Recursive Algorithms for Estimation of Hidden Markov Models and Autoregressive Models With Markov Regime

Vikram Krishnamurthy, *Senior Member, IEEE*, and George Gang Yin, *Senior Member, IEEE*

***Abstract*—This paper is concerned with recursive algorithms for the estimation of hidden Markov models (HMMs) and autoregressive (AR) models under Markov regime. Convergence and rate of convergence results are derived. Acceleration of convergence by averaging of the iterates and the observations are treated. Finally, constant step-size tracking algorithms are presented and examined.**

***Index Terms*—Convergence, hidden Markov estimation, rate of convergence, recursive estimation.**

I. Introduction

MOTIVATED by many important applications in signal processing, speech recognition, communication systems, neural physiology, and environment modeling, in this paper, we consider recursive (online) estimation of the parameters of hidden Markov models (HMMs) and jump Markov autoregressive systems (also known as autoregressive processes with Markov regime), and develop stochastic approximation algorithms to carry out the estimation task. Our main effort is to prove the convergence and rate of convergence of these recursive estimation algorithms.

An HMM is a discrete-time stochastic process with two components $\{X_n, Y_n\}$ such that $\{X_n\}$ is a finite-state Markov chain and given $\{X_n\}$, $\{Y_n\}$ is a sequence of conditionally independent random variables; the conditional distribution of Y_n depends only of X_n. It is termed a hidden Markov chain since $\{X_n\}$ is not observable and one has to rely on $\{Y_n\}$ for any statistical inference task. Such models have been widely used in several areas including speech recognition and neurobiology; see [34] and the references therein.

In this paper, we consider both the standard HMMs and the more general autoregressive models under Markov regime, in which the autoregressive parameters switch in time according to the realization of a finite-state Markov chain, which are widely used in econometrics [15]–[17], statistical signal processing, and maneuvering target tracking (see [21] and the references therein). For such models, the distribution of the observables Y_n depend not only on X_n, but also on $Y_{n-1}, \ldots, Y_{n-d}$. Equivalently, Y_n is obtained by a regression on $Y_{n-1}, \ldots, Y_{n-d}$, where d is the order of regression. The regression functions involved can be either linear or nonlinear. Our objective is to design and analyze the properties of recursive estimators for the parameters of such autoregressive (AR) processes with Markov regime. Strong consistency of the maximum-likelihood (ML) estimator for AR processes with Markov regime was recently proved in [21]. Compared with that reference, our effort here is on the analysis of asymptotic properties of recursive algorithms for parameter estimation.

Recently, Rydén [35], [36] proposed a batch recursive algorithm for parameter estimation of standard HMMs. His main idea is to use a stochastic approximation type algorithm on batches of data of length ≥ 2. He proved the consistency by using the classical result of Kushner and Clark [22]; he also suggested an averaging approach in light of the recent development due to Polyak [33] and Ruppert [37] (see also [24] and [44]). LeGland and Mevel [29] have proved the consistency of a stochastic approximation algorithm for parameter estimation of HMMs called the recursive maximum-likelihood estimation (RMLE) algorithm. The RMLE algorithm in [29] has the advantage over the recursive-batch approach in [35] in that it is truly recursive. In Dey, Krishnamurthy, and Salmon-Legagneur [11] and Holst, Lindgren, Holst, and M. Thuvesholmen [19], stochastic approximation algorithms are presented for estimating the parameters of AR processes with Markov regime. However, these papers only provide simulation results of these algorithms and no proof of convergence or asymptotic normality is given. For the special case when the observations Y_n belong to a finite set (i.e., Y_n is a probabilistic function of a Markov chain), Arapostathis and Marcus [1] derived and analyzed recursive parameter estimation algorithms.

The main contributions of this paper are as follows.

1) We present the asymptotic analysis of the *recursive maximum-likelihood estimation* (RMLE) algorithm for estimating the parameters of HMMs and AR processes with Markov regime. We extend and generalize the results in [29] to AR models with Markov regime. Note that it is not possible to extend the batch recursive algorithm and the analysis in [35] to AR models with Markov regime. This

Manuscript received July 22, 1999; revised July 8, 2001. This work was supported in part by Australian Research Council Grant. The work of V. Krishnamurthy was supported in part by the Special Research Center for Ultra Broadband Information Networks and the Center for Expertise in Networked Decision Systems. The work of G. G. Yin was supported in part by the National Science Foundation under Grant DMS-9877090.

V. Krishnamurthy is with the Department of Electrical and Electronic Engineering, University of Melbourne, Victoria 3010, Australia (e-mail: vikram@ee.mu.oz.au).

G. G. Yin is with the Department of Mathematics, Wayne State University, Detroit, MI 48202 USA (gyin@math.wayne.edu).

Communicated by S. R. Kulkarni, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(02)00315-2.

is because the algorithm in [35] requires precise knowledge of the distribution of the state given past measurements at time instants $m, 2m, 3m, \ldots$, where $m \geq 2$. In the RMLE algorithm presented in this paper, the initial distribution of the state (at time 0) given past observations is forgotten exponentially fast and hence is asymptotically negligible.

In Section III, we study the convergence and rate of convergence issues. Different from that of [29] and [35], we use state-of-the-art techniques in stochastic approximation (see Kushner and Yin [25]). As a result, the assumptions required are weaker and our formulation and results are more general than that of Rydén [35], [36] and LeGland and Mevel [29] because we are dealing with suitably scaled sequences of the iterates that are treated as stochastic processes rather than random variables. Our approach captures the dynamic evolution of the RMLE algorithm. As a consequence, using weak convergence methods we can analyze the tracking properties of the RMLE algorithms when the parameters are time varying (see Sections III and V for details).

2) In Section IV, a modified version of the RMLE algorithm that uses averaging in both the observations and the iterates for accelerating the convergence rate is given and analyzed.
3) In Section V, a constant step size version of the RMLE for tracking variations in the parameters of HMMs and AR processes with Markov regime is analyzed.
4) In Section VI, numerical examples are presented that illustrate the performance of the algorithms for both linear and nonlinear AR processes with Markov regime.

II. Problem Formulation

A. Signal Model for HMM and AR Model With Markov Regime

Our signal model is defined on the probability space $(\Omega, \mathcal{F}, P)$ as follows. Let $\{X_k\}_{k=0}^{\infty}$ be a Markov chain with finite state space $S = \{1, \ldots, r\}$, where r is fixed and known. For $i, j = 1, \ldots, r$, the transition probabilities

$$a_{ij} = P(X_n = j | X_{n-1} = i) = a_{ij}(\varphi)$$

are functions of a parameter (vector) φ in a compact subset Φ of an Euclidean space. Write $A(\varphi) = (a_{ij}(\varphi))$.

For the AR model with Markov regime, for $n \geq 1$, the observed real-valued process $\{Y_k\}_{k=-d+1}^{\infty}$ is defined by

$$Y_n = g\left(Y_{n-1}, \ldots, Y_{n-d}, e_n; \theta_{X_n}(\varphi)\right)$$

where $\{g(\cdot; \theta) : \theta \in \Theta\}$ is a family of real-valued functions on $\mathbb{R}^{d+1}$, indexed by a parameter $\theta \in \Theta$, $\{e_n\}$ is a scalar sequence of independent and identically distributed (i.i.d.) random variables, $d > 0$ is a fixed and known integer, and Θ is a Euclidean space with coordinate projections θ_i, $i = 1, 2, \ldots, r$, where θ_i: $\Phi \rightarrow \Theta$. We will discuss the distribution of the initial vector $(Y_{-d+1}, \ldots, Y_0)$ below. Assume that at each time $n \geq 0$, each conditional distribution has a density with respect to (w.r.t.) the Lebesgue measure and denote this density by $f(y_n | y_{n-1}, \ldots, y_{n-d}; \theta)$. Let p be the dimension of the vector-valued parameter φ. Other than Section V (where we consider tracking algorithms), we assume that there is a fixed $\varphi^* \in \Phi$, which is the "true" parameter. Our objective is to design a recursive algorithm to estimate φ^*.

In the HMM case, the observed real-valued process $\{Y_k\}_{k=1}^{\infty}$ is defined by

$$Y_n = g\left(e_n; \theta_{X_n}(\varphi)\right).$$

Clearly, it is a special case of the above AR model with $d = 0$.

Remark 2.1: For notational simplicity, we have assumed that $\{e_k\}$ and $\{Y_k\}$ are scalar-valued. The results straightforwardly generalize to vector-valued processes.

Notation: For notational convenience let

$$Z_n \overset{\text{def}}{=} (Y_n, \ldots, Y_{n-d}). \tag{1}$$

For the HMM case, $d = 0$, i.e., $Z_n = Y_n$. In the subsequent development, we often use κ as a generic positive constant; its values may change for different usage. For a function $h(\cdot)$, we use both $(\partial/\partial\varphi)h$ and h_φ to denote the partial derivative with respect to φ. For a vector or a matrix v, v' denotes its transpose. For an integer ℓ, let $\mathbf{1}_\ell$ and $\mathbf{0}_\ell$, respectively, denote the ℓ-dimensional column vector in which each element is 1 and 0, respectively.

