldsymbol{X}_k = (X_0, X_1, \ldots, X_k)$. The time-varying parameter $\theta_k = \theta_k(\boldsymbol{X}_{k-1})$, $k \in \mathbb{N}_0$, is allowed to depend on the past of the time series, i.e., it is assumed to be predictable with respect to the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$. Further, θ_k is assumed to take values in some convex compact subset Θ of $\mathbb{R}^d$, to be precise, $P(\theta_k(\boldsymbol{X}_{k-1}) \in \Theta) = 1$ for all $k \in \mathbb{N}_0$. We are interested in tracking the drifting parameter $\theta_k(\boldsymbol{X}_{k-1})$ which we will often abbreviate as simply θ_k. Denote from now on

$$C_\Theta = \sup_{\theta\in\Theta} \|\theta\|_2. \tag{2}$$

At time $n \in \mathbb{N}$, the underlying (growing) statistical model is $\mathcal{P}^{(n)} = \mathcal{P}^{(n)}(\Theta^{n+1})$, which we can write as

$$\mathcal{P}^{(n)}(\Theta^{n+1}) = \Big\{ \prod_{k=0}^{n} P_{\theta_k}(x_k|\boldsymbol{x}_{k-1}) : (\theta_0, \ldots, \theta_n) \in \Theta^{n+1},\, \boldsymbol{x}_n \in \mathcal{X}^{n+1} \Big\},$$

where $P_{\theta_0}(y_0|\boldsymbol{x}_{-1})$ should be understood as $P_{\theta_0}(y_0)$. For $k = 0, \dots, n$, each conditional measure belongs to

$$\mathcal{P}_k = \mathcal{P}_k(\Theta) = \big\{ P_\theta(\cdot|\boldsymbol{x}_{k-1}) : \, \theta \in \Theta, \, \boldsymbol{x}_{k-1} \in \mathcal{X}^k \big\}.$$

At time k, given $\boldsymbol{X}_k$, the model $\mathcal{P}_{k+1}$ contains all the relevant information about the next observation but we do not consider it to be (completely) known. Instead, we assume that our prior knowledge about the model is formalized as follows: for each $k \in \mathbb{N}$ we have certain $\mathbb{R}^d$-valued functions $G_k(x, \theta|\boldsymbol{x}_{k-1})$ at our disposal (which we will call *gain vectors*), $x \in \mathcal{X}$, $\boldsymbol{x}_{k-1} \in \mathcal{X}^k \subset \mathbb{R}^{lk}$, $\theta \in \mathbb{R}^d$, i.e., $G_k : \mathcal{X}^{k+1} \times \mathbb{R}^d \to \mathbb{R}^d$, and these gain vectors satisfy conditions (A1) and (A2) below.

(A1) For all $k \in \mathbb{N}$ and all $\theta, \vartheta \in \Theta$ the following statements hold almost surely:

$$g_k(\theta, \vartheta|\boldsymbol{X}_{k-1}) = \int G_k(x, \theta|\boldsymbol{X}_{k-1}) \, dP_\vartheta(x|\boldsymbol{X}_{k-1}) \tag{3}$$

is well defined, there exists a symmetric positive definite matrix $M_k = M_k(\boldsymbol{X}_{k-1})$ with (random) eigenvalues $0 < \Lambda_{(1)}(M_k) \leq \dots \leq \Lambda_{(d)}(M_k)$ and constants $0 < \lambda_1 \leq \lambda_2 < \infty$ such that

$$g_k(\theta, \vartheta|\boldsymbol{X}_{k-1}) = -M_k(\boldsymbol{X}_{k-1})(\theta - \vartheta), \tag{4}$$

with $0 < \lambda_1 \leq \mathbb{E}[\Lambda_{(1)}(M_k)|\boldsymbol{X}_{k-2}] \leq \Lambda_{(d)}(M_k) \leq \lambda_2 < \infty$.

(A2) There exists a constant $C > 0$ such that for all $k \in \mathbb{N}$ and all $\theta, \vartheta \in \Theta$,

$$\mathbb{E}\|G_k(X_k, \theta|\boldsymbol{X}_{k-1}) - g_k(\theta, \vartheta|\boldsymbol{X}_{k-1})\|_2^2 \leq C. \tag{5}$$

Note that assumption (A2) is redundant if, for example, the gain vectors $G_k(x, \theta|\boldsymbol{X}_{k-1})$ are almost surely bounded. Condition (A1) means, in a way, that, on average, the gain vector $G_k(X_k, \hat{\theta}_k|\boldsymbol{X}_{k-1})$ shifts $\hat{\theta}_k$ towards the "true" value $\theta_k = \theta_k(\boldsymbol{X}_{k-1})$:

$$\mathbb{E}\big[G_k(X_k, \hat{\theta}_k|\boldsymbol{X}_{k-1})|\mathcal{F}_{k-1}\big] = g_k(\hat{\theta}_k, \theta_k|\boldsymbol{X}_{k-1}) = -M_k(\boldsymbol{X}_{k-1})\big(\hat{\theta}_k - \theta_k\big),$$

for some symmetric, almost surely positive definite matrix $M_k(\boldsymbol{X}_{k-1})$ such that $0 < \lambda_1 \leq \mathbb{E}[\lambda_{(1)}(M_k)|\mathcal{F}_{k-2}] \leq \lambda_{(d)}(M_k) \leq \lambda_2 < \infty$.

Condition (A1) can be reformulated as ($\tilde{\text{A}}$1), which gives some intuition as to the role of the function g_k and which may, in certain situations, be simpler to verify.

($\tilde{\text{A}}$1) The quantity $g_k(\theta, \vartheta|\boldsymbol{X}_{k-1})$ defined by (3) satisfies, almost surely, the following conditions: there exist random variables $\Lambda_1(\boldsymbol{X}_{k-1})$ and $\Lambda_2(\boldsymbol{X}_{k-1})$ and constants $0 < \lambda_1 \leq \lambda_2 < \infty$, $0 < L < \infty$ such that for all $\theta, \vartheta \in \Theta$,

$$\begin{aligned} \Lambda_1(\boldsymbol{X}_{k-1})\|\theta - \vartheta\|_2^2 \leq -(\theta - \vartheta)^T g_k(\theta, \vartheta|\boldsymbol{X}_{k-1}) \leq \Lambda_2(\boldsymbol{X}_{k-1})\|\theta - \vartheta\|_2^2 \\ \|g_k(\theta, \vartheta|\boldsymbol{X}_{k-1})\|_2 \leq L\|\theta - \vartheta\|_2 \end{aligned} \tag{6}$$

with $0 < \lambda_1 \leq \mathbb{E}[\Lambda_1(\boldsymbol{X}_{k-1})|\boldsymbol{X}_{k-2}] \leq \Lambda_2\big(\boldsymbol{X}_{k-1}\big) \leq \lambda_2 < \infty$.

In view of the lemma below, if (A1) holds, then ($\tilde{\mathrm{A}}$1) will also hold (and vice versa); the values of the constants λ_1 and λ_2 appearing in the assumptions are different, though. The proof of this lemma is deferred to Section 7.

Lemma 1. *Let $x, y \in \mathbb{R}^d$. If there exists a symmetric positive definite matrix M such that $y = Mx$ and $0 < \lambda_1 \leq \lambda_{(1)}(M) \leq \lambda_{(d)}(M) \leq \lambda_2 < \infty$ for some $\lambda_1, \lambda_2 \in \mathbb{R}$, then $0 < \lambda_1' \|x\|^2 \leq \langle x, y\rangle \leq \lambda_2'\|x\|^2 < \infty$ and $\|y\| \leq C\|x\|$ for some $\lambda_1', \lambda_2', C \in \mathbb{R}$ (depending only on λ_1, λ_2) such that $0 < \lambda_1' \leq \lambda_2' < \infty$ and $C > 0$.*

Conversely, if $0 < \lambda_1'\|x\|^2 \leq \langle x, y\rangle \leq \lambda_2'\|x\|^2 < \infty$ and $\|y\| \leq C\|x\|$ for some $\lambda_1', \lambda_2', C \in \mathbb{R}$ such that $0 < \lambda_1' \leq \lambda_2' < \infty$ and $C > 0$, then there exists a symmetric positive definite matrix M such that $y = Mx$ and $0 < \lambda_1 \leq \lambda_{(1)}(M) \leq \lambda_{(d)}(M) \leq \lambda_2 < \infty$ for some constants $\lambda_1, \lambda_2 \in \mathbb{R}$ depending only on λ_1', λ_2' and C.

At each time $k \in \mathbb{N}$, the observer should be able to calculate the gain vector at $(X_k, \boldsymbol{X}_{k-1})$ and an estimator $\hat{\theta}_k$, $G_k(X_k, \hat{\theta}_k|\boldsymbol{X}_{k-1})$, in order use it to update the estimate $\hat{\theta}_k$. In Section 4 we will show how gain functions can be constructed, but before that we present in the next section our tracking algorithm based on the gain function and our main result describing the quality of the algorithm.

3 Main result

Introduce a recursive algorithm for tracking the sequence $\theta_k = \theta_k(\boldsymbol{X}_{k-1}) \in \Theta \subset \mathbb{R}^d$ from the observations (1):

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \gamma_k G_k(X_k, \hat{\theta}_k|\boldsymbol{X}_{k-1}), \quad k \in \mathbb{N}, \tag{7}$$

for some positive sequence of step sizes $\gamma_k \leq \Gamma$ and some (arbitrary) initial value $\hat{\theta}_0 \in \Theta \subset \mathbb{R}^d$.

Heuristically, since the gain vector $G_k(X_k, \hat{\theta}_k|\boldsymbol{X}_{k-1})$ moves, on average, $\hat{\theta}_k$ towards θ_k and the sequence $\theta_k \in \Theta$ (since Θ is compact) is bounded, the resulting estimating sequence $\hat{\theta}_k$ should also be well-behaved. The following lemma states that the second moment of $\hat{\theta}_k$ is uniformly bounded in $k \in \mathbb{N}$.

Lemma 2. *For sufficiently small γ_k there exists a constant $\bar{C}_\Theta$ such that*

$$\mathbb{E}\|\hat{\theta}_k\|_2^2 \leq \bar{C}_\Theta^2, \qquad k \in \mathbb{N}.$$

The proof of this lemma is given in the Section 7. In fact, it is enough to assume that γ_k is sufficiently small for all $k \geq N$ for some fixed $N \in \mathbb{N}$. This lemma will be used in the proof of the main theorem below.

Theorem 1. *Let Assumptions (A1) and (A2) hold and $p \geq 1$. Let the tracking sequence $\hat{\theta}_k$ be defined by (7) with the sequence γ_k satisfying the conditions of Lemma 2, $\delta_k = \delta_k(\boldsymbol{X}_{k-1}) = \hat{\theta}_k - \theta_k$ and $\Delta_k = \Delta_k(\boldsymbol{X}_k) = \theta_k - \theta_{k+1}$, $k \in \mathbb{N}$. Then for any $k_0, k \in \mathbb{N}$ such that $k_0 \leq k$ and $\gamma_i\lambda_2 < 1$ for all $k_0 \leq i \leq k$, the following relation holds:*

$$\mathbb{E}\|\delta_{k+1}\|_p \leq C_1 \exp\Big(-\frac{\lambda_1}{2}\sum_{i=k_0}^{k}\gamma_i\Big) + C_2\Big(\sum_{i=k_0}^{k-1}\gamma_i^2\Big)^{1/2} + C_3 \max_{i=k_0,\dots,k}\mathbb{E}\|\theta_{i+1} - \theta_{k_0}\|_2, \tag{8}$$

where $C_1 = (2d)^{1/2}\big(\bar{C}_\Theta + C_\Theta\big)$, $C_2 = d^{1/2}C^{1/2}\big(1+\lambda_2/\lambda_1\big)$, $C_3 = d^{1/2}\big(1+\lambda_2/\lambda_1\big)$ *and* C *is from Assumption (A2).*

If, in addition, $\Lambda_{(1)}(M_k) \geq \lambda_1$ *(in Assumption (A1)) and* $|G_k(X_k, \hat{\theta}_k|\boldsymbol{X}_{k-1})| \leq C$ *almost surely, then for any* $k_0, k \in \mathbb{N}$ *such that* $k_0 \leq k$ *and* $\gamma_i\lambda_2 < 1$ *for all* $k_0 \leq i \leq k$*,*

$$\mathbb{E}\|\delta_{k+1}\|_p^p \leq C_1 \exp\Big(-p\lambda_1 \sum_{i=k_0}^{k} \gamma_i\Big) + C_2\Big(\sum_{i=k_0}^{k-1}\gamma_i^2\Big)^{p/2} + C_3 \max_{i=k_0,\dots,k} \mathbb{E}\|\theta_{i+1} - \theta_{k_0}\|_p^p, \tag{9}$$

where $C_1 = 2^{p-1}K_p^p\mathbb{E}\|\delta_{k_0}\|_p^p$, $C_2 = d\,2^{2p-1}B_pC^p\big(1+K_p^2\lambda_2/\lambda_1\big)^p$ *and* $C_3 = 2^{p-1}\big(1+K_p^2\lambda_2/\lambda_1\big)^p$.

Proof. For the sake of brevity, denote $\theta_k = \theta_k(\boldsymbol{X}_{k-1})$, $G_k = G(X_k, \hat{\theta}_k|\boldsymbol{X}_{k-1})$ and $g_k = g(\hat{\theta}_k, \theta_k|\boldsymbol{X}_{k-1})$, $k \in \mathbb{N}$. Recall that $\mathcal{F}_k = \sigma(\boldsymbol{X}_k)$ is the σ-field generated by $\boldsymbol{X}_k = (X_0, X_2, \dots, X_k)$.

We have

$$\mathbb{E}[G_k|\mathcal{F}_{k-1}] = g_k(\hat{\theta}_k, \theta_k|\boldsymbol{X}_{k-1}) = g_k, \quad k \in \mathbb{N}.$$

It follows that $D_k = G_k - g_k$, $k \in \mathbb{N}$, is a (vector) martingale difference sequence with respect to the filtration $\{\mathcal{F}_k,\ k \in \mathbb{N}_0\}$.

Rewrite the algorithm equation (7) as

$$\delta_{k+1} = \delta_k + \Delta\theta_k + \gamma_k D_k + \gamma_k g_k, \quad k \in \mathbb{N}.$$

In view of Assumption (A1), we have the decomposition $g_k = -M_k\delta_k$, with a symmetric positive definite matrix $M_k = M(\hat{\theta}_k, \theta_k|\boldsymbol{X}_{k-1})$ so that

$$\delta_{k+1} = \Delta\theta_k + \gamma_k D_k + (I - \gamma_k M_k)\delta_k, \quad k \in \mathbb{N}. \tag{10}$$

By iterating the above relation, we obtain that for any $k_0 = 0, \dots, k$

$$\begin{aligned}\delta_{k+1} &= (1-\gamma_k M_k)(I - \gamma_{k-1}M_{k-1})\delta_{k-1} + \Delta\theta_k + \gamma_k D_k \\ &\quad + (1-\gamma_k M_k)(\Delta\theta_{k-1} + \gamma_{k-1}D_{k-1}) \\ &= \Big[\prod_{i=k_0}^{k}(I - \gamma_i M_i)\Big]\delta_{k_0} + \sum_{i=k_0}^{k}\Big[\prod_{j=i+1}^{k}(I-\gamma_j M_j)\Big](\Delta\theta_i + \gamma_i D_i).\end{aligned} \tag{11}$$

Denote $A_i = \sum_{j=k_0}^{i}\gamma_j D_j$, $B_i = \sum_{j=k_0}^{i}\Delta\theta_j$ and $C_i = A_i + B_i$. Applying the vector version of the Abel transformation (Lemma 4) to the second term of the right hand side of (11) yields

$$\sum_{i=k_0}^{k}\Big[\prod_{j=i+1}^{k}(I-\gamma_j M_j)\Big](\Delta\theta_i + \gamma_i D_i) = C_k - \sum_{i=k_0}^{k-1}\gamma_{i+1}M_{i+1}\Big[\prod_{j=i+2}^{k}(I-\gamma_j M_j)\Big]C_i. \tag{12}$$

Note in particular that, if we take $M_j = \lambda_1$ for $j = k_0, \dots, k$, $\Delta\theta_j = 0$, for $j = k_0, \dots, k$, $D_{k_0} = 1$ and $D_j = 0$ for $j = k_0+1, \dots, k$, we derive that (if $0 \leq \gamma_j\lambda_1 \leq 1$ for $j = k_0, \dots, k$)

$$\sum_{i=k_0}^{k-1}\lambda_1\gamma_{i+1}\prod_{j=i+2}^{k}(1-\gamma_j\lambda_1) = 1 - \prod_{j=k_0+1}^{k}(1-\gamma_j\lambda_1) \leq 1, \tag{13}$$

which we will use later.

Using (12), we can rewrite our expansion of δ_{k+1} in (11) as

$$\delta_{k+1} = \Big[\prod_{i=k_0}^{k}(I-\gamma_i M_i)\Big]\delta_{k_0} + C_k - \sum_{i=k_0}^{k-1}\gamma_{i+1}M_{i+1}\Big[\prod_{j=i+2}^{k}(I-\gamma_j M_j)\Big]C_i.$$

Take $p \in \mathbb{N}$. The previous display, the triangle inequality and the sub-multiplicative property of the operator norm ($\|MN\|_p \le \|M\|_p\|N\|_p$) imply that

$$\begin{aligned}\|\delta_{k+1}\|_p \le \|\delta_{k_0}\|_p \prod_{i=k_0}^{k}\|I-\gamma_i M_i\|_p + \|C_k\|_p \\ + \sum_{i=k_0}^{k-1}\gamma_{i+1}\|M_{i+1}\|_p\|C_i\|_p \prod_{j=i+2}^{k}\|I-\gamma_j M_j\|_p.\end{aligned} \tag{14}$$

Due to Assumption (A1), the matrix M_i has smallest and largest eigenvalues $\Lambda_{(1),i}$ and $\Lambda_{(d),i}$, respectively, such that almost surely $\gamma_i\Lambda_{(d),i} \le \gamma_i\lambda_2 < 1$, $k_0 \le i \le k$, and $\mathbb{E}[\Lambda_{(1),i}|\mathcal{F}_{i-2}] \ge \lambda_1 > 0$. By using Lemma 3 and the fact that $\gamma_i\Lambda_{(1),i}^2 \le \gamma_i\Lambda_{(d),i}\Lambda_{(1),i} \le \Lambda_{(1),i}$ almost surely, we evaluate:

$$\begin{aligned}\mathbb{E}\Big[(1-\gamma_k\Lambda_{(1),k})^2\Big|\mathcal{F}_{k-2}\Big] &= \mathbb{E}\Big[1-2\gamma_k\Lambda_{(1),k}+\gamma_k^2\Lambda_{(1),k}^2\Big|\mathcal{F}_{k-2}\Big] \\ &= 1-2\gamma_k\mathbb{E}\big[\Lambda_{(1),k}\big|\mathcal{F}_{k-2}\big]+\mathbb{E}\big[\gamma_k^2\Lambda_{(1),k}^2\big|\mathcal{F}_{k-2}\big] \\ &\le 1-\gamma_k\mathbb{E}\big[\Lambda_{(1),k}\big|\mathcal{F}_{k-2}\big] \le 1-\gamma_k\lambda_1,\end{aligned}$$

almost surely. Similarly, $\mathbb{E}\big[1-\gamma_k\Lambda_{(1),k}\big|\mathcal{F}_{k-2}\big] \le 1-\gamma_k\lambda_1$ almost surely. It then follows that

$$\begin{aligned}\mathbb{E}\prod_{i=k_0}^{k}\|1-\gamma_i M_i\|_2^2 = \mathbb{E}\prod_{i=k_0}^{k}(1-\gamma_i\Lambda_{(1),i})^2 &= \mathbb{E}\mathbb{E}\Big[\prod_{i=k_0}^{k}(1-\gamma_i\Lambda_{(1),i})^2\Big|\mathcal{F}_{k-2}\Big] \\ &= \mathbb{E}\mathbb{E}\Big[(1-\gamma_k\Lambda_{(1),k})^2\Big|\mathcal{F}_{k-2}\Big]\prod_{i=k_0}^{k-2}(I-\gamma_i\Lambda_{(1),i})^2 \\ &\le (1-\gamma_k\lambda_1)\mathbb{E}\prod_{i=k_0}^{k-1}(I-\gamma_i\Lambda_{(1),i})^2 \le \prod_{i=k_0}^{k}(1-\gamma_i\lambda_1),\end{aligned} \tag{15}$$

by iterating the recursion.

Let D_{kl} denote the l-th coordinate of the vector D_k. Clearly, for each $l = 1,\dots,d$, $\{D_{kl},\, k \in \mathbb{N}\}$ is a martingale difference with respect to the filtration $\{\mathcal{F}_k,\, k \in \mathbb{N}_0\}$. Using the fact that martingale increments are uncorrelated, we derive that for all $i = k_0,\dots,k$

$$\mathbb{E}\|A_i\|_2^2 = \mathbb{E}\sum_{l=1}^{d}\Big(\sum_{j=k_0}^{i}\gamma_j D_{jl}\Big)^2 = \sum_{l=1}^{d}\sum_{j=k_0}^{i}\gamma_j^2\mathbb{E}D_{jl}^2 = \sum_{j=k_0}^{i}\gamma_j^2\mathbb{E}\|D_j\|_2^2 \le C\sum_{j=k_0}^{k}\gamma_j^2.$$

Since B_i is a telescopic sum, we also have, for all $p \in \mathbb{N}$ and $i = k_0, \dots, k$,

$$\mathbb{E}\|B_i\|_p = \mathbb{E}\Big\| \sum_{j=k_0}^{i} \Delta\theta_j \Big\|_p = \mathbb{E}\|\theta_{i+1} - \theta_{k_0}\|_p \le \max_{i=k_0,\dots,k} \mathbb{E}\|\theta_{i+1} - \theta_{k_0}\|_p.$$

Since $\|C_i\|_2$ is $\mathcal{F}_j$-measurable for all $j \ge i$, it follows that

$$\begin{aligned}
\mathbb{E}\Big[\|C_i\|_2 \prod_{j=i+2}^{k} \|I - \gamma_j M_j\|_2\Big] &= \mathbb{E}\mathbb{E}\Big[\|C_i\|_2 \prod_{j=i+2}^{k} \|I - \gamma_j M_j\|_2 \Big| \mathcal{F}_{k-2}\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[1 - \gamma_k \Lambda_{(1),j} \big| \mathcal{F}_{k-2}\big] \|C_i\|_2 \prod_{j=i+2}^{k-1} (1 - \gamma_j \Lambda_{(1),j})\Big] \\
&\le (1 - \gamma_k \lambda_1)\mathbb{E}\Big[\|C_i\|_2 \prod_{j=i+2}^{k-1} (1 - \gamma_j \Lambda_{(1),j})\Big] \le \mathbb{E}\|C_i\|_2 \prod_{j=i+2}^{k} (1 - \gamma_j \lambda_1).
\end{aligned}$$

Combining the last three displays, relations (13), (14) and (15), Lemma 3, the Hölder and triangle inequalities and the elementary inequality $1 - x \le e^{-x}$, we finally get that

$$\begin{aligned}
&\mathbb{E}\|\delta_{k+1}\|_2 \\
&\le \Big(\mathbb{E}\|\delta_{k_0}\|_2^2 \, \mathbb{E} \prod_{i=k_0}^{k} \|I - \gamma_i M_i\|_2^2\Big)^{1/2} + \mathbb{E}\|C_k\|_2 + \mathbb{E}\Big[\sum_{i=k_0}^{k-1} \gamma_{i+1} \|M_{i+1}\|_2 \|C_i\|_2 \prod_{j=i+2}^{k} \|I - \gamma_j M_j\|_2\Big] \\
&\le \Big(\mathbb{E}\|\delta_{k_0}\|_2^2 \prod_{i=k_0}^{k} (1 - \gamma_i \lambda_1)\Big)^{1/2} + \mathbb{E}\|C_k\|_2 + \sum_{i=k_0}^{k-1} \gamma_{i+1} \lambda_2 \mathbb{E}\Big[\|C_i\|_2 \prod_{j=i+2}^{k} \|I - \gamma_j M_j\|_2^2\Big] \\
&\le \big(\mathbb{E}\|\delta_{k_0}\|_2^2\big)^{1/2} \exp\Big(-\frac{\lambda_1}{2} \sum_{i=k_0}^{k} \gamma_j\Big) + \max_{i=k_0,\dots,k} \mathbb{E}\|C_i\|_2 \Big(1 + \sum_{i=k_0}^{k-1} \gamma_{i+1} \lambda_2 \prod_{j=i+2}^{k} (1 - \gamma_j \lambda_1)\Big) \\
&\le \sqrt{2}\Big(\bar{C}_\Theta + C_\Theta\Big) \exp\Big(-\frac{\lambda_1}{2} \sum_{i=k_0}^{k} \gamma_i\Big) + \Big(1 + \frac{\lambda_2}{\lambda_1}\Big)\Big(\Big(C \sum_{i=k_0}^{k} \gamma_i^2\Big)^{1/2} + \max_{i=k_0,\dots,k} \mathbb{E}\|\theta_{i+1} - \theta_{k_0}\|_2\Big),
\end{aligned}$$

since $\mathbb{E}\|\delta_{k_0}\|_2^2 \le 2\mathbb{E}\|\hat{\theta}_{k_0}\|_2^2 + 2\mathbb{E}\|\theta_{k_0}\|_2^2 \le 2(\bar{C}_\Theta^2 + C_\Theta^2)$, by (2) and Lemma 2. Note that $\|\delta_{k_0}\|_2 \ge \|\delta_{k_0}\|_p$ for $p \ge 2$, $d^{1/2}\|\delta_{k_0}\|_2 \ge \|\delta_{k_0}\|_p$ for $1 \le p < 2$. We have established the first statement of the theorem.

Let now the components of the gain G_k be almost surely bounded, in absolute value, by a certain constant C. Using Lemma 3 and the elementary inequality $1 - x \le e^{-x}$, we have that, for each $p \in \mathbb{N}$, and then some constant K_p, we can derive the following alternative expression to (14).

$$\begin{aligned}
\|\delta_{k+1}\|_p &\le K_p \|\delta_{k_0}\|_p \prod_{i=k_0}^{k} (1 - \gamma_i \lambda_1) + \max_{i=k_0,\dots,k} \|C_i\|_p \Big(1 + K_p^2 \sum_{i=k_0}^{k-1} \gamma_{i+1} \lambda_2 \prod_{j=i+2}^{k} (1 - \gamma_i \lambda_1)\Big) \\
&\le K_p \|\delta_{k_0}\|_p \exp\Big(-\lambda_1 \sum_{i=k_0}^{k} \gamma_j\Big) + \Big(1 + K_p^2 \frac{\lambda_2}{\lambda_1}\Big) \max_{i=k_0,\dots,k} \|C_i\|_p,
\end{aligned}$$

where we again use (13). Take now the p-th power ($p \geq 1$) of both sides of the inequality and apply the Hölder inequality $(\sum_{i=1}^m a_i)^p \leq m^{p-1} \sum_{i=1}^m |a_i|^p$ for $m = 2$ to get

$$\|\delta_{k+1}\|_p^p \leq 2^{p-1} K_p^p \|\delta_{k_0}\|_p^p \exp\left(- p\lambda_1 \sum_{i=k_0}^{k} \gamma_j \right) + 2^{p-1} \left(1 + K_p^2 \frac{\lambda_2}{\lambda_1}\right)^p \max_{i=k_0,\dots,k-1} \|C_i\|_p^p,$$

Remember that the sequence $\left\{ \sum_{j=k_0}^{i} \gamma_j D_j(\boldsymbol{X}_j, \hat{\theta}_j, \theta_j),\ i \geq k_0 \right\}$ is a martingale with respect to the filtration $\{\mathcal{F}_i, i \in \mathbb{N}\}$ and that the entries of D_j verify $|D_{jl}| \leq 2C$, almost surely. Applying the maximal Burkholder-Davis-Gundy inequality (cf. Chow and Teicher [1988, Theorem 1, p.407]) we conclude that for any $p \geq 1$, with $B_p = ((18p^{5/2})/(p-1)^{3/2})^p$,

$$\begin{aligned}\mathbb{E} \max_{i=k_0,\dots,k-1} \|A_i\|_p^p &= \mathbb{E} \max_{i=k_0,\dots,k-1} \sum_{l=1}^{d} \left| \sum_{j=k_0}^{i} \gamma_j D_{jl} \right|^p \leq \sum_{l=1}^{d} \mathbb{E} \max_{i=k_0,\dots,k-1} \left| \sum_{j=k_0}^{i} \gamma_j D_{jl} \right|^p \\ &\leq B_p \sum_{l=1}^{d} \mathbb{E} \left| \sum_{j=k_0}^{k-1} \gamma_j^2 D_{jl}^2 \right|^{p/2} \leq d B_p 2^p C^p \left| \sum_{j=k_0}^{k-1} \gamma_j^2 \right|^{p/2},\end{aligned}$$

The second inequality of the theorem now follows by taking expectations on both sides of the bound on $\|\delta_{k+1}\|_p^p$ above, by using the last inequality, (13) and the fact that $\|C_i\|_p^p \leq 2^{p-1}\|A_i\|_p^p + 2^{p-1}\|B_i\|_p^p$. □

Remark 1. Sometimes we will not be interested in tracking the, say, *natural* parameter θ_k of the model but some other parameter ϑ_k which is, on average, close to θ_k. The difference $\|\theta_k - \vartheta_k\|_p$ can be seen as an approximation term in that the parameter θ_k driving the time series is actually an approximation for our parameter of interest ϑ_k. Denoting $\hat{\theta}_k - \vartheta_k$ as δ_k^*, the following expansion can be derived,

$$\begin{aligned}\delta_{k+1}^* &= \delta_k^* + \Delta\vartheta_k + \gamma_k D_k - \gamma_k M_k(\hat{\theta}_k - \theta_k) \\ &= \Delta\vartheta_k + \gamma_k M_k(\theta_k - \vartheta_k) + \gamma_k D_k + (I - \gamma_k M_k)\delta_k^* \\ &= \Big[\prod_{i=k_0}^{k} (I - \gamma_i M_i) \Big] \delta_{k_0}^* + \sum_{i=k_0}^{k} \Big[\prod_{j=i+1}^{k} (I - \gamma_j M_j) \Big] (\Delta\vartheta_k + \gamma_k M_k(\theta_k - \vartheta_k) + \gamma_i D_i).\end{aligned}$$

The same note could be made for situations where $g_k = -M_k(\hat{\theta}_k - \theta_k - \eta_k)$ where η_k is a remainder term which may be random so long as it is measurable with respect to $\sigma(\boldsymbol{X}_{k-1})$; it would then follow that

$$\delta_{k+1} = \Big[\prod_{i=k_0}^{k} (I - \gamma_i M_i) \Big] \delta_{k_0} + \sum_{i=k_0}^{k} \Big[\prod_{j=i+1}^{k} (I - \gamma_j M_j) \Big] (\Delta\theta_i - \eta_i + \gamma_i D_i).$$

Noting that $\|\gamma_k M_k\,(\theta_k - \vartheta_k)\|_p < \lambda_2 \gamma_k K_p \|\theta_k - \vartheta_k\|_p$ we conclude, for the same constants

C_1, C_2, C_3 as before and all $p \in \mathbb{N}$, that the following also hold

$$\begin{aligned}\mathbb{E}\|\delta_{k+1}\|_p \leq & C_1 \exp\Big(-\frac{\lambda_1}{2}\sum_{i=k_0}^{k}\gamma_j\Big) + C_2\Big(\sum_{i=k_0}^{k-1}\gamma_i^2\Big)^{1/2} \\ & + C_3\mathbb{E}\max_{i=k_0,\dots,k}\|\vartheta_{i+1}-\vartheta_{k_0}\|_2 + \lambda_2 K_2 \mathbb{E}\sum_{i=k_0}^{k}\gamma_i\|\eta_i\|_2,\end{aligned} \tag{16}$$

$$\begin{aligned}\mathbb{E}\|\delta_{k+1}\|_p^p \leq & C_1 \exp\Big(-p\lambda_1\sum_{i=k_0}^{k}\gamma_j\Big) + C_2\Big(\sum_{i=k_0}^{k-1}\gamma_i^2\Big)^{p/2} \\ & + C_3\mathbb{E}\Big(\max_{i=k_0,\dots,k}\|\vartheta_{i+1}-\vartheta_{k_0}\|_p + \lambda_2 K_p \sum_{i=k_0}^{k}\gamma_i\|\eta_i\|_p\Big)^p,\end{aligned} \tag{17}$$

where either a) $\delta_k = \hat{\theta}_k - \vartheta_k$ and $\eta_k = \theta_k - \vartheta_k$, b) $\delta_k = \hat{\theta}_k - \theta_k$, $\vartheta_k = \theta_k$ and η_k such that $g_k = -M_k(\delta_k - \eta_k)$; (16) and (17) generalize then the bounds in (8) and (9) where we had c) $\delta_k = \hat{\theta}_k - \theta_k$, $\vartheta_k = \theta_k$ and $\eta_k = 0$.

Remark 2. If we are interested in tracking $\vartheta_k = \varphi(\theta_k)$, a smooth functional of the parameter θ_k, then by using Taylor's Theorem, our Theorem 1 above straightforwardly delivers a bound on the expectation of $\|\hat{\vartheta}_k - \vartheta_k\|_p = \|\varphi(\hat{\theta}_k) - \varphi(\theta_k)\|_p$ and its powers.

4 Construction of gain functions

In this section we address the construction, or choice, of appropriate gain functions to be used with the algorithm (7). Any gain function for which conditions (A1) and (A2) hold may be used with our algorithm, and whether a particular gain function is suitable or not depends exclusively on the the model under study. Namely, this will depend on the way in which the distributions in the model depend on the parameter which we are interested in tracking. For certain types of models, there might be natural choices for the gain function. As before we abbreviate $\theta_k = \theta_k(\boldsymbol{X}_{k-1})$.

A situation, which essentially extends the original setup in which Robbins and Monro [1951] developed their classical algorithm, is when the data, $\boldsymbol{X}_k = (X_1, \dots, X_k)$, is such that

$$X_k = \vartheta_k(\boldsymbol{X}_{k-1}) + \xi_k(\boldsymbol{X}_{k-1}),$$

where the $\vartheta_k(\cdot)$ are functions of $\boldsymbol{X}_{k-1}$, and $\xi_k(\boldsymbol{X}_{k-1})$ are martingale difference noise terms which may also depend on $\boldsymbol{X}_{k-1}$. In this case, given $\boldsymbol{X}_{k-1}$ we may simply take

$$G_k(x, \theta|\boldsymbol{X}_{k-1}) = x - \theta \tag{18}$$

since for each θ,

$$g_k(\theta, \vartheta_k(\boldsymbol{X}_{k-1})|\boldsymbol{X}_{k-1}) = \mathbb{E}_\vartheta[G_k(X_k, \theta|X_{k-1})|\boldsymbol{X}_{k-1}] = -(\theta - \vartheta_k(\boldsymbol{X}_{k-1})).$$

Non-parametric regression is an example of a model which fits into this situation and for which our results may be used.