Define the r-dimensional vector $f(z_k; \varphi)$ and $r \times r$ matrix $F(z_k; \varphi)$ by

$$f(z_k; \varphi) = [f_1(z_k; \varphi), \ldots, f_r(z_k; \varphi)]$$

and

$$F(z_k; \varphi) = \operatorname{diag}[f_1(z_k; \varphi), \ldots, f_r(k, \varphi)]$$

where

$$f_i(z_k; \varphi) = f\left(y_k | y_{k-1}, \ldots, y_{k-d}; \theta_i(\varphi)\right), \quad i = 1, \ldots, r. \tag{2}$$

Let the conditional probability distribution of $(Y_1, \ldots, Y_n)$ under P_φ be defined as

$$\begin{aligned} &p_n\left(y_1, \ldots, y_n | y_{-d+1}, \ldots, y_0; \varphi\right) \\ &\qquad = P_\varphi\left(Y_n \in dy_n, \ldots, Y_1 \in dy_1 | y_{-d+1}, \ldots, y_0\right). \end{aligned}$$

It is straightforward to show that [21]

$$\begin{aligned} &p\left(y_1, \ldots, y_n | y_{-d+1}, \ldots, y_0; \varphi\right) \\ &\quad = \sum_{x_1=1}^{r} P_\varphi\left(X_1 = x_1 | y_0, \ldots, y_{-d+1}\right) \prod_{k=1}^{n} F(z_k; \varphi) A(\varphi) \mathbf{1}_r \end{aligned}$$

The initial choice of $P_\varphi(X_1 = x_1 | y_0, \ldots, y_{-d+1})$ is unimportant since it does not affect the convergence analysis of the estimators—it may be taken as an arbitrary stochastic vector with positive entries. (The idea of substituting the true likelihood by the conditional likelihood given an initial sequence of observations goes back to Mann and Wald [31].)

Preliminary Assumptions: Throughout the rest of the paper, we assume the following conditions hold.

C1) The transition probability matrix $A(\varphi^*)$ is positive, i.e., $a_{ij}(\varphi^*) \geq \tilde{\epsilon}$ for all $i, j \in \{1, 2, \ldots, r\}$ for some known $\tilde{\epsilon} > 0$.The process

$$\{Q_k\} = \{X_k, Z_k\} = \{X_k, Y_k, \ldots, Y_{k-d}\}$$

is a geometrically ergodic Markov chain on the state space $S \times \mathbb{R}^{d+1}$ under φ^*. Let γ denote the unique invariant measure of $\{Q_k\}$.

Remark 2.2: For the HMM case, C1) can be relaxed to the condition that the transition probability matrix $A(\varphi^*)$ is aperiodic and irreducible, see [27].

For the AR model with Markov regime, in general, it is difficult to verify the geometric ergodicity of Q_k for a given parameter φ. In [43], it is shown that for the model

$$Y_n = g(Y_{n-1}, \dots, Y_{n-d}; \theta_{X_n}(\varphi)) + e_n$$

Q_n is V-uniformly ergodic under the following conditions (note that V-uniform ergodicity implies geometric ergodicity).

i) Sublinearity: The r mappings

$$[Y_{n-1}, \dots, Y_{n-d}] \to g(Y_{n-1}, \dots, Y_{n-d}; \theta_i(\varphi)), \qquad i = 1, \dots, r$$

are continuous and there are positive constants α_i and β_i such that for some norm $|\cdot|$ on $\mathbb{R}^d$

$$|g(y_{n-1}, \dots, y_{n-d}; \theta_i(\varphi))| \le \alpha_i \left|[y_{n-1}, \dots, y_{n-d}]'\right| + \beta_i.$$

ii) For some $s \ge 1$, $E|e_n|^s < \infty$ for $s \ge 1$. The spectral radius $\rho(\overline{A}_s) < 1$ where

$$\overline{A}_s = \begin{bmatrix} \alpha_1^s a_{11} & \cdots & \alpha_r^s a_{1r} \\ \vdots & \vdots & \vdots \\ \alpha_1^s a_{r1} & \cdots & \alpha_r^s a_{rr} \end{bmatrix}.$$

iii) The marginal density of e_n is positive.

The V-uniform ergodicity in turn implies that the following strong law of large numbers and central limit theorem holds for Q_n. Let ψ: $S \times \mathbb{R}^{d+1} \to \mathbb{R}$ denote a Borel measurable function with $\psi \in \mathcal{B}(s)$ where for $s > 1$, $(x, z) \in S \times \mathbb{R}^{d+1}$

$$\mathcal{B}(s) \overset{\text{def}}{=} \{\psi\text{: } \psi(x, z) \le \text{const}\,(1 + |z|^s)\}.$$

A) The following strong law of large numbers holds for $\psi \in \mathcal{B}(s)$:

$$\frac{1}{n}\sum_{k=1}^{n} \psi(X_k, Y_k) \to \gamma(\psi) \overset{\text{def}}{=} \int_{S \times \mathbb{R}^{d+1}} \psi(x, z)\, d\gamma(x, z), \qquad P_{\varphi^*} \text{ a.s.}$$

B) Define $\overline{\psi} = \psi - \gamma(\psi)$ and

$$\Sigma_\psi = E_\gamma[\overline{\psi}^2(X_0, Z_0)] + 2\sum_{k=1}^{\infty} E_\gamma[\overline{\psi}(X_0, Z_0)\overline{\psi}(X_k, Z_k)].$$

Then for all $\psi \in \mathcal{B}(s/2)$, Σ_ψ is well defined, nonnegative, and finite. If $\Sigma_\psi > 0$ then the following central limit theorem holds:

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}[\psi(X_k, Z_k) - \gamma(\psi)] \to N(0, \Sigma_\psi) \text{ in distribution.}$$

For simplicity, sometimes we write $\psi(Q_k)$ in lieu of $\psi(X_k, Z_k)$.

C2) The mapping $\varphi \to A(\varphi)$ is twice differentiable with bounded first and second derivatives and Lipschitz continuous second derivative. For any $y_k, y_{k-1}, \dots, y_{k-d+1}$, the mapping $\varphi \to f(z_k; \varphi)$ is three times differentiable. $f(z_k; \varphi)$ is continuous on $\mathbb{R}^{d+1}$ for each $\theta \in \Theta$.

C3) For each $i = 1, \dots, r$, the conditional probability corresponding to the true parameter,

$$P_{\varphi^*}(X_k = i | Y_{k-1}, \dots, Y_{k-d})$$

is continuous in $(y_{k-1}, \dots, y_{k-d}) \in \mathbb{R}^d$ and is strictly positive w.p. 1.

Remark 2.3: Assumption C3) is a sufficient condition for identifiability of φ^* for linear AR models with Markov regime when e_k are normally distributed; see Remark 2.10 below.

Example 2.4 (Linear AR Model With Markov Regime): The fully parameterized linear case with Markov regime may be described by letting $\mathcal{A}$ be the set of $r \times r$ stochastic matrices $\Theta = \{(b_1, \dots, b_d, \sigma) \in \mathbb{R}^d \times (0, \infty)\}$

$$g(y_{n-1}, \dots, y_{n-d}, e_n; \theta) = -b_1\, y_{n-1} - \cdots - b_d y_{n-d} + \sigma e_n \tag{3}$$

and $\Phi \subset \mathcal{A} \times \Theta$ with $a_{ij}(\cdot)$ and $\theta_i(\cdot)$ being the coordinate projections, that is, $a_{ij}(\varphi) = a_{ij}$ and $\theta_i(\varphi) = (b_{i1}, \dots, b_{id}, \sigma_i)$. The innovations $\{e_n\}$ may have, for example, a standard normal distribution, in which case $f(y_n | y_{n-1}, \dots, y_{n-d}; \theta)$ is the density of the normal distribution with mean $-b_1 y_{n-1} - \cdots - b_d y_{n-d}$ and variance σ^2. C1) holds under conditions ii) and iii) of Remark 2.2. C2) is satisfied if the marginal density of e_n is continuous and has bounded derivatives w.r.t. φ. Finally, C3) holds if C1) holds and the marginal density of e_n is positive, continuous, bounded, and has bounded derivatives w.r.t. φ; see the examples provided in [21] and also [6] and [7] for further details.

Example 2.5 (HMM): Using similar notation as in the above example, this is straightforwardly described with $\mathcal{A}$ as above and $\Theta = \{(q, \sigma) \in \mathbb{R} \times (0, \infty)\}$

$$g(e_n; \theta) = q + \sigma e_n$$

where $q = (q_1, \dots, q_r)$ and where q_i, $i = 1, \dots, r$ are often referred to as the "state levels" of the HMM.

B. HMM Prediction Filter

In the sequel, our RMLE algorithm will be based on prediction filters for the state of the Markov chain. For all $n \ge 0$, define the r-dimensional column vector

$$u_n(\varphi) = [u_{n1}(\varphi), \dots, u_{nr}(\varphi)]'$$

where

$$u_{ni}(\varphi) = P_\varphi(X_n = i | y_{n-1}, \dots, y_{-d+1})$$

denotes the predicted density of the Markov chain at time n given observations until time $n-1$. It is straightforward to show that this predicted density can be recursively computed as

$$u_{n+1}(\varphi) = \frac{A'(\varphi)F(z_n; \varphi)u_n(\varphi)}{f'(z_n; \varphi)u_n(\varphi)} = T(z_n, u_n; \varphi) \tag{4}$$

initialized by some $u_0(\varphi)$. The above equation is commonly referred to as the HMM prediction filter, "forward" algorithm, or Baum's equation [34]. Let $\mathcal{P}(S)$ denote the simplex in which $u_n(\varphi)$ resides.

Let $w_n^{(l)}(\varphi) = (\partial/\partial\varphi_l)u_n(\varphi)$ denote the partial derivative of $u_n(\varphi)$ with respect to the lth component of the p-dimensional parameter vector φ. Define the $r \times p$ matrix

$$w_n(\varphi) = ((w_n^{(1)}(\varphi), \ldots, w_n^{(p)}(\varphi)).$$

Clearly, $w_n(\varphi)$ belongs to Ξ defined by

$$\Xi = \{w \in \mathbb{R}^{s \times p} : \mathbf{1}_r' w = \mathbf{0}_p'\}.$$

Differentiating $u_{n+1}(\varphi)$ with respect to φ_l yields

$$\begin{aligned} w_{n+1}^{(l)}(\varphi) =& \frac{\partial u_{n+1}(\varphi)}{\partial \varphi_l} \\ =& R_1\left(z_n,\, u_n(\varphi),\, \varphi\right) w_n^{(l)}(\varphi) + R_2^{(l)}\left(z_n,\, u_n(\varphi),\, \varphi\right) \end{aligned} \tag{5}$$

where

$$\begin{aligned} &R_1\left(z_n, u_n(\varphi), \varphi\right) \\ &\quad = A'(\varphi)\left[I - \frac{F(z_n;\, \varphi)\, u_n(\varphi)\mathbf{1}_r'}{f'(y_n;\, \varphi)\, u_n(\varphi)}\right] \frac{F(z_n;\, \varphi)}{f'(z_n;\, \varphi) u_n(\varphi)} \\ &R_2^{(l)}\left(z_n, u_n(\varphi), \varphi\right) \\ &\quad = A'(\varphi)\left[I - \frac{F(z_n;\, \varphi) u_n(\varphi)\mathbf{1}_r'}{f'(z_n;\, \varphi) u_n(\varphi)}\right] \frac{\partial F(z_n;\, \varphi)/\partial\varphi_l u_n(\varphi)}{f'(z_n;\, \varphi) u_n(\varphi)} \\ &\qquad + \frac{\partial A'(\varphi)/\partial\varphi_l\, F(z_n;\, \varphi)\, u_n(\varphi)}{f'(z_n;\, \varphi) u_n(\varphi)}. \end{aligned}$$

Under the measure P_{φ^*}, the extended Markov chain $\{(X_n,\, Z_n,\, u_n(\varphi),\, w_n(\varphi))\}$ has the transition kernel

$$\begin{aligned} &\Pi^{ij}(z,\, u,\, w,\, d\tilde{z},\, d\tilde{u},\, d\tilde{w}) \\ &\quad = P_{\varphi^*}\left(X_{n+1} = j,\, Z_{n+1} \in d\tilde{z},\, u_{n+1} \in d\tilde{u},\, w_{n+1} \in d\tilde{w}\right. \\ &\qquad \left. | X_n = i,\, Z_n = z,\, u_n = u,\, w_n = w\right). \end{aligned}$$

For any positive integer n, each $i \in S$, and any real-valued function $h = (h^i)$ on $\mathbb{R}^{d+1} \times \mathcal{P}(S) \times \Xi$, define

$$\begin{aligned} &(\Pi^n h)^i(z,\, u,\, w) \\ &\quad = E_{\varphi^*}\left\{h(X_n,\, Z_n,\, u_n,\, w_n) | X_0 = i,\, Z_0 = z,\, u_0 = u,\, w_0 = w\right\}. \end{aligned}$$

Let L denote the set of locally Lipschitz continuous functions $h_\varphi = (h_\varphi^i)$ on $S \times \mathbb{R}^{d+1} \times \mathcal{P}(S) \times \Xi$ in the sense that there exist nonnegative Lip $(h_\varphi^i,\, z)$ and $K(h_\varphi^i,\, z)$ satisfying

$$\begin{aligned} \left|h_\varphi^i(z,\, u,\, w) - h_\varphi^i(z,\, \tilde{u},\, \tilde{w})\right| \le& \mathrm{Lip}(h_\varphi^i,\, z)\left[|w - \tilde{w}| + |u - \tilde{u}|\right. \\ & \left. \cdot\, (1 + |w| + |\tilde{w}|)\right] \\ \left|h_\varphi^i(z,\, u,\, w)\right| \le& K(h_\varphi^i,\, z)(1 + |z|) \end{aligned} \tag{6}$$

for any $u,\, \tilde{u} \in \mathcal{P}(S)$ and any $w,\, \tilde{w} \in \Xi$, such that

$$\begin{aligned} \mathrm{Lip}(h_\varphi) &= \max_{i \in S} \int_{\mathbb{R}^{d+1}} \mathrm{Lip}(h_\varphi^i,\, z) f_i(z,\, \varphi^*) dz < \infty \\ K(h_\varphi) &= \max_{i \in S} \int_{\mathbb{R}^{d+1}} K(h_\varphi^i,\, z) f_i(z,\, \varphi^*) dz < \infty. \end{aligned} \tag{7}$$

We now make the following assumption.

C4) Under P_{φ^*}, the extended Markov chain

$$\{X_n,\, Z_n,\, u_n(\varphi),\, w_n(\varphi)\}$$

that resides in $S \times \mathbb{R}^{d+1} \times \mathcal{P}(S) \times \Xi$ is geometrically ergodic. Consequently, it has a unique invariant probability distribution μ_φ under the measure P_{φ^*}. Thus, for any $\xi \in S \times \mathbb{R}^{d+1} \times \mathcal{P}(S) \times \Xi$ and for any function $h_\varphi = (h_\varphi^i)$ in L, we assume

$$\left|(\Pi_\varphi^n h_\varphi)(\xi) - \lambda\right| \le C\left[\mathrm{Lip}(h_\varphi) + K(h_\varphi)\right] \frac{\rho^n}{1 - \rho} \tag{8}$$

where the constant λ is defined as

$$\lambda = \sum_{i \in S} \int h_\varphi^i(z,\, u,\, w) \mu_\varphi^i(dz,\, du,\, dw).$$

Due to the above geometric ergodicity, the initial values $u_0(\varphi)$ and $w_0(\varphi)$ are forgotten exponentially fast and are hence asymptotically unimportant in our subsequent analysis.

Remark 2.6 (HMM Case): Recall that in this case, $Z_n = Y_n$ and $d = 0$. The geometric ergodicity of

$$\{X_n,\, Z_n,\, u_n(\varphi),\, w_n(\varphi)\}$$

is proved in [27]. We briefly summarize their results here.

Define for any $\iota \ge 0$ and $s = 0,\, 1,\, 2,\, 3$

$$\begin{aligned} \delta^{(s)}(z) &= \sup_{\varphi \in \Phi} \max_{k_1,\, \ldots,\, k_s \in \{1,\, \ldots,\, p\}} \frac{\max_{i \in S} \left|\partial_{k_1,\, \ldots,\, k_s}^s f_i(z;\, \varphi)\right|}{\min_{i \in S} f_i(z;\, \varphi)} \\ \Delta_\iota^{(s)} &= \sup_{\varphi \in \Phi} \max_{i \in S} \int_{\mathbb{R}} \left[\delta^{(s)}(z)\right]^\iota f_i(z;\, \varphi)\, dz \\ \Gamma_\iota &= \sup_{\varphi \in \Phi} \max_{i \in S} \int_{\mathbb{R}} \left[\max_{j \in S} |\log f_j(z;\, \varphi)|\right]^\iota f_i(z;\, \varphi^*)\, dz \\ \overline{Y}_\iota &= \sup_{\varphi \in \Phi} \max_{i \in S} \int_{\mathbb{R}} |z|^\iota f_i(z;\, \varphi)\, dz \end{aligned} \tag{9}$$

where φ^* denotes the true parameter. It is shown in [27] that for a locally Lipschitz function h the followng holds.

A sufficient condition for geometric ergodicity of

$$\{X_n,\, Z_n,\, u_n(\varphi),\, w_n(\varphi)\}$$

is that C1) holds, and the mapping $u \to R_2^{(l)}(z,\, u,\, \varphi)$ is locally Lipschitz for any $z \in \mathbb{R}^{d+1}$ [27, Assumption C] and $\Delta_4^{(0)}$ is finite. A sufficient condition for $u \to R_2^{(l)}(z,\, u,\, \varphi)$ to be locally Lipschitz is $\Delta_2^{(0)}$ is finite (see [27, Example 4.3]). Note that if the noise density $f_i(z;\, \varphi)$, $i \in S$, is Gaussian, then $\Delta_\iota^{(s)}$, Γ_ι, and $\overline{Y}_\iota$ are finite for $\iota \ge 0$.