It could also be that $\mathbb{E}_\theta[X_k|\boldsymbol{X}_{k-1}]$, the conditional expectation of the data, given the past, is not θ but instead $\varphi(\theta)$ for some some smooth function φ. In this case, given $\boldsymbol{X}_{k-1}$, one should consider instead,

$$G_k(x,\theta|\boldsymbol{X}_{k-1}) = x - \varphi(\theta) \tag{19}$$

and then, for each θ,

$$g_k(\theta,\vartheta_k|\boldsymbol{X}_{k-1}) = \mathbb{E}_\vartheta[G_k(X_k,\theta|X_{k-1})|\boldsymbol{X}_{k-1}] = -\big(\varphi(\theta) - \varphi(\vartheta_k)\big).$$

The term on the far right should then be comparable to $-(\theta-\vartheta_k)$. Autoregressive models, for example, fall into this category (cf. 6.4).

One may also consider more dynamical situations where the observations themselves depend on our tracking sequence. An example of such a setup is the Kiefer and Wolfowitz [1952] algorithm where we would like to track the sequence of (unique) maxima of a sequence of functions $\vartheta_k : \Theta \subset \mathbb{R}^d \mapsto \mathbb{R}$, $k \in \mathbb{N}$, which we may observe at any point, corrupted with white noise. One possibility (cf. Kushner and Yin [2003]) is to use gain functions defined using random directions. Let then D_k, $k \in \mathbb{N}$, denote a random sequence of independent unit vectors. We would consider, for a positive sequence e_k, $k \in \mathbb{N}$, the gain function

$$G_k(X_k^-, X_k^+, \hat\theta_k|\boldsymbol{X}_{k-1}^-, \boldsymbol{X}_{k-1}^+, \boldsymbol{D}_{k-1}) = D_k\frac{X_k^-(\hat\theta_k) - X_k^+(\hat\theta_k)}{2e_k}, \tag{20}$$

where, with some abuse of notation, the observations $X_{k+1}^\pm(\theta_k)$, are given by

$$X_{k+1}^\pm(\hat\theta_k) = \vartheta_k\big(\hat\theta_k \pm e_kD_k\big) + \xi_k^\pm,$$

for $\hat\theta_k$ the tracking sequence defined by the gain (20) and $\xi_k^\pm$ independent, zero mean noise. Let θ_k be the unique maxima of $\vartheta_k(\cdot)$. In this case we would have, for the filtration $\mathcal{F}_k = \sigma\big(\boldsymbol{X}_k^\pm, \boldsymbol{D}_k\big)$,

$$\begin{aligned} g_k(\hat\theta_k,\theta_k|\mathcal{F}_{k-1}) &= \mathbb{E}\Big[-D_kD_k^T\nabla\vartheta_k(\hat\theta_k) + H_k(\hat\theta_k) + D_k\frac{\xi_k^- - \xi_k^+}{2e_k}\Big|\mathcal{F}_{k-1}\Big] \\ &= -\mathbb{E}\big[D_kD_k^T\big]\nabla\vartheta_k(\hat\theta_k) + \mathbb{E}\big[H_k(\hat\theta_k)|\mathcal{F}_{k-1}\big] + \frac{\mathbb{E}\big[D_k(\xi_k^- - \xi_k^+)\big]}{2e_k} \\ &= -\mathbb{E}\big[D_kD_k^T\big]\nabla^2\vartheta_k(\theta_k^*)(\hat\theta_k - \theta_k) + \eta_k, \end{aligned}$$

where $\nabla^2\vartheta_k(\cdot)$ is the Hessian of $\vartheta_k(\cdot)$, $\theta_k^* \in \Theta$ and, for $\theta \in \Theta$,

$$H_k(\theta) = D_kD_k^T\nabla\vartheta_k(\theta) - D_k\frac{\vartheta_k(\theta + e_kD_k) - \vartheta_k(\theta - e_kD_k)}{2e_k}.$$

Conditions (A1) and (A2) will hold if, for example, we assume that the random directions where chosen such that $\mathbb{E}\big[D_kD_k^T\big]$ are positive definite matrices, that the Hessian $\nabla^2\vartheta_k(\cdot)$

is positive definite over Θ and that for appropriately small e_k the expectation $\mathbb{E}[\|\eta_k\|_p]$ is appropriately small, uniformly over $\theta \in \Theta$. These conditions are comparable to the ones in the original formulation of the Kiefer-Wolfowitz algorithm, and can be significantly relaxed by, for example, considering different types of expansions for g_k depending on how large the norm of $\delta_k = \hat{\theta}_k - \theta_k$ is.

Consider now a different example. Say $\mathcal{X} \subset \mathbb{R}$ and, given the past of the process, $\boldsymbol{X}_{k-1}$, we would like to track a conditional quantile of a certain distribution, i.e., we would like to track $\vartheta_k = \vartheta_k(\boldsymbol{X}_{k-1})$ such that $\vartheta_k = \inf\big\{x \in \mathcal{X} : F_k(x|\boldsymbol{X}_{k-1}) \geq \alpha_k\big\}$, where α_k is a sequence in $(0,1)^{\mathbb{N}}$ of our choice and $F_k(\cdot|\boldsymbol{X}_{k-1})$ the cumulative distribution function of $X_k|\boldsymbol{X}_{k-1}$. In this case it makes sense to use

$$G_k(x,\theta|\boldsymbol{X}_{k-1}) = \alpha_k - I\{x - \theta \leq 0\} \tag{21}$$

since we see that

$$g_k(\theta,\vartheta_k|\boldsymbol{X}_{k-1}) = \mathbb{E}[G_k(X_k,\theta|\boldsymbol{X}_{k-1})|\boldsymbol{X}_{k-1}] = -\big(F_k(\theta - \vartheta_k|\boldsymbol{X}_{k-1}) - \alpha_k\big),$$

where we assume w.l.g. that the distribution is centred at the quantile ϑ_k. The quantity in the last display clearly has the same sign as $\vartheta_k - \theta$. Note also that the algorithm based on this gain function only requires knowledge of the values of the indicators $\mathbf{1}\{X_k - \theta \leq 0\}$ which means that we may still track the required quantiles without explicitly observing X_k. This problem is treated in detail for the case of independent observations in Belitser and Serra [2013].

For certain models it might, however, not be obvious how gain functions can be constructed, especially when tracking multi-dimensional parameters. It is therefore important to have a general procedure that can be used to construct candidate gain functions that can either be used directly or, if needed, modified to verify (A1) and (A2).

Assume that for each $k \in \mathbb{N}$, each distribution from the family of conditional distributions $\mathcal{P}_k = \{P_\theta(x|\boldsymbol{X}_{k-1}), \theta \in \Theta\}$ has a density with respect to some σ-finite dominating measure μ and denote this conditional density by $p_\theta(x|\boldsymbol{X}_{k-1})$, $\theta = (\theta_1,\ldots,\theta_d) \in \Theta$. Assume also that there is a common support $\mathcal{X}$ for these densities, and that for any $x \in \mathcal{X}$ and $\theta \in \Theta \subset \mathbb{R}^d$, the partial derivatives $\partial p_\theta(x|\boldsymbol{X}_{k-1})/\partial\theta_i$, $i = 1,\ldots,d$, exist and are finite, almost surely. Under these assumptions, the *conditional* gradient vector

$$\nabla_\theta \log p_\theta(x|\boldsymbol{X}_{k-1}) = \Big(\partial \log p_\theta(x|\boldsymbol{X}_{k-1})/\partial\theta_1, \ldots, \partial \log p_\theta(x|\boldsymbol{X}_{k-1})/\partial\theta_d\Big) \tag{22}$$

and the square, random matrices $I_k(\theta|\boldsymbol{X}_{k-1})$ with entries

$$I_{k,i,j}(\theta|\boldsymbol{X}_{k-1}) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i}p_\theta(x|\boldsymbol{X}_{k-1}) \cdot \frac{\partial}{\partial\theta_j}p_\theta(x|\boldsymbol{X}_{k-1})\right] \tag{23}$$

for $i,j = 1,\ldots,d$, can be defined, almost surely. A possible gain function is simply the conditional score of the model, i.e. the gradient vector

$$G_k(x,\theta|\boldsymbol{X}_{k-1}) = \nabla_\theta \log p_\theta(x|\boldsymbol{X}_{k-1}). \tag{24}$$

If (23) is almost surely non-singular then one might also consider

$$G_k(x,\theta|\boldsymbol{X}_{k-1}) = I_k^{-1}(\theta|\boldsymbol{X}_{k-1})\nabla_\theta \log p_\theta(x|\boldsymbol{X}_{k-1}). \tag{25}$$

We justify now why these choices are reasonable. Take $\vartheta = (\vartheta_1,\ldots,\vartheta_d) \in \Theta$. It is not uncommon for the Kullback-Leibler divergence $K\big(P_\vartheta(x|\boldsymbol{X}_{k-1}), P_\theta(x|\boldsymbol{X}_{k-1})\big)$ to be a quadratic form in the distance between the parameters θ and ϑ, i.e., equal to a multiple of $(\theta-\vartheta)^T M(\theta-\vartheta)$ for some (eventually random) positive semi-definite matrix M. If so, under the assumption that we can interchange integration and differentiation and that M does not depend on θ, $g_k(\theta,\vartheta|\boldsymbol{X}_{k-1})$ will almost surely reduce to

$$\begin{aligned}
&\int \nabla_\theta \log p_\theta(x|\boldsymbol{X}_{k-1}) dP_\vartheta(x|\boldsymbol{X}_{k-1}) = \nabla_\theta \int \log p_\theta(x|\boldsymbol{X}_{k-1}) dP_\vartheta(x|\boldsymbol{X}_{k-1})\\
=&\nabla_\theta\Big(\int \log \frac{p_\theta(x|\boldsymbol{X}_{k-1})}{p_\vartheta(x|\boldsymbol{X}_{k-1})} dP_\vartheta(x|\boldsymbol{X}_{k-1}) + \int \log p_\vartheta(x|\boldsymbol{X}_{k-1}) dP_\vartheta(x|\boldsymbol{y}_{k-1})\Big)\\
=&\nabla_\theta \int \log \frac{p_\theta(x|\boldsymbol{X}_{k-1})}{p_\vartheta(x|\boldsymbol{X}_{k-1})} dP_\vartheta(x|\boldsymbol{X}_{k-1}) = -\nabla_\theta K\big(P_\vartheta(x|\boldsymbol{X}_{k-1}), P_\theta(x|\boldsymbol{X}_{k-1})\big)\\
=&-\nabla_\theta(\theta-\vartheta)^T M(\theta-\vartheta) = -2M(\theta-\vartheta).
\end{aligned}$$

The score will in principle depend on the past of the chain $\boldsymbol{X}_{k-1}$ and the previous argument might only be valid for a certain subset of values $\boldsymbol{X}_{k-1}$ in $\mathcal{X}^{k-1}$. This dependence could prevent (A1) from holding. In these cases, using the form (25) might be a good alternative since the matrix $I_k^{-1}(\theta|\boldsymbol{X}_{k-1})$ will act as an appropriate scaling factor.

The dependence of the gain function on the past of the time series is in fact one of the main issues one has to deal with when checking (A1) and (A2). On one hand, to ensure that the gain function has, on average, the right direction, as required by (3), the gain will often need to depend on previous observations. This might, however, affect either the range or the variance of the gain. Gain function, such as (24) and (25), can be modified, or rescaled, to ensure that the respective conditional expectation $g_k(\theta,\vartheta|\boldsymbol{X}_{k-1})$ verifies the assumptions of Theorem 1. One can for example truncate certain entries or factors in both $G_k(x,\theta|\boldsymbol{X}_{k-1})$ and $I_k(\theta|\boldsymbol{X}_{k-1})$ to ensure that the resulting $g_k(\theta,\vartheta|\boldsymbol{X}_{k-1})$ follows the required assumptions. Another possibility is to rescale, or directly truncate, the length of a given gain vector and consider, for example, one of the following gains

$$\begin{aligned}
\tilde{G}_k(x,\theta|\boldsymbol{X}_{k-1}) &= \frac{G_k(x,\theta|\boldsymbol{X}_{k-1})}{1+\|G_k(x,\theta|\boldsymbol{X}_{k-1})\|_2},\\
\mathring{G}_k(x,\theta|\boldsymbol{X}_{k-1}) &= G_k(x,\theta|\boldsymbol{X}_{k-1})\Big(1+\frac{\kappa-\|G_k(x,\theta|\boldsymbol{X}_{k-1})\|_2}{\|G_k(x,\theta|\boldsymbol{X}_{k-1})\|_2}\mathbf{1}_{\{\|G_k(x,\theta|\boldsymbol{X}_{k-1})\|_2\geq\kappa\}}\Big),\\
\bar{G}_k(x,\theta|\boldsymbol{X}_{k-1}) &= G_k(x,\theta|\boldsymbol{X}_{k-1})\frac{\min\big(s(\boldsymbol{X}_{k-1}),\kappa\big)}{s(\boldsymbol{X}_{k-1})},
\end{aligned}$$

for G_k an arbitrary gain function, $\kappa > 0$ and some function $s : \mathcal{X}^k \mapsto \mathbb{R}^+$. Note that $\tilde{G}_k$, $\mathring{G}_k$ and $\bar{G}_k$ all preserve the direction of G_k and have norm bounded by respectively 1, κ, and the norm of G_k, almost surely.

The gain $\bar{G}_k$ is specifically rescaled for situations where we have a conditional gain g_k almost surely of the form $g_k = -s(\boldsymbol{X}_{k-1})M_k(\theta-\vartheta)$, where M_k has eigenvalues as prescribed by (A1). Consequently we will have that $\bar{g}_k = -\min\big(s(\boldsymbol{X}_{k-1}),\kappa\big)M_k(\theta-\vartheta)$ from where it follows that the largest eigenvalue of the matrix $\min\big(s(\boldsymbol{X}_{k-1}),\kappa\big)M_k$ will then be almost surely upper-bounded; in certain situations it will be possible to use the fact that almost surely $E[\Lambda_{(1)}(M_k)|\boldsymbol{X}_{k-2}] \geq \lambda_1$, to show that $E[\min\big(s(\boldsymbol{X}_{k-1}),\kappa\big)\Lambda_{(1)}(M_k)|\boldsymbol{X}_{k-2}] \geq c\lambda_1$ for some $0 < c \leq 1$ and sufficiently large κ. Using the fact that the function $\min(x,\kappa)/x \leq 1$ we have, again abbreviating $\bar{G}_k(X_k,\theta|\boldsymbol{X}_{k-1})$ and $\bar{g}_k(\theta,\vartheta|\boldsymbol{X}_{k-1})$

$$\begin{aligned}\mathbb{E}\mathbb{E}_\vartheta\big[\|\bar{G}_k-\bar{g}_k\|_2^2|\boldsymbol{X}_{k-1}\big] &= \\ \mathbb{E}\Big[\Big(\frac{\min\big(s(\boldsymbol{X}_{k-1}),\kappa\big)}{s(\boldsymbol{X}_{k-1})}\Big)^2&\mathbb{E}_\vartheta\big[\|G_k-g_k\|_2^2|\boldsymbol{X}_{k-1}\big]\Big] \leq \mathbb{E}\|G_k-g_k\|_2^2,\end{aligned} \tag{26}$$

such that if G_k verifies (A2) then so will $\bar{G}_k$.

Another possible modification one might consider, is to truncate the iterates of the our algorithm (7). This might be motivated by practical considerations in the case where the parameter being tracked has some sort of physical meaning and is therefore bounded; it stands to reason then that the algorithm itself should be restricted as well. We would then, for a parameter set Θ, consider the sequence

$$\hat{\theta}_{k+1} = \Pi_{\bar{\Theta}}\big(\hat{\theta}_k + \gamma_k G_k(X_k,\hat{\theta}_k|\boldsymbol{X}_{k-1})\big), \quad k \in \mathbb{N}, \tag{27}$$

where $\Pi_{\bar{\Theta}}(\cdot)$ acts as a projection on a convex set $\bar{\Theta} \supset \Theta$ in that $\Pi_{\bar{\Theta}}(\cdot)$ is an identity on $\bar{\Theta}$ and maps points in $\bar{\Theta}^c$ to $\bar{\Theta}$.

We will provide concrete examples of gain functions later in Section 6. Before this, we present in Section 5 some examples of different types of variation that the parameter of the model may have such that our algorithm is capable of adequately tracking it.

5 Variational setups for the drifting parameter

It is clear – and in fact explicit in (8) and (9) – that the changes in the parameter have a non-negligible contribution to the accuracy of our tracking algorithm. This is reasonable since, if the parameter changes arbitrarily in-between observations, we should not expect it to the "trackable". We must then specify how the parameter is allowed to vary and, based on that assumption, pick an appropriate sequence γ_k which minimizes the general bounds in (8) or (9). We will specify in this section what these bounds reduce to for concrete examples for the variation of the parameter being tracked. These examples refer only to how the parameter is assumed to change and are unrelated to the actual model in question; examples of specific models can be found in Section 6.