Remark 2.7 (AR Model With Markov Regime): The above conditions on Δ_ι do not directly apply to AR models with Markov regime. In [13], weaker sufficient conditions are given for the exponential forgetting of $u_n(\varphi)$. We summarize this result and outline how it can be used to show geometric ergodicity of $(X_n,\, Z_n,\, u_n(\varphi),\, w_n(\varphi))$.

As in [13, Assumption A1], assume in addition to C1) that for $y \in \mathbb{R}$, $\xi \in \mathbb{R}^d$, and $z = (y,\, \xi)$

$$0 < \min_i \inf_{\varphi \in \Phi} f_i(z;\, \varphi) \quad \text{and} \quad \max_i \sup_{\varphi \in \Phi} f_i(z;\, \varphi) < \infty.$$

Suppose $u_n(\varphi)$ and $\tilde{u}_n(\varphi)$ are predictors with initial conditions $u_0(\phi)$ and $\tilde{u}_0(\varphi)$, respectively. Then it is proved in [13, Corollary 1] that

$$|u_n(\varphi)-\tilde{u}_n(\varphi)| < \rho^n, \quad \rho \overset{\text{def}}{=} 1-\frac{\inf_{\varphi\in\Phi}\min_{i,j} a_{ij}(\varphi)}{\sup_{\varphi\in\Phi}\max_{i,j} a_{ij}(\varphi)}. \qquad (10)$$

Compared with the results of [27], ρ is observation independent. As a consequence, starting from (10) one can obtain the geometric ergodicity of $(X_n, Z_n, u_n(\varphi), w_n(\varphi))$ along the lines of [27, Secs. 3 and 5] as follows.

- By exactly the same steps as the proof of [27, Proposition 3.8], one can establish that for any locally Lipschitz function $h = (h)^i$, $i \in S$

$$\begin{aligned}\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} &\left|h^{i_{n+1}}\left(z_{n+1}, u_n(\varphi)\right) - h^{i_{n+1}}\left(z_{n+1}, \tilde{u}_n(\varphi)\right)\right| \\ &\cdot f_{i_m}(z_m;\varphi^*)\cdots f_{i_{n+1}}(z_{n+1};\varphi^*)\,dy_m\cdots dy_{n+1} \\ \le \mathrm{Lip}(h)&\int_{\mathbb{R}}\cdots\int_{\mathbb{R}} |u_n(\varphi)-\tilde{u}_n(\varphi)| \\ &\cdot f_{i_m}(z_m;\varphi^*)\cdots f_{i_{n+1}}(z_{n+1};\varphi^*)\,dy_m\cdots dy_{n+1}. \end{aligned} \qquad (11)$$

  In [27], the exponent $\delta(y)$ for the exponential forgetting of $|u_n(\varphi)-\tilde{u}_n(\varphi)|$ depends on the observations—hence integrability conditions such as $\Delta_1 < \infty$ are required. In comparison with the exponential forgetting (10), it follows from (11) that [27, Proposition 3.8] holds.
- The proof of geometric ergodicity of $(X_n, Z_n, u_n(\varphi))$ then follows along the lines of [27, Theorem 3.6], but the argument is much simpler because ρ is observation independent.
- As in [27], assume that $R_2(z, u(\varphi), \varphi)$ is Lipschitz continuous in u. Assuming C1) and (10), the geometric ergodicity of $(X_n, Z_n, u_n(\varphi), w_n(\varphi))$ follows along the same lines as [27, Sects. 4 and 5]. In particular, for $z = (y, \xi)$ where $\xi \in \mathbb{R}^d$, define $\bar{\delta}^{(0)}(z) = \max_i \sup_\varphi f_i(z;\varphi)$ and $\bar{\delta}^{(1)}(z) = \max_i \sup_\varphi \|\partial_\varphi f_i(z;\varphi)\|$. Then the integrability conditions

$$\max_i \sup_\varphi \int_{\mathbb{R}} \left[\bar{\delta}^{(0)}(z)\right]^2 f_i(z;\varphi)\,dy < \infty$$

  and

$$\max_i \sup_\varphi \int_{\mathbb{R}} \left[\bar{\delta}^{(1)}(z)\right]^2 f_i(z;\varphi)\,dy < \infty$$

  for all $\xi \in \mathbb{R}^d$ are sufficient for geometric ergodicity of $(X_n, Z_n, u_n(\varphi), w_n(\varphi))$.

C. Kullback–Leibler Information

The conditional log-likelihood function (suitably normalized) based on the observations $y_{-d+1}, \ldots, y_n$ is

$$l_n(\varphi) = \frac{1}{n+1}\log p_n\left(y_1, \ldots, y_n | y_{-d+1}, \ldots, y_0; \varphi\right).$$

It is straightforward to show that the conditional log likelihood can be expressed as the sum of terms involving the observations and the *observed state* (prediction filter) as follows:

$$l_n(\varphi) = \frac{1}{n+1}\sum_{k=0}^{n}\log\left[f'(z_k;\varphi)u_k(\varphi)\right]. \qquad (12)$$

For the HMM case, under assumptions C1) and C4), [27, Example 3.4] shows that the $\log(f'(z_k;\varphi)u_k(\varphi)$ is locally Lipschitz if $\Delta_1^{(0)}$ and Γ_1 are finite. This and the geometric ergodicity yield that the following strong law of large numbers holds.

Proposition 2.8 (HMM Case): Under assumptions C1), C2), and C4), if $\Delta_1^{(0)}$ and Γ_1 are finite, then for any $\varphi \in \Phi$ there exists a finite $l(\varphi)$ such that

$$l_n(\varphi) \to l(\varphi), \qquad P_{\varphi^*} \text{ w.p. 1 as } n \to \infty$$

where

$$l(\varphi) = \int_{\mathbb{R}\times\mathcal{P}(S)} \log\left[f'(z;\varphi)u_n(\varphi)\right]\nu_\varphi(dz, du)$$

and ν_φ denotes the marginal density of the invariant measure μ_φ defined on $\mathbb{R}\times\mathcal{P}(S)$.

For the AR case with Markov regime the following strong law of large numbers holds—see [13, Proposition 1] for proof.

Proposition 2.9 (AR With Markov Regime): Under C1), C2), and C4), with $z = (y, \xi)$ if $\sup_\varphi \sup_{\xi, y} f_i(z;\varphi) < \infty$ and $E_{\varphi^*}(\log(\inf_\varphi f_i(z;\varphi))) < \infty$ for all $i \in S$ then for any $\varphi \in \Phi$ there exists a finite $l(\varphi)$ such that

$$l_n(\varphi) \to l(\varphi), \qquad P_{\varphi^*} \text{ w.p. 1 as } n \to \infty$$

where

$$l(\varphi) = \int_{\mathbb{R}^{d+1}\times\mathcal{P}(S)} \log\left[f'(z;\varphi)u(\varphi)\right]\nu_\varphi(dz, du)$$

and ν_φ denotes the marginal density of the invariant measure μ_φ defined on $\mathbb{R}^{d+1}\times\mathcal{P}(S)$.

Recall that φ^* is the true parameter that we are seeking. Under C1)–C4), define for any $\varphi \in \Phi$ the Kullback–Leibler information as

$$K(\varphi) = -\left[l(\varphi) - l(\varphi^*)\right] \ge 0.$$

We have proved in [21] that φ^* belongs to the set L_{ML} of global minima of $K(\varphi)$

$$L_{\mathrm{ML}} = \arg\min_{\varphi\in\Phi} K(\varphi). \qquad (13)$$

In addition, the ML estimator (MLE)

$$\hat{\varphi}_{\mathrm{ML}} = \arg\max_{\varphi\in\Phi} l_n(\varphi)$$

is strongly consistent.

Remark 2.10 (Identifiability in Linear AR Case): Consider the linear AR process with Markov regime of Example 2.4. Assume e_k is normally distributed. Assume that the true model vectors $\{(b_{i1}^*, \ldots, b_{id}^*, \sigma_i^*)\}_{i=1}^r$ are distinct, so that for each n, there exists a point $(y_{n-1}, \ldots, y_{n-d}) \in \mathbb{R}^d$ such that $\{(b_{i1}^* y_{n-1} + \cdots + b_{id}^* y_{n-d}, \sigma_i^*)\}_{i=1}^r$ are distinct. Then using C3, it is proved in [21, Example 3] that φ^* is uniquely identifiable, in the sense that $K(\varphi) = 0$ implies that $\varphi = \varphi^*$ up to a permutation of indices.

D. RMLE Algorithm

To estimate φ^*, one can search for the minima of the Kullback–Leibler divergence $K(\varphi)$. Assuming the function $K(\cdot)$ is sufficiently smooth, the parameter estimation problem is converted to finding the zeros of $(\partial/\partial\varphi)K(\cdot)$. In this paper, we use a recursive algorithm of stochastic approximation type to carry out the task.