5.1 Static parameter

We assume in this section that $\theta_j(\boldsymbol{X}_{j-1}) = \theta_0$, almost surely, $\forall j \in \mathbb{N}$ for some unknown $\theta_0 \in \Theta$ such that in fact $\Delta\theta_j = \mathbf{0}$, almost surely, and we are actually in a parametric setup. Note that, in this case, the second terms in both (8) and (9) obviously vanish.

Take then $\gamma_j = C_\gamma j^{-1}\log j$ and for $q \in (0,1)$, $n_0 = [qn]$, where $[a]$ is the whole part of $a \in \mathbb{R}$. Let $n \geq 2/q = N_q$ such that $n_0 \geq 2$. For large enough C_γ and all $n \geq N_q$ we have,

$$\sum_{j=n_0}^{n} \gamma_j \geq c_\gamma \log n_0 \sum_{j=n_0}^{n} \frac{1}{k} \geq \frac{\log n}{2\lambda_1},$$

from where for all $p \in \mathbb{N}$,

$$\exp\Big(- p\lambda_1 \sum_{j=n_0}^{n} \gamma_j\Big) \leq n^{-p/2}.$$

Note that in the case where we have $\mathbb{E}\|\delta_{n_0}\|_p^p \leq C_0 n_0^p$ we can take the constant C_γ to be larger (say take rC_γ, $r > 2$) in which case

$$C_1 \exp\Big(- p\lambda_1 \sum_{j=n_0}^{n} \gamma_j\Big) \leq c_1 n^p n^{-rp/2} \leq C_1 n^{-p/2}.$$

Using now the fact that $\sum_{j=n_0}^{n} \gamma_j^2 \leq c(\log n)^2 n^{-1}$ for some constant $c > 0$ we have

$$\left(\sum_{j=n_0}^{n} \gamma_j^2\right)^{p/2} \leq (n^{-1/2}\log n)^p.$$

We conclude that we can rewrite (8) and (9) as respectively,

$$\max_{n \geq N_q} \mathbb{E}\frac{\sqrt{n}}{\log n}\|\delta_n\|_p \leq C \quad \text{and} \quad \max_{n \geq N_q} \mathbb{E}\Big(\frac{\sqrt{n}}{\log n}\|\delta_n\|_p\Big)^p \leq C,$$

for all $p \in \mathbb{N}$. The log term in the rate cannot be avoided and is a consequence of the recursiveness of the algorithm.

Note that by taking $p > \epsilon^{-1}$ and, by using Markov's inequality and the second bound in the previous display, we conclude that

$$\begin{aligned} \sum_{n=1}^{\infty} P\big(n^{1/2-\epsilon}\|\hat{\theta}_n - \theta_0\|_1 > c\big) &\leq \sum_{n=1}^{\infty} P\big(d^{\frac{p-1}{p}} n^{1/2-\epsilon}\|\hat{\theta}_n - \theta_0\|_p > c\big) \\ &\leq \sum_{n=0}^{\infty} \frac{d^{p-1} n^{p/2-p\epsilon}\mathbb{E}\|\delta_n\|_p^p}{c^p} \leq C \sum_{n=1}^{\infty} \frac{(d\log n)^p}{n^{-1/2-\epsilon}} < \infty. \end{aligned} \tag{28}$$

By application of the Borel-Cantelli Lemma, we conclude that $\|\hat{\theta}_n - \theta_0\|_1 \to 0$ as $n \to 0$ takes place with probability 1 at a rate $n^{1/2-\epsilon}$ for all $\epsilon > 0$.

The particular setup presented in this section, where the parameter is fixed, might seem out of place since we are mainly concerned with tracking time-changing parameters. We would like to point out, however, that our algorithm is recursive and, as such, always produces estimates in a fast, straightforward fashion. This is an advantage especially over "offline" estimators obtained, say, as solutions to a certain system, which require iterative likelihood or least squares optimization or are obtained via other indirect methods, a situation which is common when dealing with Markov models (cf. Section 6.4.)

5.2 Stabilizing parameter

Suppose now that the parameter we want to track is stabilizing. This situation might arise if the expectation of the sequence of values that the parameter takes is converging to some limiting value. It could also be the case that the data is being sampled, with increasing frequency, from an underlying, continuous time process which depends on a parameter varying continuously; in this case, the parameter varies less and less since it is allowed less time to change. Regardless, we assume that $\Delta\theta_i = \theta_i(\boldsymbol{X}_{i-1}) - \theta_{i+1}(\boldsymbol{X}_i)$ verifies

$$\mathbb{E}\|\Delta\theta_i\|_p^p \leq \rho_i^p, i \in \mathbb{N}$$

for $p \geq 1$ and some decreasing sequence ρ_i. Assume then that we have $\rho_i = c_\rho i^{-\beta}$ for some constant $c_\rho > 0$ and $\beta \geq 0$.

Consider first the case $\beta \geq 3/2$. In this case, the variation of the parameter vanishes so quickly that we are essentially in the setup of the previous section. Indeed, take γ_i and n_0 as in the previous section. The first and third term in both (8) and (9) can be bounded in the same way as in the previous section. As for the second term, by using the Hölder inequality

$$\begin{aligned}\mathbb{E}\Big(\sum_{i=n_0}^{n} \|\Delta\theta_i\|_p\Big)^p &\leq (n-n_0)^{p-1}\sum_{i=n_0}^{n}\mathbb{E}\|\Delta\theta_i\|_p^p \leq C(n-n_0)^p\rho_{n_0}^p \\ &\leq c\big((n-n_0)n_0^{-\beta}\big)^p \leq Cn^{-(\beta-1)p} \leq Cn^{-p/2},\end{aligned} \tag{29}$$

leading to the same bounds as in the previous section

Consider now the case where $0 < \beta < 3/2$. Let $\gamma_i = C_\gamma(\log i)^{1/3} i^{-2\beta/3}$, $n_0 = n - n^{2\beta/3}(\log n)^{2/3}$. By using the elementary inequality $(1+x)^\alpha \leq 1 + \alpha x$ for $0 < \alpha < 1$ and $x \geq -1$, we obtain that for sufficiently large n (i.e., $n \geq N_1 = N_1(\beta)$) and sufficiently large constant C_γ

$$\begin{aligned}\sum_{i=n_0}^{n}\gamma_i &\geq C_\gamma(\log n_0)^{1/3}\sum_{i=n_0}^{n}\frac{1}{i^{2\beta/3}} \geq C_\gamma(\log n_0)^{1/3}\int_{n_0}^{n}\frac{dx}{x^{2\beta/3}} \\ &= \frac{C_\gamma(\log n_0)^{1/3}}{1-2\beta/3}\Big[n^{1-2\beta/3} - n^{1-2\beta/3}\big(1-n^{2\beta/3-1}(\log n)^{2/3}\big)^{1-2\beta/3}\Big] \\ &\geq \frac{C_\gamma(\log n_0)^{1/3}}{1-2\beta/3}\Big[n^{1-2\beta/3} - n^{1-2\beta/3}\big(1-n^{2\beta/3-1}(\log n)^{2/3}(1-2\beta/3)\big)\Big] \\ &= C_\gamma(\log n_0)^{1/3}(\log n)^{2/3} \geq \frac{\log n}{2h}.\end{aligned}$$

This yields the same bound for the first term in (8) and (9): for $n \geq N_1$, sufficiently large constant C_γ and all $p \in \mathbb{N}$,

$$C_1 \exp\Big(-ph\sum_{i=n_0}^{n}\gamma_i\Big) \leq C_1 n^{-p/2}.$$

Let us bound now the last term in (8) and (9): for $n \geq N_2 = N_2(\beta)$ and all $p \in \mathbb{N}$

$$\Big(\sum_{i=n_0}^{n} \gamma_i^2\Big)^{p/2} \leq C\big((\log n)^{2/3} n_0^{-4\beta/3}(n-n_0)\big)^{p/2} \leq c\big((\log n)^{2/3} n^{-\beta/3}\big)^p.$$

For sufficiently large n (i.e., $n \geq N_3 = N_3(\beta)$) the second term in (8) and (9) are bounded similarly to (29) by

$$\mathbb{E}\Big(\sum_{i=n_0}^{n} \|\Delta\theta_i\|_p\Big)^p \leq c\big((n-n_0)n_0^{-\beta}\big)^p \leq C\big((\log n)^{2/3} n^{-\beta/3}\big)^p.$$

Finally we obtain that for $0 < \beta < 3/2$ and sufficiently large constant C_γ in the algorithm step $\gamma_i = C_\gamma(\log i)^{1/3} i^{-2\beta/3}$, (8) and (9) can be rewritten as respectively

$$\max_{n \geq N_\beta} \mathbb{E}\frac{n^{\beta/3}}{(\log n)^{2/3}}\|\delta_n\|_2 \quad \text{and} \quad \max_{n \geq N_\beta} \mathbb{E}\bigg(\frac{n^{\beta/3}}{(\log n)^{2/3}}\|\delta_n\|_p\bigg)^p \leq c,$$

where $N_\beta = \max(N_1, N_2, N_3)$ is the burn-in period of the algorithm.

Remark 3. If we choose $\gamma_i = C_\gamma(\log i)^{\alpha_1} i^{-\alpha}$ and $n_0 = n - n^\alpha(\log n)^{\alpha_2}$, $0 < \alpha < 1$, $\alpha_1, \alpha_2 \geq 0$, $\alpha_1 + \alpha_2 \geq 1$ in case $0 < \beta < 3/2$, then we get the following bound of the convergence rate: for sufficiently large n and sufficiently large constant C_γ

$$\mathbb{E}\|\delta_n\|_p^p \leq C\Big(n^{-\min\{\beta-\alpha,\alpha/2\}}(\log n)^{\max\{\alpha_2,\alpha_1+\alpha_2/2\}}\Big)^p.$$

Thus , the choice $\alpha = 2\beta/3$, $\alpha_1 = 1/3$, $\alpha_2 = 2/3$ is optimal in the sense of the minimum of the right-hand side of the above inequality.

Remark 4. Much in the same way as for (28), we can establish that for any $\epsilon > 0$, $\lim_{n\to\infty} n^{\beta/3-\epsilon}\|\delta_n\|_1 = 0$ with probability 1.

Finally, consider the case $\beta = 0$, i.e., we assume the following weak requirement: $\mathbb{E}\|\Delta\theta_i\|_p^p \leq c$, $i \in \mathbb{N}$, for some uniform constant c. Take $n - n_0 = N$, $\gamma_i = \gamma$ for some $N \in \mathbb{N}$, $\gamma > 0$. Then Theorem 1 implies that

$$\max_{n \geq N} \mathbb{E}\|\delta_n\|_p^p \leq C_1 e^{-phN\gamma} + C_2 N^{p/2}\gamma^p + C_3 N^p c = D.$$

We thus have that the algorithm will track down the parameter in the proximity of size D, which we can try to minimize by choosing appropriate constants N and γ.

5.3 Lipschitz signal with asymptotics in the sampling frequency

We consider now a slightly different setup where we assume that the parameter is changing, on average, like a Lipschitz function. In this setup we let the time series (1) be sampled from a continuous time process X_t, $t \in [0, 1]$ which we observe with frequency n. This means that for each $n \in \mathbb{N}$ we have a different model, namely,

$$X_0^n \sim P_{\theta_0^n}, \qquad X_k^n|\boldsymbol{X}_{k-1}^n \sim P_{\theta_k^n}(\cdot|\boldsymbol{X}_{k-1}^n), \quad k \leq n \in \mathbb{N}, \tag{30}$$

where the parameter $\theta_k^n = \theta_k^n(\boldsymbol{X}_{k-1}^n)$ verifies, for some $p \in \mathbb{N}$, $\kappa_{d,p} < \infty$

$$\mathbb{E}\|\theta_k^n(\boldsymbol{X}_{k-1}^n) - \theta_{k_0}^n(\boldsymbol{X}_k^n)\|_p^p \leq \kappa_{d,p}^p \Big(\frac{k-k_0}{n}\Big)^{\beta p}.$$

We could have for example that $\theta_k^n(\boldsymbol{X}_{k-1}^n) = \vartheta(k/n)$, almost surely, where $\vartheta(\cdot) \in \mathcal{L}(L,\beta) = \{g(\cdot) : \|g(t_1) - g(t_2)\|_1 \leq L|t_1 - t_2|^\beta, t_1, t_2 \in [0,1]\}$ for some $0 < \beta \leq 1$ and $L > 0$, a space of vector valued Lipschitz functions.

Let $\gamma_k \equiv C_\gamma (\log n)^{(2\beta-1)/(2\beta+1)} n^{-2\beta/(2\beta+1)}$, (constant in k) for $k = 1, \dots, n$, and

$$k_0 = k_0(n) = k - (\log n)^{2/(2\beta+1)} n^{2\beta/(2\beta+1)},$$

for $k \geq K_n = (\log n)^{2/(2\beta+1)} n^{2\beta/(2\beta+1)}$. Note that for $K_n/n \to 0$ as $n \to \infty$ for any $0 < \beta \leq 1$.

For sufficiently large C_γ

$$\sum_{i=k_0}^k \gamma_i = C_\gamma (\log n)^{(2\beta-1)/(2\beta+1)} n^{2\beta/(2\beta+1)} (k - k_0) \geq C_\gamma \log n \geq \frac{\log n}{3\lambda_1},$$

leading to

$$\exp\bigg(- p\lambda_1 \sum_{i=k_0}^k \gamma_i \bigg) \leq c n^{-p/3}.$$

In much the same way, we have

$$\bigg(\sum_{i=k_0}^k \gamma_i^2 \bigg)^{p/2} \leq C \bigg((\log n)^{\frac{2\beta-1}{2\beta+1}} n^{-\frac{2\beta}{2\beta+1}} (k-k_0)^{1/2} \bigg)^p = C \bigg((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}} \bigg)^p.$$

¿From our assumption on the variation of the parameter, we have

$$\max_{i=k_0,\dots,k} \mathbb{E}\|\theta_{i+1}^n - \theta_{k_0}^n\|_p^p \leq c\Big(\frac{k-k_0}{n}\Big)^{-p\beta} \leq C\bigg((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}} \bigg)^p.$$

Combining the three bounds, we get that (8) and (9) imply

$$\sup_{\vartheta \in \mathcal{L}(L,\beta)} \max_{i \geq K_n} \mathbb{E}\|\delta_i\|_2 \leq C (\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}}, \tag{31}$$

$$\sup_{\vartheta \in \mathcal{L}(L,\beta)} \max_{i \geq K_n} \mathbb{E}\|\delta_i\|_p^p \leq C \bigg((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}} \bigg)^p. \tag{32}$$

6 Some applications of the main result

In this section we present some examples of particular models to which our algorithm may be applied. We start with two toy examples and present thereafter some more involved examples. The toy examples illustrate the type of results that can be obtained from our main result and its extensions, how a gain function can be picked and modified, and how conditions (A1) and (A2) checked.

6.1 Tracking the intensity function of a Poisson process

Lets say that we are monitoring n independent Poisson processes on $[0,1]$ with unknown intensity function $\lambda(\cdot)$, for fixed $n \in \mathbb{N}$. This is equivalent to observing $N(t) = N(t,n)$, a Poisson process with intensity $n\lambda(t)$, $0 \le t \le 1$. We would like to track the intensity function $\lambda(\cdot)$ which we will assume is uniformly upper-bounded by L.

Lets say that we observe the process with frequency n, in that our observations are $X_k^n = N(k/n)$, such that for each $n \in \mathbb{N}$ we have the model

$$X_0^n = 0, \quad X_{k+1}^n | X_k^n \sim P_{\theta_k^n}(\cdot | X_k^n) = P_{\theta_k^n}(\cdot - X_k^n), \quad k = 1, \dots, n,$$

where $P_\theta(\cdot)$ represents a Poisson law with parameter $\theta \in \mathbb{R}^+$. We will work then with $p_\theta(\cdot|y)$ a conditional, shifted Poisson mass function given by

$$p_\theta(x|y) = \exp(-\theta)\frac{\theta^{x-y}}{(x-y)!},$$

for $x \in \mathbb{N}$, $x \ge y$. The moving parameter θ_k^n is given, for $k = 1, \dots, n$, by

$$\theta_k^n = \int_{\frac{k-1}{n}}^{\frac{k}{n}} n\lambda(t)\, dt.$$

Consider now the gain function G_k of the type (25) and its conditional expectation g_k, respectively given by

$$\begin{aligned} G_k(x,\theta|X_{k-1}^n) &= x - X_{k-1}^n - \theta, \\ g_k(\theta,\vartheta|X_{k-1}^n) &= \mathbb{E}_\vartheta[X_k^n - X_{k-1}^n - \theta | X_{k-1}^n] = -(\theta - \vartheta). \end{aligned} \tag{33}$$

with $\mathbb{E}_\vartheta[\,\cdot\,|X_{k-1}^n]$ the expectation with respect to $p_\vartheta(\cdot|X_{k-1}^n)$. Its also simple to see that,

$$\mathbb{E}|G(X_k^n,\theta|X_{k-1}^n) - g(\theta,\vartheta|X_{k-1}^n)|^2 = \mathbb{E}\mathbb{E}_\vartheta[|X_k^n - X_{k-1}^n - \vartheta|^2 | X_{k-1}^n] = \vartheta \le L.$$

We conclude then that the gain function displayed in (33) satisfies both (A1) and (A2).

This gain function can now be used for the three setups outlined in Section 5 and attains the rates indicated there. For a constant intensity function $\lambda(\cdot) \equiv \vartheta$, $0 < \vartheta \le L$, the parameter of the model θ_k^n reduces to the constant ϑ and we simply track the rate of the process. Note that this happens since we matched the sampling frequency $1/n$ with the sample size n. If we were to have sampled the process with frequency $2/n$, say, then $\theta_k^n = 2\vartheta$ in which case the algorithm would track 2ϑ and not ϑ. The tracking sequence would then have to be recalled by a factor $1/2$ to obtain a tracking sequence for ϑ itself.

In the setup where we assume that the parameter is stabilizing, take $n = 1$ and call $\vartheta_k = \theta_n^1 = \int_{k-1}^k \lambda(t)\, dt$ the mean number of events per time unit. Note that

$$|\Delta\vartheta_k| = \Big|\int_{k-1}^k \lambda(t)\, dt - \int_k^{k+1} \lambda(t)\, dt\Big| = |\theta_k^1 - \theta_{k+1}^1|.$$

We then assume that the average number of events is stabilizing in such a way that the previous display is upper bounded by say $c_\beta k^{-\beta}$, for $\beta \geq 0$ and $c_\beta > 0$. The algorithm will then track the mean number of events per time unit.

We can also assume that the intensity function $\lambda(\cdot)$ belongs to $\mathcal{L}(L,\beta) = \{g(\cdot) : |g(t_1) - g(t_2)| \leq L|t_1 - t_2|^\beta, t_1, t_2 \geq 0\}$ for some $0 < \beta \leq 1$ and $L > 0$. Call $\vartheta_k^n = \lambda(k/n)$, $k, n \in \mathbb{N}$. It follows that

$$\begin{aligned} \big|\Delta\vartheta_k^n\big| &= \big|\lambda\big(k/n\big) - \lambda\big((k+1)/n\big)\big| \leq L\, n^{-\beta}, \\ \big|\theta_k^n - \vartheta_k^n\big| &= \Big|\int_{(k-1)/n}^{k/n} n\lambda(t)\, dt - \lambda(k/n)\Big| \leq n\int_{(k-1)/n}^{k/n} \big|\lambda(t) - \lambda(k/n)\big|\, dt \leq L\, n^{-\beta}. \end{aligned}$$

The tracking sequence based on the gain (33) will then track the sequence $\vartheta_k^n = \lambda(k/n)$, $k, n \in \mathbb{N}$ (as well as θ_k^n) with the asymptotics seen in Section 5 (cf. Remark 1.)

6.2 Tracking the mean function of a conditionally Gaussian process

Assume that we observe, with fixed frequency $n \in \mathbb{N}$, a process X_t, $t \in [0,1]$, taking values on $\mathcal{X} \subset \mathbb{R}^d$, $d \in \mathbb{N}$. In this way, for $k = 1, \ldots, n$, the observations available to us at time k/n will be a random vector $\boldsymbol{X}_k^{(n)} = \big(X_0, X_{1/n}, \ldots, X_{k/n}\big)$. The increments $X_{k/n} - X_{(k-1)/n}$ will be assumed to be conditionally Gaussian in the sense that given the past of the process, each increment has a multivariate normal distribution, and so,

$$X_0^n \sim N\Big(\theta_0^n,\, \Sigma_0^n\Big), \quad X_{(k+1)/n}|\boldsymbol{X}_k^n \sim N\Big(\theta_k^n\big(\boldsymbol{X}_{k-1}^n\big),\, \Sigma_k^n\big(\boldsymbol{X}_{k-1}^n\big)\Big), \quad k = 1, \ldots, n.$$

The dependence on the past in the model comes from the fact that both the mean and the covariance of the process are allowed to depend on the past of the process. Here, for each $n \in \mathbb{N}$, θ_k^n is an arbitrary sequence in k depending eventually on $\boldsymbol{X}_{k-1}^n$ and Σ_k^n a sequences in $k \in \mathbb{N}$ of (positive-definite) covariance matrices or order d which, as already mentioned, may also depend on $\boldsymbol{X}_{k-1}^n$.

In the case where the covariance structure of the process is known, we can use the gain (24) which it is straightforward to check verifies, given $\boldsymbol{X}_{k-1}^n$, for $\boldsymbol{x}, \theta, \vartheta \in \mathbb{R}^d$, $k = 1, \ldots, n$,

$$\begin{aligned} G_k\big(\boldsymbol{x}, \theta|\boldsymbol{X}_{k-1}^n\big) &= \big(\Sigma_k^n(\boldsymbol{X}_{k-1}^n)\big)^{-1}\big(\boldsymbol{x} - \theta\big), \\ g_k\big(\theta, \vartheta(\boldsymbol{X}_{k-1}^n)|\boldsymbol{X}_{k-1}^n\big) &= -\big(\Sigma_k^n(\boldsymbol{X}_{k-1}^n)\big)^{-1}\big(\theta - \vartheta(\boldsymbol{X}_{k-1}^n)\big). \end{aligned} \tag{34}$$

If this gain is used, we assume that, almost surely, for $k = 1, \ldots, n$, the eigenvalues of the covariance matrices $\Sigma_k^n(\boldsymbol{X}_{k-1}^n)$ are $0 < \Lambda_{(1),k}^n(\boldsymbol{X}_{k-1}^n) \leq \cdots < \Lambda_{(d),k}^n(\boldsymbol{X}_{k-1}^n) < \infty$, such that for constants λ_1^n, λ_2^n we have,

$$0 < \lambda_1^n \leq \Lambda_{(1),k}^n(\boldsymbol{X}_{k-1}^n) \leq \Lambda_{(d),k}^n(\boldsymbol{X}_{k-1}^n) \leq \lambda_2^n < \infty,$$

almost surely. We then have for all $\theta, \vartheta \in \mathbb{R}^d$,

$$\begin{aligned} &\mathbb{E}\|G(X_k^n, \theta|\boldsymbol{X}_{k-1}^n) - g(\theta, \vartheta(\boldsymbol{X}_{k-1}^n)|\boldsymbol{X}_{k-1}^n)\|_2^2 = \\ &\quad = \mathbb{E}\big((\Sigma_k^n(\boldsymbol{X}_{k-1}^n))^{-1}\big(X_k^n - \vartheta(\boldsymbol{X}_{k-1}^n)\big)\big)^T(\Sigma_k^n(\boldsymbol{X}_{k-1}^n))^{-1}\big(X_k^n - \vartheta(\boldsymbol{X}_{k-1}^n)\big) \\ &\quad \leq (\lambda_1^n)^{-2}\, \mathbb{E}\mathbb{E}_\vartheta\big[\|X_k^n - \vartheta(\boldsymbol{X}_{k-1}^n)\|_2^2\big|\boldsymbol{X}_k^n\big] = (\lambda_1^n)^{-2}\mathbb{E}\,\mathrm{tr}\,\big(\Sigma_k^n(\boldsymbol{X}_{k-1}^n)\big) \leq d\,\lambda_2^n\,(\lambda_1^n)^{-2}. \end{aligned}$$

Assumptions (A1) and (A2) are then met for the gain in (34).

Let us now assume that the covariance matrix of the process is unknown, difficult to invert or that assumption on the eigenvalues of the covariance matrix does not hold. In this case we can use the gain (25) which gives us, for $\boldsymbol{x}, \theta, \vartheta \in \mathbb{R}^d$, $k = 1, \ldots, n$,

$$\begin{aligned} G(\boldsymbol{x}, \theta | \boldsymbol{X}_{k-1}^n) &= \boldsymbol{x} - \theta, \\ g(\theta, \vartheta(\boldsymbol{X}_{k-1}^n) | \boldsymbol{X}_{k-1}^n) &= -\big(\theta - \vartheta(\boldsymbol{X}_{k-1}^n)\big). \end{aligned} \tag{35}$$

If we now assume that, almost surely, for $k = 1, \ldots, n$, the largest eigenvalue of the covariance matrices $\Sigma_k^n(\boldsymbol{X}_{k-1}^n)$ is upper bounded by some constant $\lambda_2^n < \infty$, then, for all $\theta, \vartheta \in \mathbb{R}^d$,

$$\begin{aligned} &\mathbb{E}_\vartheta \| G(X_k^n, \theta | \boldsymbol{X}_{k-1}^n) - g(\theta, \vartheta(\boldsymbol{X}_{k-1}^n) | \boldsymbol{X}_{k-1}^n) \|_2^2 = \\ &\quad = \mathbb{E}\mathbb{E}_\vartheta \big[\| X_k^n - \vartheta(\boldsymbol{X}_{k-1}^n) \|_2^2 \big| \boldsymbol{X}_{k-1}^n \big] = \mathbb{E} \operatorname{tr} \big(\Sigma_k^n(\boldsymbol{X}_{k-1}^n) \big) \leq d \, \lambda_2^n, \end{aligned}$$

and so assumptions (A1) and (A2) are met for the gain in (35).

The results of Section 5 can be applied to the algorithm based on the gain functions presented above. If, for each $n \in \mathbb{N}$, the mean of the process is constant, $\theta_k^n(\boldsymbol{X}_{k-1}^n) \equiv \vartheta^n$ then the algorithm will track the (fixed) mean of the process. Alternatively we may assume that the parameter isn't constant but it stabilizing. We take then $n = 1$, and assume that the changes in the mean vector of the process are such that, for $k \in \mathbb{N}$,

$$\mathbb{E}\|\Delta\theta_k^n\|_2^2 = \mathbb{E}\|\theta_k^n(\boldsymbol{X}_{k-1}^n) - \theta_{k+1}^n(\boldsymbol{X}_k^n)\|_2^2 \leq c_\beta k^{-\beta},$$

for some $\beta \geq 0$, and a constant $c_\beta > 0$. The other possibility is to assume that for $n \in \mathbb{N}$, the mean of the process is obtained from a function $\theta(\cdot, \boldsymbol{X}_{k-1}^n)$ which is, on average, Lipschitz in the sense that it belongs to $\mathcal{L}\big(L, \beta, \boldsymbol{X}_{k-1}^n\big) = \{g : \mathbb{E}\|g(t_1, \boldsymbol{X}_{k-1}^n) - g(t_2, \boldsymbol{X}_{k-1}^n)\|_1 \leq L|t_1 - t_2|^\beta, t_1, t_2 \geq 0\}$ for some $0 < \beta \leq 1$ and $L > 0$. Call $\vartheta_k^n = \theta(k/n, \boldsymbol{X}_{k-1}^n)$, $k, n \in \mathbb{N}$. It follows that

$$\mathbb{E}\big\|\Delta\vartheta_k^n\big\|_1 = \mathbb{E}\big\|\theta\big(k/n, \boldsymbol{X}_{k-1}^n\big) - \theta\big((k+1)/n, \boldsymbol{X}_k^n\big)\big\|_1 \leq L n^{-\beta}.$$

In this case the algorithm tracks the mean function $\theta(k/n, \boldsymbol{X}_{k-1}^n)$ at times k/n, with $k \in \mathbb{N}$.

6.3 Tracking an ARCH(1) parameter

Consider the following ARCH(1) model with drifting parameter

$$X_k = \big(1 + \theta_k X_{k-1}^2\big)^{1/2} \xi_k, \quad X_0 = 0 \text{ (a.s.)}, \tag{36}$$

where ξ_k, $k \in \mathbb{N}$, form a martingale difference sequence with variance $\sigma^2 > 0$. The drifting parameter θ_k belongs to some interval $[0, \rho]$ for some ρ such that $\rho^2 \mathbb{E}\epsilon_k^4 \leq 1$ for all $k \in \mathbb{N}$.

Consider the gain function

$$G(X_k, \theta | X_{k-1}) = \frac{\min(X_{k-1}^2, c\sigma^2)}{X_{k-1}^2} (X_k^2 - 1 - \theta X_{k-1}^2) \tag{37}$$

such that, since $\mathbb{E}_\vartheta X_k = 0$ and $\mathbb{E}_\vartheta[X_k^2|X_{k-1}] = 1 + \vartheta X_{k-1}^2$,

$$g(\theta, \vartheta|X_{k-1}) = \mathbb{E}_\vartheta\Big[\frac{\min(X_{k-1}^2, c\sigma^2)}{X_{k-1}^2}(X_k^2 - 1 - \theta X_{k-1}^2)\Big|X_{k-1}\Big] = -\min(X_{k-1}^2, c\sigma^2)(\theta - \vartheta),$$

For some constant $c > 0$. We then have that $\Lambda_{(1)} \le c\sigma^2$, almost surely. Note that

$$\mathbb{E}\Big[\min(X_{k-1}^2, c\sigma^2)\Big|X_{k-2}\Big] = \mathbb{E}\Big[\min\big((1 + \theta_{k-1}X_{k-2}^2)\epsilon_{k-1}^2, c\sigma^2\big)\Big|X_{k-2}\Big] \ge \mathbb{E}\Big[\min\big(\epsilon_{k-1}^2, c\sigma^2\big)\Big].$$

By using the fact that $\min(a, b) = (a + b)/2 - |a - b|$ and the Hölder inequality, it is straightforward to check that

$$2\,\mathbb{E}\Big[\min\big(\epsilon_{k-1}^2, c\sigma^2\big)\Big] = (c+1)\sigma^2 - \mathbb{E}|\epsilon_{k-1}^2 - c\sigma^2| \ge (c+1)\sigma^2 - \big(\mathbb{E}\big[(\epsilon_{k-1}^2 - c\sigma^2)^2\big]\big)^{1/2} \ge \sigma^2,$$

as long as for every $k \in \mathbb{N}$, $2c\sigma^2 \ge \mathbb{E}\epsilon_k^4$. We conclude (A1) holds for the gain (37).

To check (A2) note first that

$$\mathbb{E}[X_k^2|X_{k-1}^2] = \sigma^2(1 + \theta_k X_{k-1}^2)$$

and then

$$\mathbb{E}X_k^2 \le \sigma^2(1 + \rho\mathbb{E}X_{k-1}^2).$$

Since $\rho^2\mathbb{E}\epsilon_k^4 \le 1$ then $\rho\sigma^2 \le 1$ by Jensen's inequality. Using this recursion we get that

$$\mathbb{E}X_k^2 \le \sigma^2 + \sigma^2\rho\mathbb{E}X_{k-1}^2 \le \sigma^2 + \sigma^4\rho + \sigma^4\rho^2\mathbb{E}X_{k-2}^2 \le \sigma^2\sum_{i=1}^k(\rho\sigma^2)^{i-1} \le \frac{\sigma^2}{1 - \sigma^2\rho}.$$

In the same way,

$$\mathbb{E}[X_k^4|X_{k-1}] = (1 + 2\theta_k X_{k-1}^2 + \theta_k^2 X_{k-1}^4)\mathbb{E}\epsilon_k^4,$$

and then, since $\rho^2\mathbb{E}\epsilon_k^4 \le 1$,

$$\mathbb{E}X_k^4 \le (1 + 2\frac{\sigma^2\rho}{1 - \sigma^2\rho} + \rho^2\mathbb{E}X_{k-1}^4)\mathbb{E}\epsilon_k^4 \le \mathbb{E}\epsilon_k^4(1 + 2\frac{\sigma^2\rho}{1 - \sigma^2\rho})\sum_{i=1}^k(\rho^2\mathbb{E}X_{k-1}^4)^{i-1} < \infty.$$

Using the same argument as for (26) we see that (A2) holds since by the Hölder inequality

$$\mathbb{E}G^2(X_k, \theta|X_{k-1}) \le 3(\mathbb{E}X_k^2 + \rho^2\mathbb{E}X_{k-1}^2 + 1),$$

which is bounded, uniformly over $k \in \mathbb{N}$.

6.4 Tracking an AR(d) parameter

We consider now an autoregressive model with d time-varying auto-regressive parameters:

$$X_k = \sum_{i=1}^{d} \theta_{k,i} X_{k-i} + \xi_k, \quad k \in \mathbb{N},\ k \geq d, \tag{38}$$

where $X_0, X_1, \ldots, X_{d-1}$ have p bounded moments (cf. the end of this section). We would like to track the vector $\theta_k = (\theta_{k,1}, \theta_{k,2}, \ldots, \theta_{k,d})$, which may be random but must be measurable with respect to the σ algebra generated by $\boldsymbol{X}_{k-2d-1}$. In this section we will use the notation $\boldsymbol{X}_{k,d} = \big(X_k, X_{k-1}, \ldots, X_{k-(d-1)}\big)$ for the vector of the d observations leading up to X_k.

In analogy with the non-drifting AR(d) model, we can associate with the model its (drifting) autoregressive polynomial $z \mapsto 1 - \sum_{i=1}^{d} \theta_{k,i} z^i$; write then

$$t(z, \theta) = 1 - \sum_{i=1}^{d} \theta_i z^i, \quad z \in \mathbb{C}. \tag{39}$$

It is well know that an AR(p) model with autoregressive parameters θ is stationary if, and only if, the (complex) zeros of the polynomial $t(z, \theta)$ are outside the unit circle. This motivates the definition of the parameter sets $\Theta(\rho)$, (cf. Moulines et al. [2005]) which we define as the closure of [is it already closed???]

$$\big\{\theta \in \mathbb{R}^d : \text{for all } |z| < \rho^{-1}, t(z, \theta) \neq 0\big\}, \tag{40}$$

for any $0 < \rho < 1$. One can show that if $\mathcal{B}(r)$ is a uniform ball in $\mathbb{R}^d$ with radius $r > 0$, then the following embeddings hold:

$$\mathcal{B}\big((\rho^{-2} + \cdots + \rho^{-2d})^{-1/2}\big) \subseteq \Theta(\rho) \subseteq \mathcal{B}\big((1+\rho)^d - 1\big),$$

which gives us some feeling as to the size of the parameter set. This implies in particular that for all $\rho \in (0, 1)$, the set $\Theta(\rho)$ is non-empty and bounded.