Recalling that the symbol $'$ denotes transpose and differentiating the terms within the summation in (9) with respect to φ_l yields the p-dimensional "incremental score vector"

$$S(\tilde{Y}_n;\,\varphi)=\left(S^{(1)}(\tilde{Y}_n;\,\varphi),\,\ldots,\,S^{(p)}(\tilde{Y}_n;\,\varphi)\right)'$$

with

$$S^{(l)}(\tilde{Y}_n;\,\varphi)=\frac{f'(z_n;\,\varphi)w_n^{(l)}(\varphi)}{f'(z_n;\,\varphi)u_n(\varphi)}+\frac{[(\partial/\partial\varphi_l)f'(z_n;\,\varphi)]\,u_n(\varphi)}{f'(z_n;\,\varphi)u_n(\varphi)} \tag{14}$$

where

$$\tilde{Y}_n \overset{\text{def}}{=} (Z_n,\,u_n(\varphi),\,w_n(\varphi)) \tag{15}$$

with u_n and w_n defined by (4) and (5), respectively. The RMLE algorithm takes the form

$$\varphi_{n+1}=\Pi_G\left(\varphi_n+\varepsilon_n S(\tilde{Y}_n;\,\varphi_n)\right). \tag{16}$$

In (13), $\{\varepsilon_n\}$ is a sequence of step sizes satisfying $0\le\varepsilon_n\to 0$ and $\sum_n \varepsilon_n=\infty$, G is a convex and compact set, and Π_G denotes the projection of the estimate to the set G. More precise conditions will be given later. Note that in (13), following the usual approach in stochastic approximation, we have collected $(Z_n,\,u_n,\,w_n)$ in $\tilde{Y}_n$. This enables us to treat $\tilde{Y}_n$ as a noise process. Our task to follow is to analyze the asymptotic properties of (13). Moreover, we also examine its variant algorithms.

III. Asymptotic Properties

The objective of this section is to analyze the convergence and rate of convergence of the RMLE algorithm proposed in the previous section. In what follows, we use the results in [25] whenever possible with appropriate references noted. For the convergence analysis, we use the ordinary differential equation (ODE) approach that relates the discrete-time iterations of the RMLE algorithm to an ODE. For rate of convergence, we present a weak convergence analysis to examine the dependence of the estimation error $(\varphi_n-\varphi^*)$ on the step size ε_n. We answer the question for what real number α, $\varepsilon_n^{\alpha}(\varphi_n-\varphi^*)$ converges to a nontrivial limit.

Note that our formulation and results are more general than that of Rydén [35], [36] because we are dealing with suitably scaled sequences of the iterates that are treated as stochastic processes rather than random variables. Our approach captures the dynamic evolution of the RMLE algorithm. As a consequence, we can analyze the tracking properties of the RMLE algorithms when the parameters are time varying, which is done in Section V.

A. Preliminaries

First rewrite the first equation in (13) as

$$\varphi_{n+1}=\varphi_n+\varepsilon_n S(\tilde{Y}_n;\,\varphi_n)+\varepsilon_n M_n \tag{17}$$

where M_n is the projection or correction term, i.e., it is the vector of shortest Euclidean length needed to bring $\varphi_n+\varepsilon_n S(\tilde{Y}_n;\,\varphi_n)$ back to the constraint set G if it ever escapes from G (see [25, p. 89] for more discussion). For future use, denote by $\mathcal{F}_n$ the σ algebra generated by $\{\varphi_0,\,\tilde{Y}_j,\,j<n\}$, and let E_n denote the conditional expectation with respect to $\mathcal{F}_n$.

Constraint Set: Let $q_i(\cdot)$, $i=1,\,\ldots,\,p$, be continuously differentiable real-valued functions on $\mathbb{R}^p$. Without loss of generality, let $(\partial/\partial\varphi)q_i(\varphi)\ne 0$ if $q_i(\varphi)=0$. Let the constraint set be

$$G=\{\varphi;\,q_i(\varphi)\le 0,\,i=1,\,\ldots,\,p\}$$

and assume it is connected, compact, and nonempty. A constraint $q_i(\cdot)$ is active at φ if $q_i(\varphi)=0$. Define $\mathcal{A}(\varphi)$, the set of indexes of the active constraints at φ, by $\mathcal{A}(\varphi)=\{i;\,q_i(\varphi)=0\}$. Define $C(\varphi)$ to be the convex cone generated by the set of outward normals $\{\psi;\,\psi=(\partial/\partial\varphi)q_i(\varphi),\,i\in\mathcal{A}(\varphi)\}$. Suppose for each φ, $\{(\partial/\partial\varphi)q_i(\varphi),\,i\in\mathcal{A}(\varphi),\,\mathcal{A}(\varphi)\ne\emptyset\}$ is linearly independent. If $q_i(\varphi)\ne 0$ for all i, then $C(\varphi)$ contains only the zero element.

To prove the convergence of the algorithm, we use the ODE approach (see Kushner and Clark [22]); the following development follows the framework setup in [25]). Take a piecewise-constant interpolation of φ_n as follows. Define $t_0=0$ and $t_n=\sum_{i=0}^{n-1}\varepsilon_i$, and

$$m(t)=\begin{cases}\text{unique } n;\ t_n\le t<t_{n+1}, & \text{for } t\ge 0\\ 0, & \text{for } t<0.\end{cases}$$

Let

$$\varphi^0(t)=\begin{cases}\varphi_0, & \text{for } t\le 0\\ \varphi_n, & \text{for } t_n\le t<t_{n+1},\ \text{for } t\ge 0.\end{cases}$$

Define the sequence of shifted process $\varphi^n(\cdot)$ by

$$\varphi^n(t)=\varphi^0(t_n+t),\qquad \text{for } t\in(-\infty,\,\infty).$$

Define $M^0(\cdot)$ and $M^n(\cdot)$ by

$$M^0(t)=\begin{cases}\sum\limits_{i=0}^{m(t)-1}\varepsilon_i M_i, & t\ge 0\\ 0, & \text{for } t<0\end{cases}$$

and

$$M^n(t)=\begin{cases}M^0(t_n+t)-M^0(t), & \text{for } t\ge 0\\ -\sum\limits_{i=m(t_n+t)}^{n-1}\varepsilon_i M_i, & t<0.\end{cases}$$

Using such interpolations, one then aims to show $\{\varphi^n(\cdot),M^n(\cdot)\}$ is equicontinuous in the extended sense [25, p. 73] and uniformly bounded. By the Ascoli–Arzelá theorem, we can extract a convergent subsequence such that its limit satisfies a projected ODE, which is one whose dynamics are projected onto the constraint set G.

Projected ODE: Consider the projected ODE

$$\dot{\varphi}=H(\varphi)+\tilde{m},\qquad \varphi(0)=\varphi_0,\ \tilde{m}\in -C(\varphi) \tag{18}$$

where $H(\varphi)=(\partial/\partial\varphi)K(\varphi)$, and $\tilde{m}(\cdot)$ is the projection or constraint term. The term $\tilde{m}(\cdot)$ is the minimum force needed to keep $\varphi(\cdot)\in G$. Let $L_G=\{\varphi;\,\varphi$ be a limit point of (15), $\varphi_0\in G\}\}$, and $\hat{L}_G=\{\varphi\in G;\,H(\varphi)+\tilde{m}=0\}$. The points in $\hat{L}_G$ are termed stationary points. When $\varphi\in G^0$, the interior of G, the stationary condition is $(\partial/\partial\varphi)K(\varphi)=H(\varphi)=0$, and when $\varphi\in\partial G$, the boundary of G, $H(\varphi)\in C(\varphi)$. For more discussion on projected ODE, see [25, Sec. 4.3] for details. A set

$\hat{A} \subset G$ is locally asymptotically stable in the sense of Liapunov for (15), if for each $\delta > 0$, there is a $\delta_1 > 0$ such that all trajectories starting in $N_{\delta_1}(\hat{A})$ never leave $N_\delta(\hat{A})$ and ultimately stay in $N_{\delta_1}(\hat{A})$, where $N_\eta(\hat{A})$ denotes an η neighborhood of $\hat{A}$.

B. Convergence

Assume the following conditions are satisfied.

A1) Conditions C1)–C4) hold.

A2) For each $\varphi \in G$, $\{S(\tilde{Y}_j;\, \varphi)\}$ is uniformly integrable, $ES(\tilde{Y}_j;\, \varphi) = H(\varphi) = (\partial/\partial\varphi)K(\varphi)$, $H(\cdot)$ is continuous, and $S(\tilde{Y};\cdot)$ is continuous for each $\tilde{Y}$. There exist nonnegative measurable functions $\tilde{\rho}(\cdot)$ and $\hat{\rho}(\cdot)$ such that $\tilde{\rho}(\cdot)$ is bounded on bounded φ set, and

$$\left|S(\tilde{Y};\, \varphi) - S(\tilde{Y};\, \psi)\right| \le \tilde{\rho}(\varphi - \psi)\hat{\rho}(\tilde{Y})$$

such that $\tilde{\rho}(\varphi) \to 0$ as $\varphi \to 0$ and

$$P\left(\limsup_n \sum_{i=n}^{m(t_n+s)} \varepsilon_i \hat{\rho}(\tilde{Y}_i) < \infty\right) = 1, \quad \text{for some } s > 0.$$

In the above, the expectation is taken w.r.t. the φ-parameterized stationary distribution.