The AR(d) model (38) can also be described by the following inhomogeneous difference equation

$$\boldsymbol{X}_{k,d} = C(\theta_k) \boldsymbol{X}_{k-1,d} + \boldsymbol{I} \boldsymbol{e}_1 \xi_k, \tag{41}$$

where $\boldsymbol{e}_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$ and, for any $\theta \in \mathbb{R}^d$, $C(\theta)$ is the square matrix of order d

$$C(\theta) = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_{d-1} & \theta_d \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{42}$$

This matrix is usually called the *companion matrix* to the autoregressive polynomial $t(z, \theta)$; it is also sometimes called the *state transition matrix*. One can show that the eigenvalues of

$C(\theta)$ are exactly the reciprocals of the zeros of $t(z,\theta)$. This means that all the eigenvalues of $C(\theta)$ for $\theta \in \Theta(\rho)$ are at most $\rho < 1$. This in turn implies that for any sequence of vectors $\theta_d, \theta_{d+1}, \cdots \in \Theta(\rho)$, the pair of sequences

$$\Big(\big(C(\theta_d), C(\theta_{d+1}), \cdots\big), \big(I_d, I_d, \cdots\big)\Big)$$

forms a so called *exponentially stable* pair. Among other things, this gives us that so long as the p-th moments of both the initial $\boldsymbol{X}_{d-1,d}$ and the noise terms ξ_k are bounded, then the p-th moments of all X_k, $k \geq d$ will be bounded as well (cf. Proposition 10 of Moulines et al. [2005]).

A particular gain function which can be used to track the parameters of an autoregressive model can be found in Moulines et al. [2005]. The gain function considered there is an appropriately rescaled version of the gain (19), namely,

$$G(X_k, \theta|\boldsymbol{X}_{k-1,d}) = (X_k - \theta^T \boldsymbol{X}_{k-1,d}) \frac{\boldsymbol{X}_{k-1,d}}{1 + \mu \boldsymbol{X}_{k-1,d}^T \boldsymbol{X}_{k-1,d}},$$

for an appropriately chosen $\mu > 0$. Its straightforward to check that the conditions in ($\tilde{\text{A}}$1) on the corresponding conditional gain g hold in this case; the lower bound in (6) is established in Lemma 17 of Moulines et al. [2005] and the upper bounds are straightforward to check; assumption (A2) can be reduced to moment conditions on the observations of the autoregressive process which are verified if the signal θ lives in $\Theta(\rho)$ as mentioned above. In a sense, conditions (A1) and (A2) capture the essential properties that a gain function must have such that resulting tracking algorithm behaves properly and, in fact, these conditions will hold even if the noise terms are not Gaussian; we discuss this issue again at the end of this section. In the following we propose an alternative gain function. We will first treat the one dimensional case where we can obtain a stronger result.

Consider then $d = 1$ and assume that the sequence $\theta_k(\boldsymbol{X}_{k-1}) \in \Theta(\rho)$ such that it is almost surely bounded, in absolute value, by $\rho < 1$. Assume also that $\mathbb{E}X_0^2$ and $\mathbb{E}X_0^4$ are bounded and that for all $k \in \mathbb{N}$, $\mathbb{E}\xi_k = \mathbb{E}\epsilon_k^3 = 0$, $\mathbb{E}\xi_k^2 = \sigma^2 > 0$ and $\mathbb{E}\xi_k = c\,\sigma^4$, for some constant $0 \leq c < 5$. (We can take $c = 3$ if the noise is Gaussian, for example.) Lets say we would like to use the following gradient type gain function

$$\begin{aligned} G_k\big(X_k, \theta|X_{k-1}\big) &= X_{k-1}^2 (X_k/X_{k-1} - \theta), \\ g_k\big(\theta, \vartheta|X_{k-1}\big) &= -X_{k-1}^2(\theta - \vartheta), \end{aligned} \tag{43}$$

almost surely. The random eigenvalues in (A1) reduce, in this case, to just $\Lambda_{(1)}(X_{k-1}) = X_{k-1}^2$. Note that if for all $k \in \mathbb{N}$ X_{k-2} were integrable, we have

$$\mathbb{E}\big[X_{k-1}^2\big|X_{k-2}\big] = \mathbb{E}\big[(X_{k-2}\theta_{k-2} + \xi_{k-1})^2\big|X_{k-2}\big] = X_{k-2}^2\theta_{k-2}^2 + \mathbb{E}\xi_{k-1}^2 \geq \sigma^2, \tag{44}$$

but still X_{k-1}^2 would not be almost surely upper-bounded by a constant. To remedy this we will truncate X_{k-1}^2 and consider

$$\begin{aligned} G_k\big(X_k, \theta|X_{k-1}\big) &= \min\Big(X_{k-1}^2, \frac{9-c}{4}\sigma^2\Big)(X_k/X_{k-1} - \theta), \\ g_k\big(\theta, \vartheta|X_{k-1}\big) &= -\min\Big(X_{k-1}^2, \frac{9-c}{4}\sigma^2\Big)(\theta - \vartheta). \end{aligned} \tag{45}$$

(Note that this is a rescaled gain function of the same type as $\bar{G}$ at the end of Section 4.) We now have an almost sure upper-bound for $\Lambda_{(1)}(X_{k-1}) = \min\big(X_{k-1}^2, (9-c)\sigma^2/4\big)$; we truncate X_{k-1}^2 at this specific value since one can prove (cf. Lemma 5) that

$$\mathbb{E}\Big[\min\Big(X_{k-1}^2, \frac{9-c}{4}\sigma^2\Big)\Big|X_{k-2}\Big] \geq \frac{5-c}{4}\sigma^2 > 0,$$

such that (A1) holds. Assumption (A2) also holds since

$$\begin{aligned}\mathbb{E}\big|G_k\big(X_k, \theta|X_{k-1}\big) - g_k\big(\theta, \vartheta|X_{k-1}\big)\big|^2 &= \mathbb{E}\frac{\min\big(X_{k-1}^2, \frac{5-c}{4}\sigma^2\big)^2}{X_{k-1}^2}\mathbb{E}_\vartheta\big[|X_k - \vartheta X_{k-1}|^2\big|X_{k-1}\big]\\ &\leq \Big(\frac{5-c}{4}\Big)^2\sigma^4\mathbb{E}\xi_k^2 = \Big(\frac{5-c}{4}\Big)^2\sigma^6.\end{aligned}$$

The previous truncation argument is still valid if we truncate X_{k-1}^2 at a higher value. In that case, we also still have that (A1) and (A2) hold, with a larger constant λ_2 in (A1) and larger C in (A2). This means that in order to use the previous gain function we don't need to know the exact value of σ^2 but only an upper bound for it. Also, in practice, for a truncation at a high enough value, the effect of the truncation will be innocuous and trajectories of (43) and (45) will coincide, with high probability; the truncation is simply an artifice to enforce the fulfillment of (A1) and should be of little practical importance. Up to the requirement that the distribution of the noise be symmetrical about 0, the previous result generalizes that of Belitser [2000] where the noise terms are assumed to be almost surely bounded.

We turn our attention now to the general AR(d) model. As we will see in what follows, assumptions (A1) and (A2) can be easily checked. In the d dimensional case we assume that the noise terms ξ_k in (38) form a Gaussian white noise sequence with mean zero and variance $\sigma^2 > 0$. Assume first that the autoregressive parameters do not depend on k, i.e. $\theta_k \equiv \theta = (\theta_1, \ldots, \theta_d) \in \Theta(\rho) \subset \mathbb{R}^d$. Given the vector of past observations $\boldsymbol{X}_{k-d,d} = \big(X_{k-d}, X_{k-d-1}, \ldots, X_{k-2d+1}\big)$, we can see $\boldsymbol{X}_{k,d}$ as a system of d equations in $X_k, X_{k-1}, \ldots, X_{k-(d-1)}$, depending on $X_{k-d}, X_{k-d-1}, \ldots, X_{k-2d+1}$ and θ, which, for $\boldsymbol{\xi}_{k,d} = \big(\xi_k, \xi_{k-1}, \ldots, \xi_{k-(d-1)}\big)$, can be written as

$$A(\theta)\boldsymbol{X}_{k,d} = B(\theta)\boldsymbol{X}_{k-d,d} + \boldsymbol{\xi}_{k,d}; \tag{46}$$

the matrices $A(\theta)$ and $B(\theta)$ are Toeplitz matrices created respectively from the vectors $\boldsymbol{a}(\theta) = (0, \ldots, 0, 1, -\theta_1, \ldots, -\theta_{d-1})$ and $\boldsymbol{b}(\theta) = (\theta_1, \ldots, \theta_{d-1}, \theta_d, 0, \ldots, 0)$. (We remind that for a vector $\boldsymbol{m} = (m_{-(d-1)}, m_{-(d-2)}, \ldots, m_{-1}, m_0, m_1, \ldots, m_{d-2}, m_{d-1})$, the Toeplitz matrix of order d associated with that vector is the square matrix M of order d with entries $m_{i,j} = m_{i-j}$, such that the entries of the matrix are constant over descending diagonals.) The matrix $A(\theta)$ is upper triangular with a diagonal consisting of ones whence invertible. We conclude, then, that given the full past of the process, $\boldsymbol{X}_{k-d}$,

$$\boldsymbol{X}_{k,d}|\boldsymbol{X}_{k-d} \sim N\big(A^{-1}(\theta)B(\theta)\boldsymbol{X}_{k-d,d},\ \sigma^2 A^{-1}(\theta)A^{-T}(\theta)\big). \tag{47}$$

Alternatively, we could have derived the (46) by applying the recursion in (41) d times.

In this case we will consider a gain of the type (24) such that

$$G_{dk}\big(\boldsymbol{x}, \theta | \boldsymbol{X}_{d(k-1),d}\big) = \nabla_\theta \log p_\theta\big(\boldsymbol{x} | \boldsymbol{X}_{d(k-1)}\big), \tag{48}$$

where $p_\theta(\cdot | \boldsymbol{X}_{d(k-1)})$ is the conditional density of (47). At the end of this section we explicitly compute (48); see (50).

For us, each data point will be a vector $\boldsymbol{X}_{dk,d}$, $k \in \mathbb{N}$ such that the tracking sequence is updated with batches of d observations from the autoregressive process. (Below, to ease the notation, we will mostly write $\boldsymbol{x}$ and $\boldsymbol{y}$ instead of $\boldsymbol{X}^n_{dk,d}$ and $\boldsymbol{X}^n_{d(k-1),d}$, respectively.) This is necessary to make sure the representation (47) is valid even if the parameter θ is allowed to change among different batches of observations; otherwise the system (46) would be under-determined. We must now establish that this gain function verifies (A1).

As explained in Section 4, the expectation g_{dk} can be seen as minus the gradient of the Kullback-Leibler divergence between the transition kernel with two different parameters. This observation is particularly useful if we are able to write this Kullback-Leibler divergence as an appropriate quadratic form. The Kullback-Leibler divergence between two d-dimensional multivariate normal distributions $P_0 = N(\boldsymbol{\mu}_0, \Sigma_0)$ and $P_1 = N(\boldsymbol{\mu}_1, \Sigma_1)$ is given (cf. **?**) by,

$$K(P_0, P_1) = \frac{1}{2}\left(\log \frac{\det \Sigma_1}{\det \Sigma_0} + \operatorname{tr}(\Sigma_1^{-1}\Sigma_0) - d + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \Sigma_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right). \tag{49}$$

Write, for $\boldsymbol{y} \in \mathbb{R}^d$, $\mu(\theta, \boldsymbol{y}) = A^{-1}(\theta)B(\theta)\boldsymbol{y}$ and $\Sigma(\theta) = \sigma^2 A^{-1}(\theta)A^{-T}(\theta)$. Let also $S = S_d$ be the Toeplitz matrix associated with the vector $\boldsymbol{s} = (0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^{2d-1}$ where the 1 occupies the $(d+1)$-th position; these are sometimes called *shift matrices.* For $i = 2, \dots, d-1$, the powers S^i are the Toeplitz matrices associated with the vectors $(0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^{2d-1}$ where the 1 occupies the $(d+i)$-th position; S^d is $O = O_d$, the null matrix of order d, and S^0 should be read as $I = I_d$, the identity matrix of order d. It follows from the definition of the matrix $A(\cdot)$ that for $\theta, \vartheta \in \Theta_d$,

$$A(\theta) - A(\vartheta) = S(\vartheta_1 - \theta_1) + S^2(\vartheta_2 - \theta_2) + \cdots + S^d(\vartheta_d - \theta_d),$$

from where we conclude

$$A(\theta)A^{-1}(\vartheta) = I + SA^{-1}(\vartheta)(\vartheta_1 - \theta_1) + S^2A^{-1}(\vartheta)(\vartheta_2 - \theta_2) + \cdots + S^dA^{-1}(\vartheta)(\vartheta_d - \theta_d).$$

We will compute now $K\big(N(\mu(\vartheta, \boldsymbol{y}), \Sigma(\vartheta)),\, N(\mu(\theta, \boldsymbol{y}), \Sigma(\theta))\big)$. For all θ, the matrices $A(\theta)$ have all eigenvalues equal to one (so then also their inverses) whence $\det \Sigma(\theta) \equiv \sigma^{2d}$; we conclude that the logarithm in (49) is null. Also, using basic properties properties of the trace and the representation for $A(\theta)A^{-1}(\vartheta)$ derived above,

$$\begin{aligned}
&\operatorname{tr}\Big(\Sigma^{-1}(\theta)\Sigma(\vartheta)\Big) - d = \operatorname{tr}\Big(\big(A^{-1}(\theta)A^{-T}(\theta)\big)^{-1}\big(A^{-1}(\vartheta)A^{-T}(\vartheta)\big)\Big) - d \\
&= \operatorname{tr}\Big(A^T(\theta)A(\theta)A^{-1}(\vartheta)A^{-T}(\vartheta)\Big) - d = \operatorname{tr}\Big(\big(A(\theta)A^{-1}(\vartheta)\big)^T A(\theta)A^{-1}(\vartheta)\Big) - d \\
&= 2\sum_{i=1}^d \operatorname{tr}\big(S^iA^{-1}(\vartheta)\big)(\vartheta_i - \theta_i) + \sum_{i=1}^d\sum_{j=1}^d \operatorname{tr}\big(A^{-T}(\vartheta)(S^i)^TS^jA^{-1}(\vartheta)\big)(\vartheta_i - \theta_i)(\vartheta_j - \theta_j).
\end{aligned}$$

The inverse of an upper-triangular matrix is upper-triangular and so, for all $i = 1, \dots, d$ and all ϑ, the matrices $S^i A^{-1}(\vartheta)$ have null trace. Denote now for any n by m matrix M, $\text{vect}(M)$ as the column vector containing the nm entries of M in any (fixed) order. Write then for $i = 1, \dots, d$, $v_i(\vartheta) = \text{vect}\big(S^i A^{-1}(\vartheta)\big)$; $v_d(\vartheta)$ is always a null vector. Note that the i, j-the element of the double sum in the previous display is given by $v_i^T(\vartheta) v_j(\vartheta)$, for $i, j = 1, \dots, d$. We conclude that the previous display can be written as

$$(\vartheta - \theta)^T \big[v_1(\vartheta) v_2(\vartheta) \dots v_d(\vartheta)\big]^T \big[v_1(\vartheta) v_2(\vartheta) \dots v_d(\vartheta)\big] (\vartheta - \theta),$$

where the matrices are written by columns.

The quadratic form in the Kullback-Leibler divergence (49) can be written, for any $\theta, \vartheta \in \Theta_d$ and $\boldsymbol{y} \in \mathbb{R}^d$, as

$$\begin{aligned}&\big(\mu(\theta, \boldsymbol{y}) - \mu(\vartheta, \boldsymbol{y})\big)^T \Sigma^{-1}(\theta) \big(\mu(\theta, \boldsymbol{y}) - \mu(\vartheta, \boldsymbol{y})\big) = \\ &\quad = \sigma^{-2} \boldsymbol{y}^T \big(B(\theta) - A(\theta) A^{-1}(\vartheta) B(\vartheta)\big)^T \big(B(\theta) - A(\theta) A^{-1}(\vartheta) B(\vartheta)\big) \boldsymbol{y}.\end{aligned}$$

Note that the matrix $B(\cdot)$ is linear in its argument and so se can write the expansion

$$B(\theta) - B(\vartheta) = B(\theta - \vartheta) = \big(S^{d-1}\big)^T (\vartheta_1 - \theta_1) + \dots + S^T (\vartheta_{d-1} - \theta_{d-1}) + I(\vartheta_d - \theta_d),$$

from where, using the representation for $A(\theta) A^{-1}(\vartheta)$ derived above, we have

$$B(\theta) - A(\theta) A^{-1}(\vartheta) B(\vartheta) = C_1(\vartheta)(\vartheta_1 - \theta_1) + C_2(\vartheta)(\vartheta_2 - \theta_2) + \dots + C_d(\vartheta)(\vartheta_d - \theta_d),$$

where $C_i(\vartheta) = \big(S^{d-i}\big)^T - S^i A^{-1}(\vartheta) B(\vartheta)$ for $i = 1, \dots, d$. We can then write, for $\theta, \vartheta \in \Theta(\rho)$ and $\boldsymbol{y} \in \mathbb{R}^d$,

$$\big(B(\theta) - A(\theta) A^{-1}(\vartheta) B(\vartheta)\big) \boldsymbol{y} = \big[C_1(\vartheta) \boldsymbol{y} \cdots C_d(\vartheta) \boldsymbol{y}\big] \big(\theta - \vartheta\big).$$

We conclude that the following representation holds

$$\begin{aligned}g_{dk}\big(\theta, \vartheta | \boldsymbol{X}_{d(k-1),d}\big) &= -\sigma^{-2} \Big(\sigma^2 \big[v_1(\vartheta) v_2(\vartheta) \dots v_d(\vartheta)\big]^T \big[v_1(\vartheta) v_2(\vartheta) \dots v_d(\vartheta)\big] + \\ &+ \big[C_1(\vartheta) \boldsymbol{X}_{d(k-1),d} \cdots C_d(\vartheta) \boldsymbol{X}_{d(k-1),d}\big]^T \big[C_1(\vartheta) \boldsymbol{X}_{d(k-1),d} \cdots C_d(\vartheta) \boldsymbol{X}_{d(k-1),d}\big]\Big) \big(\theta - \vartheta\big).\end{aligned}$$

Note that the matrix that precedes the vector $(\theta - \vartheta)$ does not depend on θ and is clearly positive semi-definite. We bound now the eigenvalues of this sum of matrices.

The first matrix in the sum above is positive semi-definite but has at least one null eigenvalue. It is also clear that the entries of this matrix are polynomials in $\vartheta_1, \dots, \vartheta_{d-1}$, such that, since $\Theta(\rho)$ is a bounded set, we have that the largest eigenvalue of this matrix is upper bounded, uniformly over Θ_d, by some constant, say, K_1, depending only on d and the diameter of $\Theta(\rho)$; we remind that this diameter is at most $(1 + \rho)^d - 1 < 2^d - 1$.

We have that

$$\begin{aligned}\text{tr}\Big(\big[C_1(\vartheta) \boldsymbol{y} \quad \cdots \quad C_d(\vartheta) \boldsymbol{y}\big]^T \big[C_1(\vartheta) \boldsymbol{y} \quad \cdots \quad C_d(\vartheta) \boldsymbol{y}\big] \Big) = \\ \boldsymbol{y}^T C_1^T(\vartheta) C_1(\vartheta) \boldsymbol{y} + \dots + \boldsymbol{y}^T C_d^T(\vartheta) C_d(\vartheta) \boldsymbol{y}.\end{aligned}$$

For each $i = 1, \dots, d-1$, the entries of the matrices $C_i^T(\vartheta)C_i(\vartheta)$, are polynomials in $\vartheta_1, \dots, \vartheta_d$; the previous display is then also upper-bounded uniformly over $\Theta(\rho)$ by, say, $K_2\boldsymbol{y}^T\boldsymbol{y}$, where K_2 is a constant which like K_1 above, depends only on d and the diameter of $\Theta(\rho)$.

To derive a lower bound on the smallest eigenvalue of the matrix in the representation for g_{dk} note that this matrix can be rewritten in the form,

$$\left[\begin{array}{ccc|c} v_{1,1}(\vartheta) & \cdots & v_{1,d-1}(\vartheta) & 0 \\ \vdots & \ddots & \vdots & \vdots \\ v_{d-1,1}(\vartheta) & \cdots & v_{d-1,d-1}(\vartheta) & 0 \\ \hline c_{d,1}(\vartheta) & \cdots & c_{d,d-1}(\vartheta) & \boldsymbol{y}^T\boldsymbol{y} \end{array}\right] + \left[\begin{array}{ccc|c} c_{1,1}(\vartheta) & \cdots & c_{1,d-1}(\vartheta) & c_{1,d}(\vartheta) \\ \vdots & \ddots & \vdots & \vdots \\ c_{d-1,1}(\vartheta) & \cdots & c_{d-1,d-1}(\vartheta) & c_{d-1,d}(\vartheta) \\ \hline 0 & \cdots & 0 & 0 \end{array}\right]$$

for $v_{i,j}(\vartheta) = \sigma^2 v_i^T(\vartheta)v_j(\vartheta)$ and $c_{i,j}(\vartheta) = \boldsymbol{y}^T C_{d-i+1}^T(\vartheta)C_{d-j+1}(\vartheta)\boldsymbol{y}$, where we swapped the last rows of the matrices. (Note that $C_d(\vartheta) \equiv I$ such that $c_{d,d}(\vartheta) = \boldsymbol{y}^T\boldsymbol{y}$ and also $v_d(\vartheta) = \mathbf{0}^T$.)

Note that the top left matrices in the block matrices above are Gram matrices and therefore positive semi-definite; the full block matrices are triangular by blocks. The matrix $[v_{i,j}(\vartheta)]_{i,j=1,\dots,d-1}$ is the Gram matrix associated with the vectors $v_1(\vartheta), \dots, v_{d-1}(\vartheta)$. Its simple to see that these vectors are linearly independent (this follows from the fact that $A^{-1}(\vartheta)$ is a triangular matrix with 1's in its main diagonal) whence the associated Gramian is actually positive definite for each ϑ. Note also that the determinant of this Gramian is a polynomial in the entries of the matrix which in turn are a polynomial in $\vartheta_1, \dots, \vartheta_d$. Since $\vartheta \in \Theta(\rho)$, which is a compact set, we conclude that the infimum of the determinant of this matrix over $\vartheta \in \Theta(\rho)$ is lower bounded by some positive constant say, K_3. Using the same reasoning we can see that its determinant is upper bounded by some constant K_4. A lower bound on the smallest eigenvalue can then be obtained by noting that for any positive definite matrix M of order d,

$$\lambda_{(1)}(M) \geq \frac{\det(M)}{\lambda_{(d)}^{d-1}(M)} \geq \frac{K_3}{K_4^{d-1}} \geq \nu > 0,$$

for some constant ν depending only on d and say, the diameter of the parameter set $\Theta(\rho)$.

We conclude that the smallest eigenvalue of the block matrix on the left is at least $\min(\nu, \boldsymbol{y}^T\boldsymbol{y})$. The block matrix on the right is clearly positive semi-definite. We conclude that the smallest eigenvalue of the matrix in the representation for g_{dk} above is lower bounded by $\min(\nu, \boldsymbol{y}^T\boldsymbol{y})$ by using Weyl's Monotonicity Theorem; cf. Bathia [1997].

Condition (A2) is simpler to check. Let $D(\boldsymbol{X}_{dk,d}, \boldsymbol{X}_{d(k-1),d}) = D(\boldsymbol{X}_{dk,2d})$ be the Toeplitz matrix associated with the vector $\boldsymbol{d}(\boldsymbol{X}_{dk,2d}) = (X_{d(k-2)+1}, \dots, X_{dk-2}, X_{dk-1})$ which is simply the vector $\boldsymbol{X}_{dk-1,2d-1}$ written backwards. Based on this, the gain (48) can be written, up to a constant depending only on σ^2 and with $\boldsymbol{x} = \boldsymbol{X}_{dk,d}$, $\boldsymbol{y} = \boldsymbol{X}_{d(k-1),d}$, in the following form

$$
\begin{aligned}
G_{dk}\big(\boldsymbol{x},\theta|\boldsymbol{y}\big) &= -\nabla_\theta\big(A(\theta)\boldsymbol{x}-B(\theta)\boldsymbol{y}\big)^T\big(A(\theta)\boldsymbol{x}-B(\theta)\boldsymbol{y}\big)\\
&= -2\big(A(\theta)\boldsymbol{x}-B(\theta)\boldsymbol{y}\big)^T\frac{\boldsymbol{\partial}\big(A(\theta)\boldsymbol{x}-B(\theta)\boldsymbol{y}\big)}{\boldsymbol{\partial}\theta}\\
&= 2\big(A(\theta)\boldsymbol{x}-B(\theta)\boldsymbol{y}\big)^T D(\boldsymbol{x},\boldsymbol{y})J,
\end{aligned}
\tag{50}
$$

where $\boldsymbol{\partial}/\boldsymbol{\partial}\theta$ represents the Jacobian operator. To verify (A2) it suffices to check that the expectation of the norm of G_{dk} is bounded. We omit the details but it is clear from the expression derived above that the norm of the gain function squared is a polynomial of degree 4 in the elements of $\boldsymbol{X}_{dk-1,2d-1}$. We've already mentioned that so long as the initial values for the autoregressive process and the noise terms have uniformly bounded p-th moments, then this transfers to the each observation X_k, so long as the sequence of parameters of the model, θ_k, lives in the parameter set $\Theta(\rho)$, for some $\rho < 1$.

As we saw above, the eigenvalues of the matrix appearing in the conditional gain vector g_{dk} are upper and lower bounded by multiples of $\|\boldsymbol{X}_{d(k-1),d}\|_2^2$. We can easily get rid of this dependence by using the scaled gain $\bar{G}_{dk}$ defined at the end of Section 4, for $s(x) = \|x\|_2^2$ and large enough κ. The derivation in (26) shows that (A2) still holds for this rescaled gain. The largest eigenvalue of the matrix in $\bar{g}_{dk}$ corresponding to $\bar{G}_{dk}$ is going to be almost surely bounded by construction. We need then to verify that the smallest eigenvalue of the matrix in $\bar{g}_{dk}$ has conditional expectation bounded away from zero such that (A1) holds. Note that

$$
\begin{aligned}
&\mathbb{E}\big[\|\boldsymbol{X}_{d(k-1),d}\|_2^2\big|\boldsymbol{X}_{d(k-2),d}\big] \geq \mathbb{E}\big[X_{d(k-1)}^2\big|\boldsymbol{X}_{d(k-2),d}\big] =\\
&\mathbb{E}\big[\big(\theta_{dk,1}X_{d(k-1)-1}+\cdots+\theta_{dk,d}X_{d(k-2)}+\xi_{d(k-1)}\big)^2\big|\boldsymbol{X}_{d(k-2),d}\big].
\end{aligned}
$$

There are three different types of terms in the sum above: a) error terms which are independent of the filtration, b) observations which are measurable with respect to the filtration, and c) observations which can be written as an error term which is independent of the filtration and a linear combination of previous observations of the process. The sum can therefore be written as the sum of two terms, namely: a) a linear combination of terms which are measurable with respect to the filtration, and b) a linear combination of error terms which are independent of the filtration. This can then be bounded in the same way as (43). We conclude that the previous display is lower bounded by σ^2.

One can then proceed as in Lemma 5 to show that for an appropriately large κ, $\mathbb{E}\big[\min(X_{d(k-1)}^2,\kappa)\big|\boldsymbol{X}_{d(k-2),d}\big]$ is positive; we omit this derivation.

For the most part, the requirement that the errors be Gaussian is not used extensively so we expect the same results hold simply under appropriate moment assumptions: one could still use the gain (50) and bound its conditional expectation directly instead of using the Kullback-Leibler representation in (49) and assure the validity of (A1) and (A2) based on moment assumptions on the error terms and on the initial conditions for the model as we did in the one dimensional case.

7 Proofs of the lemmas

Proof of Lemma 1. First suppose that $y = Mx$ for some symmetric positive definite matrix M such that $0 < \lambda_1 \leq \lambda_{(1)}(M) \leq \lambda_{(d)}(M) \leq \lambda_2 < \infty$. Then $\langle x, y\rangle = x^T Mx$ and therefore

$$0 < \lambda_1 \|x\|_2^2 \leq \lambda_{(1)}(M)\|x\|_2^2 \leq \langle x, y\rangle \leq \lambda_{(d)}(M)\|x\|_2^2 \leq \lambda_2\|x\|_2^2$$

and

$$\|y\|_2^2 = \langle y, y\rangle = x^T M^T Mx = x^T M^2 x \leq \lambda_2^2 \|x\|_2^2.$$

Now we prove the converse assertion. Suppose $x, y \in \mathbb{R}^d$ and $0 < \lambda_1'\|x\|_2^2 \leq \langle x, y\rangle \leq \lambda_2'\|x\|_2^2 < \infty$ for some $\lambda_1', \lambda_2' \in \mathbb{R}$ such that $0 < \lambda_1' \leq \lambda_2' < \infty$ and that $\|y\|_2 \leq C\|x\|_2$. Let $V = \{v = ax + by : a, b \in \mathbb{R}\}$ be the linear space spanned by x and y. First consider the case $\dim(V) = 1$, i.e., $y = \alpha x$ for some $\alpha \in \mathbb{R}$. Then $\langle y, x\rangle = \alpha\|x\|_2^2$ so that $0 < \lambda_1' \leq \alpha \leq \lambda_2' < \infty$. Thus $y = \alpha x = Mx$ with symmetric and positive $M = \alpha I$ so that $0 < \lambda_1' \leq \alpha = \lambda_{(1)}(M) = \lambda_{(d)}(M) \leq \lambda_2' < \infty$.

Now consider the case $\dim(V) = 2$. Let $e_1 = x/\|x\|_2$ and $\{e_1, e_2\}$ be an orthonormal basis of V. Then

$$\begin{aligned} x &= \|x\|_2 e_1 \\ y &= \alpha e_1 + \beta e_2. \end{aligned}$$

The conditions $\lambda_1'\|x\|_2^2 \leq \langle x, y\rangle = \alpha\|x\|_2 \leq \lambda_2'\|x\|_2^2$ and $\|y\|_2 = \sqrt{\alpha^2 + \beta^2} \leq C\|x\|_2$ imply that

$$\lambda_1'\|x\|_2 \leq \alpha \leq \min\{\lambda_2', C\}\|x\|_2, \quad |\beta| \leq C\|x\|_2.$$

Let e_2 be chosen in such a way that $\beta > 0$ (which is always possible.) Now, we change the basis of V as follows:

$$\begin{aligned} e_1' &= \cos(\theta)e_1 - \sin(\theta)e_2, \\ e_2' &= \sin(\theta)e_1 + \cos(\theta)e_2. \end{aligned}$$

We thus rotate the basis $\{e_1, e_2\}$ by the angle θ. In these new basis we have

$$\begin{aligned} x &= \|x\|_2 \cos(\theta)e_1' + \|x\|_2 \sin(\theta)e_2' = \alpha_x e_1' + \beta_x e_2', \\ y &= (\alpha\cos(\theta) - \beta\sin(\theta))e_1' + (\alpha\sin(\theta) + \beta\cos(\theta))e_2' = \alpha_y e_1' + \beta_y e_2'. \end{aligned}$$

Recall that $\alpha, \beta > 0$. Take $\theta \in (0, \pi/2)$ such that $\alpha\cos(\theta) - \beta\sin(\theta) = \frac{1}{2}\alpha\cos(\theta)$ (i.e., $\tan(\theta) = \frac{\alpha}{2\beta}$). Then we have that

$$\frac{\lambda_1'}{2} \leq \frac{\alpha}{2\|x\|_2} = \frac{\alpha_y}{\alpha_x} \leq \frac{\min\{\lambda_2', C\}}{2}, \; \lambda_1' \leq \frac{\alpha}{\|x\|_2} \leq \frac{\beta_y}{\beta_x} \leq \frac{\alpha}{\|x\|_2} + \frac{2\beta^2}{\alpha\|x\|_2} \leq \min\{\lambda_2', C\} + \frac{2C^2}{\lambda_1'}.$$

Take then $\lambda_1 = \lambda_1'/2$ and $\lambda_2 = \min\{\lambda_2', C\} + 2C^2/\lambda_1'$.

Let $\{e_3, \ldots, e_d\}$ be the orthonormal basis of $V^\perp$, so that $b = \{e_1', e_2', e_3, \ldots, e_d\}$ is an orthonormal basis of $\mathbb{R}^d$. Take

$$M' = \left[\begin{array}{c|c} D & 0 \\ \hline 0 & I_{d-2} \end{array}\right] \quad \text{with} \quad D = \begin{bmatrix} \alpha_y/\alpha_x & 0 \\ 0 & \beta_y/\beta_x \end{bmatrix}$$

where the 0's indicate null matrices of the appropriate orders. We then have that $y = M'x$ in the basis b and $\lambda_1 \le \lambda_{(1)}(M') \le \lambda_{(d)}(M') \le \lambda_2$. We can finally obtain M by using the (orthogonal) change of basis matrix E from basis b to the canonical basis of $\mathbb{R}^d$ as $M = E^{-1}M'E = E^T M'E$. Note that M has the same eigenvalues as M' (which are all positive and finite) and is symmetric. □

Proof of Lemma 2. For the sake of brevity, we use the notations $\theta_k = \theta_k(\boldsymbol{X}_{k-1})$, $G_k = G(X_k, \hat{\theta}_k|\boldsymbol{X}_{k-1})$ and $g_k = g(\hat{\theta}_k, \theta_k|\boldsymbol{X}_{k-1})$, $k \in \mathbb{N}$, $\mathcal{F}_k = \sigma(\boldsymbol{X}_k)$ is the σ-field generated by $\boldsymbol{X}_k = (X_0, X_2, \dots, X_k)$.