A3) Suppose that L_G^1 is a subset of L_G and L_{ML} is locally asymptotically stable. For any initial condition $\varphi_0 \notin L_G^1$, the trajectories of (15) goes to L_{ML}.

Remark 3.1: For the HMM case, A2) holds if the marginal density of e_k is Gaussian. A sufficient condition for the uniform integrability and Lipschitz continuity in A2) is that $\Delta_2^{(1)}$, $\overline{Y}_2$, and Γ_2 in (8) are finite; see [29].

Consider the AR case with Markov regime: A2) is easily verifiable for the AR(1) linear case (i.e., $d = 1$ in (3))

$$Y_{k+1} = -b_{X_{k+1}} Y_k + e_k, \qquad \text{where } |b_{i1}| < 1,\ i = 1, 2 \ldots, r.$$

Suppose $\{e_k\}$ is a sequence of i.i.d. Gaussian random variables with zero mean and finite variance σ^2. It is easily seen that for each $(Y_k,\, Y_{k-1}) \in \mathbb{R}^2$, $f(z_k;\cdot)$ is continuously differentiable w.r.t. θ with bounded derivatives and hence it is Lipschitz continuous. It is also clear that the Lipschitz constant depends on $(Y_k,\, Y_{k-1})$. Thus, by using (4), (5), and (11), A2) is verified. Higher order linear AR models with Markov regime (i.e., $d > 1$) can be treated in a similar way with more complex notation.

Regarding the uniform integrability, suppose that for each $\varphi \in G$, $E|S(\tilde{Y}_j;\, \varphi)|^{1+\Delta} < \infty$ for some $\Delta > 0$. Then the uniform integrability is verified. If $\{S(\tilde{Y}_j;\, \varphi)\}$ is bounded by an integrable random variable $\tilde{U}$ in the sense

$$P(|S(\tilde{Y}_j;\, \varphi)| \ge \alpha) \le P(|\tilde{U}| \ge \alpha)$$

then $\{S(\tilde{Y}_j;\, \varphi)\}$ is also uniformly integrable. More specifically, if h satisfies the condition (6) with $K(h^i,\, z)$ verifies

$$\max_{i\in S} \int_{\mathbb{R}^{d+1}} \left[K(h^i,\, z)\right]^2 f_i(z,\, \varphi^*)\, dz < \infty. \tag{19}$$

in lieu of (7), and $\{Z_n\}$ given by (1) is uniformly integrable, then the desired uniform integrability can be verified via the use of Cauchy–Schwarz inequality. In [13, Theorem 2 and Lemma 10] sufficient conditions are given for $E|S(\tilde{Y}_j;\, \varphi)|^2 < \infty$ for the AR case with Markov regime. Such conditions also guarantee the uniform integrability of $\{S(\tilde{Y}_j;\, \varphi)\}$.

Lemma 3.2: Under the conditions A1) and A2), for each φ, each $\mu > 0$, and some $T > 0$

$$\lim_{n\to\infty} P \left\{ \sup_{j\ge n} \max_{0\le t\le T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \varepsilon_i \left(S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right) \right| \ge \mu \right\} = 0. \tag{20}$$

Remark 3.3: To prove the consistency of stochastic approximation algorithms, a crucial step is to verify that condition (17) holds. Such conditions were first brought in by Kushner and Clark in [22]; it is summarized in the current form and referred to as "asymptotic rate of change" in [25, Secs. 5.3 and 6.1]. These conditions appear to be close to the minimal requirement needed, and have been proved to be necessary and sufficient condition in certain cases [42]. To verify this condition, we use the idea of perturbed state or perturbed test function methods. Note that our conditions are weaker than that of [35]. Only finite first moment is needed.

Proof of Lemma 3.2: We use a discounted perturbation. The use of perturbed test function for stochastic approximation was initiated by Kushner, and the discounted modification was suggested in Solo and Kong [39]. For future use, define

$$B_{in} = \begin{cases} \prod_{k=n}^{i} (1 - \varepsilon_k), & i \ge n \\ 1, & i < n. \end{cases}$$

For each φ, define $v_n(\varphi)$ as

$$\begin{aligned} v_0(\varphi) &= 0 \\ v_{n+1}(\varphi) &= v_n(\varphi) + \varepsilon_n \left(S(\tilde{Y}_n;\, \varphi) - H(\varphi)\right), \qquad n \ge 0 \\ \delta v_n(\varphi) &= \sum_{i=n}^{\infty} \varepsilon_i B_{i(n+1)} E_n \left(S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right) \\ \tilde{v}_n(\varphi) &= v_n(\varphi) + \delta v_n(\varphi) \end{aligned}$$

and

$$\begin{aligned} \delta N_n(\varphi) = {} & \sum_{i=n+1}^{\infty} \varepsilon_i B_{i(n+2)} E_{n+1} \left(S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right) \\ & - \sum_{i=n+1}^{\infty} \varepsilon_i B_{i(n+2)} E_n \left(S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right). \end{aligned}$$

Then by noting that $B_{i(n+2)} - B_{i(n+1)} = \varepsilon_{n+1} B_{i(n+2)}$

$$\begin{aligned} & \tilde{v}_{n+1}(\varphi) - \tilde{v}_n(\varphi) \\ & \quad = (v_{n+1}(\varphi) - v_n(\varphi)) + (\delta v_{n+1}(\varphi) - \delta v_n(\varphi)) \\ & \quad = \varepsilon_n \left(S(\tilde{Y}_n;\, \varphi) - H(\varphi)\right) \\ & \qquad + \sum_{i=n+1}^{\infty} \varepsilon_i B_{i(n+2)} E_{n+1} \left(S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right) \\ & \qquad - \sum_{i=n}^{\infty} \varepsilon_i B_{i(n+1)} E_n \left(S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right) \\ & \quad = \delta N_n(\varphi) + \varepsilon_{n+1} E_n \delta v_{n+1}(\varphi). \end{aligned}$$

Note that by telescoping

$$\sup_n \sum_{i=n}^{\infty} \varepsilon_{i+1} B_{i(n+1)} = \sup_n \sum_{i=n}^{\infty} \left(B_{in} - B_{i(n+1)}\right) < \infty.$$

It yields that

$$\sup_n \sum_{i=n}^{\infty} \varepsilon_i B_{i(n+1)} = \sup_n \sum_{i=n}^{\infty} \frac{\varepsilon_i - \varepsilon_{i+1}}{\varepsilon_{i+1}}\, \varepsilon_{i+1} B_{i(n+1)} + \sup_n \sum_{i=n}^{\infty} \varepsilon_{i+1} B_{i(n+1)} < \infty.$$

Therefore,

$$\sum_{i=n}^{\infty} \varepsilon_i B_{in} \to 0 \text{ as } n \to \infty.$$

Owing to A2)

$$\sum_{i=0}^{\infty} \varepsilon_i B_{i1} < \infty \quad \text{and} \quad \sum_{i=0}^{\infty} \varepsilon_i B_{i1} E\left|S(\tilde{Y}_i;\, \varphi)\right| < \infty$$

so

$$\sum_{i=0}^{\infty} \varepsilon_i B_{i1} S(\tilde{Y}_i;\, \varphi) \text{ converges w.p. 1}$$

and

$$\sum_{i=n}^{\infty} \varepsilon_i B_{i(n+1)} \left|S(\tilde{Y}_i;\, \varphi)\right| \to 0 \text{ w.p. 1}.$$

Similarly,

$$\sum_{i=0}^{\infty} \varepsilon_i B_{i1} E\left|S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right| < \infty$$

$$\sum_{i=n}^{\infty} \varepsilon_i B_{i(n+1)} \left|S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right| \to 0 \text{ w.p. } 1$$

and

$$\sum_{i=n}^{\infty} \varepsilon_i B_{i(n+1)} \left|E_n S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right| \to 0 \text{ w.p. } 1.$$

Consequently, $\delta v_n(\varphi) \to 0$ w.p. 1 as $n \to \infty$. Likewise,

$$\sum_{i=n+1}^{\infty} \varepsilon_i B_{i(n+2)} E\left|E_{n+1} S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right| + \sum_{i=n+1}^{\infty} \varepsilon_i B_{i(n+2)} E\left|E_n S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right| < \infty$$

and hence $\delta N_n(\varphi) \to 0$ w.p. 1 as $n \to \infty$. As a result,

$$\varepsilon_{n+1} E_n \delta v_{n+1}(\varphi) \to 0 \text{ w.p. 1}$$

and

$$\tilde{v}_{n+1}(\varphi) - \tilde{v}_n(\varphi) \to 0 \text{ w.p. 1}.$$

Therefore, the asymptotic rate of change of

$$\sum_{i=0}^{m(t)-1} \varepsilon_i \left(S(\tilde{Y}_i;\, \varphi) - H(\varphi)\right) \text{ is 0 w.p. 1 as } t \to \infty.$$

The proof of the lemma is concluded. □

Theorem 3.4: Assume conditions A1) and A2). There is a null set $\tilde{N}$ such that for all $\omega \notin \tilde{N}$, $\{\varphi^n(\omega,\, \cdot),\, M^n(\omega,\, \cdot)\}$ is equicontinuous (in the extended sense as in [25, p. 73]). Let $(\varphi(\omega,\, \cdot),\, M(\omega,\, \cdot))$ denote the limit of some convergent subsequence. Then the pair satisfies the projected ODE (15), and $\{\varphi_n\}$ converges to an invariant set of the ODE in G.

Assume A3). Then the limit points are in $L_G^1 \cup \hat{A}_G$ w.p. 1. If, in particular, $L_G^1 \cup \hat{A}_G = \{\varphi^*\}$, and φ_n visit $L_G^1 \cup \hat{A}_G$ infinitely often w.p. 1, then $\varphi_n \to \varphi^*$ w.p. 1.

Proof: The proof follows from Lemma 3.2, [25, Theorems 6.1.1 and 5.2.2]. □

Remark 3.5: In view of [22, Theorem 5.3.1], the set of stationary points of (15) is the set of Kuhn–Tucker points

$$\text{KT} = \left\{\varphi;\ \text{there exist } \lambda_i \geq 0 \text{ such that } -H(\varphi) + \sum_{i;\, q_i(\varphi)=0} \lambda_i \frac{\partial q_i(\varphi)}{\partial \varphi} = 0\right\}.$$

As observed in [21], for linear AR processes with Markov regime, the only global minima of $K(\varphi)$ are φ^* and possibly also parameters equal to φ^* up to a permutation of states.

C. Rate of Convergence

Since our main concern here is the convergence rate, we assume that $\varphi^* \in G^0$, the interior of G, and that φ_n converges to φ^* w.p. 1. Suppose the following conditions hold.

A4) $(\varepsilon_n/\varepsilon_{n+1})^{1/2} = 1 + \mu_n$ where either
  a) $\mu_n = 1/(2n) + o(\varepsilon_n)$ or
  b) $\mu_n = o(\varepsilon_n)$.

A5) For each $\varphi \in G$, $E|S(\tilde{Y}_j;\, \varphi)|^{2+\Delta} < \infty$ for some $\Delta > 0$ and $\{(\partial/\partial\varphi) S(\tilde{Y}_j;\, \varphi)\}$ is uniformly integrable.

A6) $\varphi_n \to \varphi^*$ w.p. 1 and $\{(\varphi_n - \varphi^*)/\sqrt{\varepsilon_n}\}$ is tight.

A7) a) $S(\tilde{Y};\cdot)$ has continuous partial derivatives for each $\tilde{Y}$, $H(\cdot)$ is continuously differentiable, and $H_\varphi(\varphi^*) = (\partial/\partial\varphi) H(\varphi^*)$ is Hurwitz (i.e., all of its eigenvalues have negative real parts).
  b) If A4) a) holds (in this case, $\varepsilon_n = O(1/n)$), then $H_\varphi(\varphi^*) + I/2$ is also Hurwitz.

A8) Denote $S_i = S(\tilde{Y}_i;\, \varphi^*)$. For $i \geq j \geq n$, define $r_n(i-j) = E_n S_i' S_j$ and $R(i-j) = |E r_n(i-j)|$ such that

$$\sum_{i \geq j} R(i-j) < \infty.$$

Remark 3.6: Assumption A4) is a condition on the step size. Strictly speaking, it is not an assumption since the step size is at our disposal. Typical examples include $\varepsilon_n = A_0/n$ for some $A_0 > 0$, which satisfies a) in A4), and $\varepsilon_n = A_0/n^\gamma$ for some $0 < \gamma < 1$, which satisfies b) in A4). It also covers a wide variety of other cases.

In the HMM case, a sufficient condition for A5) to hold is that $\overline{Y}_2$, $\Delta_2^{(1)}$, $\Delta_2^{(2)}$, and $\Delta_2^{(3)}$ are finite, see [29, Assumption B′]. These hold for example, when $\{e_n\}$ is a sequence of Gaussian noises.

Condition A5) can be verified for linear Gaussian autoregressive processes with Markov regime. Consider the AR(1) case $Y_{n+1} = -b_{X_{n+1}} Y_n + e_n$, where the meaning and conditions of the parameters are as in Remark 3.1. In view of the discussion in Remark 2.4, it is easily seen that

$$\frac{\partial f(z_n;\,\theta)}{\partial b_1} = -\frac{y_{n-1}}{\sqrt{2\pi}\sigma^3}(y_n+b_1y_{n-1})\exp\left(-\frac{(y_n+b_1y_{n-1})^2}{2\sigma^2}\right)$$

and

$$\frac{\partial f(z_n;\,\theta)}{\partial \sigma} = \frac{1}{\sqrt{2\pi}\sigma^4}(y_n+b_1y_{n-1})^2\exp\left(-\frac{(y_n+b_1y_{n-1})^2}{2\sigma^2}\right).$$

Since a normal distribution has finite moments of any order

$$E\left|\frac{\partial f(z_n;\,\theta)}{\partial b_1}\right|^{2+\Delta} < \infty \quad \text{and} \quad E\left|\frac{\partial f(z_n;\,\theta)}{\partial \sigma}\right|^{2+\Delta} < \infty.$$

In view of (4), (5), and (11), $E|S(\tilde{Y}_n;\,\varphi)|^{2+\Delta} < \infty$. Higher order linear AR models with Markov regime (i.e., $d > 1$) can be treated in a similar way with more complex notation. The moment condition is needed in functional central limit theorem; see [5], [14]. If the noise $\tilde{Y}_n$ has moment generating function, then all moments are finite. In the Gaussian case, it is characterized by the first two moments.

Again, we can supply sufficient conditions ensuring the uniform integrability. For example, as in the discussion in Remark 3.1, in view of (6), if $\{Z_n\}$ is uniformly integrable and $K(h^i,\,z)$ verifies

$$\max_{i\in S}\int_{\mathbb{R}^{d+1}}\left[K(h^i,\,z)\right]^{(2+\Delta)/(1+\Delta)} f_i(z,\,\varphi^*)dz < \infty \tag{21}$$

in lieu of (16), the uniform integrability can be verified by use of Hölder inequality.

Condition A7) ensures the limit stochastic differential equation (22) is asymptotically stable. That is, R is a stable matrix (see the definition of R in Theorem 3.9). Such a stability is necessary for the rate of convergence study; see the discussion after Theorem 3.9, in particular, the asymptotic covariance representation (23).

The smoothness condition of $H(\varphi)$ is used for convenience only. Aiming to obtaining local limit result, the only requirement is that $H(\varphi)$ is locally linearizable by a stable matrix. The smoothness assumption can be replaced by

$$H(\varphi) = \tilde{H}(\varphi-\varphi^*) + o(|\varphi-\varphi^*|) \tag{22}$$

where $\tilde{H}$ is a stable matrix (the real parts of its eigenvalues are all negative). Under (19), all the subsequent development goes through with $H_\varphi(\varphi^*)$ replaced by $\tilde{H}$. Note also that the form of (19) is a standard condition used in stochastic approximation. Finally, [13] provides a central limit theorem for the score vector.

Condition A8) simply says that the correlation decays sufficiently fast. It is shown in [27, Example 5.2] that for the HMM case, if $\Delta_2^{(0)}$ defined in (8) is finite, then $S(\tilde{Y};\,\varphi)$ is locally Lipschitz. As a result, $S(\tilde{Y}_j;\,\varphi)$ is geometrically ergodic. For the AR case if $S(\tilde{Y};\,\varphi)$ is locally Lipschitz in $\tilde{Y}$, i.e., (6) holds then $S(\tilde{Y}_j;\,\varphi)$ is geometrically ergodic and satisfies (8). For the linear Gaussian AR case with Markov regime, A8) is easily verifiable; see also the remarks about A2). The condition we pro-pose models that of a mixing process. As indicated in [5, p. 168], for example, the mixing condition is satisfied for a Markov chain that verifies a Doeblin condition, has one ergodic class, and is aperiodic; the condition is also satisfied for certain functions of mixing processes.

The tightness of the rescaled sequence can be verified by using a perturbed Liapunov function method. Sufficient conditions can be given. In fact, in Section IV, we will prove such an assertion. We assume this condition here for simplicity.