Recall that Θ is compact so that $\sup_{\theta\in\Theta} \|\theta\|_2 \le C_\Theta$. First assume $\mathbb{E}\|\hat{\theta}_k\|_2^2 \le KC_\Theta^2$. By (5) and (6), we obtain

$$\mathbb{E}\|G_k\|_2^2 = \mathbb{E}\|G_k - g_k + g_k)\|_2^2 \le 2C + 4L(\mathbb{E}\|\theta_k\|_2^2 + \mathbb{E}\|\hat{\theta}_k\|_2^2) \le 2C + 4L(K+1)C_\Theta^2 = C_1,$$

which implies, in view of (7) and $\gamma_k \le \Gamma$,

$$\mathbb{E}\|\hat{\theta}_{k+1}\|_2^2 \le 2\mathbb{E}\|\hat{\theta}_k\|_2^2 + 2\gamma_k^2\mathbb{E}\|G_k\|_2^2 \le 2KC_\Theta^2 + 2\Gamma^2 C_1 = C_2.$$

Next, consider the case $\mathbb{E}\|\hat{\theta}_k\|_2^2 > KC_\Theta^2$ which of course implies $\mathbb{E}\|\hat{\theta}_k\|_2^2 > K\mathbb{E}\|\theta_k\|_2^2$. As M_k is a symmetric positive definite matrix such that $0 < A \le \lambda_{(1)}(M_k) \le \lambda_{(d)}(M_k) \le B < \infty$, by the Cauchy-Schwarz inequality, we have that

$$\hat{\theta}_k^T M_k \theta_k \le |\hat{\theta}_k^T M_k \theta_k| \le \big(\hat{\theta}_k^T M_k \hat{\theta}_k\big)^{1/2}\big(\theta_k^T M_k \theta_k\big)^{1/2} \le B\|\hat{\theta}_k\|_2\|\theta_k\|_2.$$

By using the last relation, (2), (5), (6) and (7), we evaluate $\mathbb{E}\|\hat{\theta}_{k+1}\|_2^2$:

$$\begin{aligned}
\mathbb{E}\|\hat{\theta}_{k+1}\|_2^2 \le& \mathbb{E}\|\hat{\theta}_k\|_2^2 + 2\gamma_k\mathbb{E}\big[\hat{\theta}_k^T\mathbb{E}(G_k|\mathcal{F}_{k-1})\big] + \gamma_k^2\mathbb{E}\|G_k\|_2^2\\
\le& \mathbb{E}\|\hat{\theta}_k\|_2^2 - 2\gamma_k\mathbb{E}\big(\hat{\theta}_k^T M_k(\hat{\theta}_k - \theta_k)\big) + \gamma_k^2\big[2C + 4L(\mathbb{E}\|\theta_k\|_2^2 + \mathbb{E}\|\hat{\theta}_k\|_2^2)\big]\\
\le& \mathbb{E}\|\hat{\theta}_k\|_2^2 - 2\gamma_k\big[A\mathbb{E}\|\hat{\theta}_k\|_2^2 - \mathbb{E}\big(\hat{\theta}_k^T M_k \theta_k\big)\big] + \gamma_k^2\big[2C + 4LC_\Theta^2 + 4L\mathbb{E}\|\hat{\theta}_k\|_2^2)\big]\\
\le& \mathbb{E}\|\hat{\theta}_k\|_2^2 - 2\gamma_k\big[A\mathbb{E}\|\hat{\theta}_k\|_2^2 - B\mathbb{E}\big(\|\hat{\theta}_k\|_2\|\theta_k\|_2\big)\big] + \gamma_k^2\big[2C + 4LC_\Theta^2 + 4L\mathbb{E}\|\hat{\theta}_k\|_2^2)\big].
\end{aligned}$$

From $\mathbb{E}\|\hat{\theta}_k\|_2^2 > K\mathbb{E}\|\theta_k\|_2^2$ and the Cauchy-Schwarz inequality, it follows that $\mathbb{E}\|\hat{\theta}_k\|_2\|\theta_k\|_2 \le \big(\mathbb{E}\|\hat{\theta}_k\|_2^2\mathbb{E}\|\theta_k\|_2^2\big)^{1/2} \le \mathbb{E}\|\hat{\theta}_k\|_2^2/\sqrt{K}$. Using this, we proceed by bounding the previous display as follows:

$$\begin{aligned}
\le& \mathbb{E}\|\hat{\theta}_k\|_2^2 - 2\gamma_k\big(A\mathbb{E}\|\hat{\theta}_k\|_2^2 - B(\mathbb{E}\|\theta_k\|_2^2\mathbb{E}\|\hat{\theta}_k\|_2^2)^{1/2}\big) + \gamma_k^2\big[2C + 4LC_\Theta^2 + 4L\mathbb{E}\|\hat{\theta}_k\|_2^2)\big]\\
\le& \mathbb{E}\|\hat{\theta}_k\|_2^2 - \gamma_k\mathbb{E}\|\hat{\theta}_k\|_2^2\Big(2A - \frac{2B}{\sqrt{K}} - \gamma_k 4L\Big) + \gamma_k^2(2C + 4LC_\Theta^2)\\
\le& \mathbb{E}\|\hat{\theta}_k\|_2^2 - \gamma_k\mathbb{E}\|\hat{\theta}_k\|_2^2\Big(2A - \frac{2B}{\sqrt{K}} - \gamma_k 4L\Big) + \gamma_k^2(2C + 4LC_\Theta^2)\frac{\mathbb{E}\|\hat{\theta}_k\|_2^2}{KC_\Theta^2}\\
=& \mathbb{E}\|\hat{\theta}_k\|_2^2 - \gamma_k\mathbb{E}\|\hat{\theta}_k\|_2^2\Big(2A - \frac{2B}{\sqrt{K}} - \gamma_k\frac{4LC_\Theta^2(K+1) + 2C}{KC_\Theta^2}\Big) \le \mathbb{E}\|\hat{\theta}_k\|_2^2,
\end{aligned}$$

for sufficiently large K and sufficiently small γ_k. Thus, for sufficiently large K and sufficiently small γ_k, $\mathbb{E}\|\hat{\theta}_{k+1}\|_2^2 \le C_2$. □

Lemma 3. *Let M be a symmetrical positive-definite matrix of order d with (increasing) eigenvalues $\lambda_{(i)}(M)$, the smallest and largest of which we denote as $\lambda_{(1)}(M)$ and $\lambda_{(d)}(M)$ respectively. Then, for $\gamma > 0$ such that $\gamma\lambda_{(d)}(M) < 1$, and constants $K_p > 0$, $p \in \mathbb{N}$,*

$$\|M\|_p \leq K_p \|M\|_2 = K_p \lambda_{(d)}(M),$$
$$0 < \lambda_{(1)}(I - \gamma M) \leq \lambda_{(d)}(I - \gamma M) = 1 - \gamma\lambda_{(1)}(M) < 1,$$

where for $p \in \mathbb{N}$, $\|M\|_p$ is the operator norm induced by l_p.

Proof. Note that for $x \in \mathbb{R}^d$, if

$$R_2^p = \max_{x \neq 0} \frac{\|x\|_p}{\|x\|_2}, \quad R_p^2 = \max_{x \neq 0} \frac{\|x\|_2}{\|x\|_p},$$

then it follows (cf. Horn and Johnson [1988, Theorem 5.6.18])

$$\max_{M \neq 0} \frac{\|M\|_p}{\|M\|_2} = R_2^p R_p^2 = K_p.$$

We then have (c.f. Horn and Johnson [1988, Section 5.6.6]) that for M a real, symmetrical, positive definite matrix, where $\lambda_{(i)}(M)$ is the i-th largest eigenvalue of a matrix M,

$$\|M\|_2 = \max_i \sqrt{\lambda_i(M^T M)} = \max_i \sqrt{\lambda_{(i)}(M^2)} = \lambda_{(d)}(M).$$

The first statement then follows. Note that by application of the Hölder inequality, we have $\|x\|_p \leq d^{(q-p)/(qp)} \|x\|_q$ for $p \leq q$ and so we can take $K_p = d^{(p-1)/(2p)}$ if $p \geq 2$ and $K_p = d^{1/2}$ if $p = 1$.

Its straightforward to check that the matrix $I - \gamma M$ has eigenvalues $1 - \gamma\lambda_i$. Now, if $\gamma\lambda_{(d)}(M) < 1$ then for all $i = 1, \ldots, d$, $0 < \gamma\lambda_{(1)}(M) \leq \gamma\lambda_i \leq \gamma\lambda_{(d)}(M) < 1$ implying $1 > 1 - \gamma\lambda_{(1)}(M) \geq 1 - \gamma\lambda_i \geq 1 - \gamma\lambda_{(d)}(M) > 0$ and so $\max_{i=1,\ldots,d} |1 - \gamma\lambda_i| = 1 - \gamma\lambda_{(1)}(M) < 1$. □

Lemma 4 (Abel Tranformation). *For $k_0, k \in \mathbb{N}$ such that $k_0 \leq k$, let $a_i \in \mathbb{R}^d$, $i = k_0, \ldots, k$, B_i, $i = k_0, \ldots, k$, be square $d \times d$ matrices and $A_i = \sum_{j=k_0}^{i} a_j$, $i = k_0, \ldots, k$. Then*

$$\sum_{i=k_0}^{k} B_i a_i = \sum_{i=k_0}^{k-1} (B_i - B_{i+1}) A_i + B_k A_k.$$

Proof. We shall prove this by induction on k. For $k = k_0$ we simply have $B_{k_0} a_{k_0} = B_{k_0} A_{k_0} = B_{k_0} a_{k_0}$ and the assertion holds. Let us assume then that the equality holds for

$k = n$ and let us prove the result for $k = n + 1$. We have

$$\begin{aligned}\sum_{i=k_0}^{n+1} B_i a_i = \sum_{i=k_0}^{n} B_i a_i + B_{n+1}a_{n+1} &= \sum_{i=k_0}^{n-1} (B_i - B_{i+1})A_i + B_n A_n + B_{n+1}a_{n+1} \\ &= \sum_{i=k_0}^{n} (B_i - B_{i+1})A_i - (B_n - B_{n+1})A_n + B_n A_n + B_{n+1}a_{n+1} \\ &= \sum_{i=k_0}^{n} (B_i - B_{i+1})A_i + B_{n+1}A_{n+1}.\end{aligned}$$

□

Lemma 5. *Consider an AR(1) model with a random, drifting parameter θ_k,*

$$X_k = X_{k-1}\theta_k + \xi_k, \quad k \in \mathbb{N},$$

where the random variables ξ_k are independent of $\sigma(X_0, \dots, X_{k-1})$, the σ algebra generated by $\boldsymbol{X}_{k-1}$ and for all $k \in \mathbb{N}$, $\mathbb{E}\xi_k = \mathbb{E}\xi_k^3 = 0$, $\mathbb{E}\xi_k^2 = \sigma^2 > 0$ and, for some constant $0 \le c < 5$, $\mathbb{E}\xi_k^4 = c\,\sigma^4$. Let also X_0 be such that $\mathbb{E}X_0^2$ and $\mathbb{E}X_0^4$ are bounded. We assume that the drifting parameter θ_k is measurable with respect to $\sigma(\boldsymbol{X}_{k-1})$, and verifies $|\theta_k| \le q < 1$, almost surely, for every $k \in \mathbb{N}$. Then, for any s such that $4s \ge (9-c)\sigma^2$,

$$\mathbb{E}\big[\min\big(X_t^2, s\big)\big|X_{t-1}\big] \ge \frac{5-c}{4}\sigma^2.$$

Proof. Note first that since $\sigma^2 > 0$, if $\mathbb{E}X_0^2$ and $\mathbb{E}X_0^4$ are bounded then we can write $\mathbb{E}X_0^2 \le c_1\sigma^2$ and $\mathbb{E}X_0^4 \le c_2\sigma^4$ for some $c_1, c_2 \ge 0$. Using the independence of the noise and the bound on the norm of the autoregressive parameters we have that

$$\mathbb{E}X_k^2 = \mathbb{E}(X_{k-1}\theta_k + \xi_k)^2 = \mathbb{E}[X_{k-1}^2\theta_k^2] + 2\mathbb{E}[X_{k-1}\theta_k]\mathbb{E}\xi_k + \mathbb{E}\xi_k^2 \le q^2\mathbb{E}X_{k-1}^2 + \sigma^2,$$

and by using this recursion we conclude that

$$\mathbb{E}X_k^2 \le q^{2k}\mathbb{E}X_0^2 + \sigma^2\sum_{i=1}^{k-1} q^{2i} \le \sigma^2\left(c_1 + \frac{1}{1-q^2}\right) < \infty.$$

Using the previous display and proceeding in the same way,

$$\mathbb{E}X_k^4 = \mathbb{E}(X_{k-1}\theta_k + \xi_k)^4 \le q^4\mathbb{E}X_{k-1}^4 + 6q^2\sigma^2\mathbb{E}X_{k-1}^2 + c\,\sigma^4 \le q^4\mathbb{E}X_{k-1}^4 + \sigma^4\kappa,$$

with $\kappa = c + 6q^2c_1 + 6q^2/(1-q^2)$. Using this recursion we have that

$$\mathbb{E}X_k^4 \le q^{4k}\mathbb{E}X_0^4 + \sigma^4\kappa\sum_{i=1}^{k-1} q^{4i} \le \sigma^4\left(c_2 + \frac{\kappa}{1-q^4}\right) < \infty.$$

We can now use basic properties of the conditional expectation to see that,

$$\begin{aligned}\mathbb{E}\big[X_k^2\big|X_{t-1}\big] &= X_{k-1}^2\theta_k^2 + 2X_{k-1}\theta_k\mathbb{E}\xi_k + \mathbb{E}\xi_k^2 = X_{k-1}^2\theta_k^2 + \sigma^2,\\ \mathbb{E}\big[X_k^4\big|X_{t-1}\big] &= X_{k-1}^4\theta_k^4 - 4X_{k-1}^3\theta_k^3\mathbb{E}\xi_k + 6X_{k-1}^2\theta_k^2\mathbb{E}\xi_k^2 - 4X_{k-1}\theta_k\mathbb{E}\xi_k^3 + \mathbb{E}\xi_k^4 =\\ &= X_{k-1}^4\theta_k^4 + 6X_{k-1}^2\theta_k^2\sigma^2 + c\,\sigma^4.\end{aligned}$$

For $a, b \in \mathbb{R}$ we have $\min(a,b) = (a+b)/2 - |a-b|/2$ and so, by the Cauchy-Schwarz inequality and the last display,

$$\begin{aligned}\mathbb{E}\Big[\min\Big(X_t^2, \rho\,\sigma^2\Big)\Big|X_{t-1}\Big] &= \mathbb{E}\Big[\frac{1}{2}X_k^2 + \frac{\rho}{2}\sigma^2 - \frac{1}{2}\big|X_k^2 - \rho\sigma^2\big|\Big|X_{t-1}\Big]\\ &\geq \frac{1}{2}X_{k-1}^2\theta_k^2 + \frac{\rho+1}{2}\sigma^2 - \frac{1}{2}\Big(\mathbb{E}\Big[\big(X_k^2-\rho\sigma^2\big)^2\Big|X_{t-1}\Big]\Big)^{1/2},\end{aligned}$$

for $\rho > 0$. We now have, by plugging in the expressions derived above and simplifying,

$$\begin{aligned}\mathbb{E}\Big[\big(X_k^2-\rho\sigma^2\big)^2\Big|X_{t-1}\Big] &= \mathbb{E}\big[X_k^4\big|X_{t-1}\big] - 2\rho\sigma^2\mathbb{E}\big[X_k^2\big|X_{t-1}\big] + \rho^2\sigma^4 =\\ &= X_{k-1}^4\theta_k^4 + 2(3-\rho)X_{k-1}^2\theta_k^2\sigma^2 + (c - 2\rho + \rho^2)\sigma^4 = \Big(X_{k-1}^2\theta_k^2 + \frac{c+3}{4}\sigma^2\Big)^2,\end{aligned}$$

if we pick $\rho = (9-c)/4 > 1$. Combining the previous two displays we conclude that

$$\mathbb{E}\Big[\min\Big(X_t^2, \frac{9-c}{4}\sigma^2\Big)\Big|X_{t-1}\Big] \geq \frac{5-c}{4}\sigma^2,$$

and the statement of the lemma follows a fortiori. □

References

R. Bathia. *Matrix Analysis*, volume 169 of *Graduate texts in Mathematics.* Springer-Verlag, New York, 1997.

E. Belitser. Recursive estimation of a drifting autoregressive parameter. *Ann. Statist.*, 28 (3):860–870, 2000.

Eduard Belitser and Paulo Serra. On properties of the algorithm for pursuing a drifting quantile. *Automation and Remote Control*, 74(4):613–627, April 2013.

A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximation.* New York, Berlin: Springer-Verlag., 1990.

J.-M. Brossier. *Egalization Adaptive et Estimation de Phase: Application aux Communications Sous-Marines.* PhD thesis, Institut National Polytechnique de Grenoble, 1992.

Yuan Shih Chow and Henry Teicher. *Probability Theory. Independence, Interchangeability, Martingales.* Springer texts in Statistics. Springer Verlag, New York, second edition, 1988.

B. Delyon and A. Juditsky. Asymptotical study of parameter tracking algorithms. *SIAM Journal on Control and Optimization*, 33(1):323–345, January 1995.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, 1988.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

H. J. Kushner and D. S. Clark. *Stochastic Approximation for Constrained and Unconstrained Systems.* Springer Verlag, New York, 1978.

H. J. Kushner and J. Yang. Analysis of adaptive step-size sa algorithms for parameter tracking. *IEEE Trans. Autom. Control*, 40:1403–1410, 1995.

H. J Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications.* Berlin and New York: Springer-Verlag, 2003.

L. Ljung and T. Söderström. *Theory and Practice of Recursive Identification.* MIT Press, Cambridge, MA., 1983.

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

M.B. Nevelson and R. Z. Khasminskii. *Stochastic Approximation and Recursive Estimation.*, volume 47 of *Translation of Mathematical Monographs.* American American Mathematical Society, 1976.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

YaZ. Tsypkin. *Adaptation and Learning in Automatic Systems.* Academic Press, New York, 1971.

M. T. Wasan. *Stochastic Approximation.* Cambridge University Press, 1969.