By the smoothness of $S(\cdot)$, a Taylor expansion leads to

$$\begin{aligned} S(\tilde{Y};\,\varphi) = {}& S(\tilde{Y};\,\varphi^*) + \frac{\partial}{\partial\varphi}S(\tilde{Y};\,\varphi^*)(\varphi-\varphi^*) \\ & + \int_0^1 \frac{\partial}{\partial\varphi}S\left(\tilde{Y};\,\varphi^*+s(\varphi-\varphi^*)\right)ds(\varphi-\varphi^*) \end{aligned}$$

and

$$\begin{aligned} \varphi_{n+1} = {}& \varphi_n + \varepsilon_n S(\tilde{Y}_n;\,\varphi^*) + \varepsilon_n\frac{\partial S(\tilde{Y}_n;\,\varphi^*)}{\partial\varphi}(\varphi_n-\varphi^*) \\ & + \varepsilon_n\int_0^1\frac{\partial S\left(\tilde{Y};\,\varphi^*+s(\varphi_n-\varphi^*)\right)}{\partial\varphi}ds(\varphi_n-\varphi^*) \\ & + \varepsilon_n M_n. \end{aligned}$$

Since $\varphi_n \to \varphi^*$ w.p. 1 and $\varphi^* \in G^0$, the reflection term can be effectively dropped for the consideration of rate of convergence; we do so henceforth. Using $U_n = (\varphi_n-\varphi^*)/\sqrt{\varepsilon_n}$, we obtain

$$\begin{aligned} U_{n+1} = {}& \sqrt{\frac{\varepsilon_n}{\varepsilon_{n+1}}}U_n + \sqrt{\frac{\varepsilon_n}{\varepsilon_{n+1}}}\Bigg(\varepsilon_n\frac{\partial}{\partial\varphi}S(\tilde{Y}_n;\,\varphi^*)U_n \\ & + \sqrt{\varepsilon_n}\left(S(\tilde{Y}_n;\,\varphi^*)-H(\varphi^*)\right)+\varepsilon_n\rho_n\Bigg) \end{aligned} \tag{23}$$

where $\rho_n = o(|U_n|)$ in probability due by use of the Taylor expansion and the uniform integrability of $(\partial/\partial\varphi)S(\tilde{Y}_n,\,\varphi)$. Now define

$$\begin{aligned} W^n(t) &= \sum_{i=n}^{m(t_n+t)-1}\sqrt{\varepsilon_i}S_i, \qquad t \geq 0 \\ &= -\sum_{i=m(t_n+t)}^{n-1}\sqrt{\varepsilon_i}S_i, \qquad t < 0 \end{aligned}$$

where $S_i = S(\tilde{Y}_i;\,\varphi^*) - H(\varphi^*)$. Let $U^n(\cdot)$ be the piecewise-constant interpolation of $\{U_i,\,i \geq n\}$ on $[0,\infty)$.

Lemma 3.7: Under A1), A5), and A7), for each μ

$$\frac{1}{n}\sum_{j=\mu}^{\mu+n}S(\tilde{Y}_j;\,\varphi^*) \to 0 \text{ w.p. 1, as } n\to\infty$$

$$\frac{1}{n}\sum_{j=\mu}^{\mu+n}\frac{\partial}{\partial\varphi}S(\tilde{Y}_j;\,\varphi^*) \to \frac{\partial}{\partial\varphi}H(\varphi^*) \text{ w.p. 1, as } n\to\infty.$$

Proof: Note

$$S(\tilde{Y}_j;\,\varphi^*) = S(g(\theta^*,\,\tilde{Y}_{j-1},\,\ldots,\,\tilde{Y}_{j-d+1},\,e_j);\,\varphi^*).$$

By A1), $\{Q_k\}$ is stationary and ergodic. Thus

$$\frac{1}{n}\sum_{j=\mu}^{\mu+n}S(\tilde{Y}_j;\,\varphi^*)$$

converges w.p. 1 to

$$ES(\tilde{Y}_j;\, \varphi^*) = H(\varphi^*) = 0.$$

The first assertion is verified.

Since $\{(\partial/\partial\varphi)S(\tilde{Y}_j;\, \varphi)\}$ is uniformly integrable, by A5), the dominated convergence theorem then yields

$$\frac{1}{n}\sum_{j=\mu}^{\mu+n} \frac{\partial}{\partial\varphi}\, S(\tilde{Y}_j;\, \varphi^*) \to E\frac{\partial}{\partial\varphi}\, S(\tilde{Y}_j;\, \varphi^*) = \frac{\partial}{\partial\varphi}\, H(\varphi^*) \text{ w.p. } 1.$$

The second assertion is also proved. □

Lemma 3.8: Under A1), A5), and A8), $W^n(\cdot)$ converges weakly to a Brownian motion $W(\cdot)$ with covariance Σ where

$$\Sigma = ES_1S_1' + \sum_{i=2}^{\infty} ES_1S_i' + \sum_{i=2}^{\infty} ES_iS_1'. \tag{24}$$

Proof: Note that

$$\begin{aligned}
&E_{m(t_n+t)}\left|W^n(t+s) - W^n(t)\right|^2 \\
&\quad\le \kappa \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \sum_{i\ge j} \sqrt{\varepsilon_i}\sqrt{\varepsilon_j} E_{m(t_n+t)} S_i'S_j \\
&\quad\le \kappa \sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \varepsilon_j \sum_{i\ge j} E_{m(t_n+t)} S_i'S_j.
\end{aligned}$$

Since

$$\left| E\sum_{i\ge j} E_{m(t_n+t)} S_i'S_j \right| \le \sum_{i\ge j} R(i-j) \le \kappa$$

by virtue of A8), and

$$\begin{aligned}
&\sum_{j=m(t_n+t)}^{m(t_n+t+s)-1} \varepsilon_j \\
&\quad= O(s) \lim_{s\to 0} \limsup_{n} EE_{m(t_n+t)}\left|W^n(t+s) - W^n(t)\right|^2 = 0.
\end{aligned}$$

It follows from the tightness criterion (see [14], [23]), $W^n(\cdot)$ is tight in $D^{r(r+1)}(-\infty,\, \infty)$.

By Prohorov's theorem (see [5], [14]), we can extract a convergent subsequence. Do so and for simplicity, still index it by n with limit denoted by $W(\cdot)$. For any bounded and continuous function $h(\cdot)$, any integer k, any real numbers $t,\, s > 0$, and any $t_j \le t$ for $j \le k$, we have that $W^n(t_j)$ are $\mathcal{F}_{m(t_n+t)}$ measurable and

$$\begin{aligned}
&Eh\left(W^n(t_j), j \le k\right)\left(W^n(t+s) - W^n(t)\right) \\
&\quad= Eh\left(W^n(t_j), j \le k\right) E_{m(t_n+t)} \sum_{i=m(t_n+t)}^{m(t_n+t+s)-1} \sqrt{\varepsilon_i} S_i \to 0, \\
&\qquad\qquad \text{as } n \to \infty.
\end{aligned}$$

Owing to A5), $\{|S_i|^2\}$ is uniformly integrable. This together with the estimate above and the weak convergence implies

$$Eh\left(W(t_j),\, j \le k\right)\left(W(t+s) - W(t)\right) = 0.$$

Thus, $W(\cdot)$ is a continuous martingale. Next consider its quadratic variation. We have

$$\begin{aligned}
&Eh\left(W^n(t_j), j \le k\right) \\
&\quad \cdot \left(W^n(t+s) - W^n(t)\right)\left(W^n(t+s) - W^n(t)\right)' \\
&\quad \to Eh\left(W(t_j), j \le k\right)\Sigma s, \qquad \text{as } n \to \infty
\end{aligned}$$

where Σ is given by (21). Therefore, the limit $W(\cdot)$ is a Brownian motion with covariance Σ as desired. Since the limit does not depend on the chosen subsequence, the lemma follows. □

To carry out the weak convergence analysis, an appropriate way is to truncate the dynamics of $U^n(\cdot)$ and works with a truncated version $U^{n,\,M}(\cdot)$ with $M < \infty$ first. One then proceeds with proving the tightness and weak convergence of $U^{n,\,M}(\cdot)$ and finally passing the limit to $M \to \infty$ (see [23], [25] for references). Using the lemmas above, carrying out the details as in [25, Sections 10.1 and 10.2], we establish the following theorem.

Theorem 3.9: A1), and A4)–A8). Then the sequence $\{U^n(\cdot),\, W^n(\cdot)\}$ converges weakly in

$$D^{r(r+1)}[0,\, \infty) \times D^{r(r+1)}(-\infty,\, \infty)$$

to $(U(\cdot),\, W(\cdot))$, where $W(\cdot)$ is a Brownian motion with covariance Σ and $U(\cdot)$ is stationary such that

$$dU = RU\, dt + dW \tag{25}$$

where

$$R = \begin{cases} H_\varphi(\varphi^*), & \text{under A4) a)} \\ H_\varphi(\varphi^*) + I/2, & \text{under A4) b).} \end{cases}$$

Remark 3.10: In the sense of equivalence of probability distribution on $D^{r(r+1)}[0,\, \infty) \times D^{r(r+1)}(-\infty,\, \infty)$, we can write

$$U(t) = \int_{-\infty}^{t} \exp\left(R(t-s)\right)\, dW(s)$$

which is the stationary solution of (22). The reason for using $D^{r(r+1)}(-\infty,\, \infty)$ is mainly because it allows us to write $U(t)$ as in the above representation involving the entire past of the Brownian motion (see [25, Ch. 10] for further details).

Note that the above theorem effectively gives the rate of convergence result. That is, it gives the order of the scaling, namely, $\sqrt{\varepsilon_k}$ and the asymptotic covariance. To further illustrate, rewrite (22) as

$$dU = RU\, dt + \Sigma^{1/2}\, d\tilde{w}$$

where $\tilde{w}(\cdot)$ is a standard Brownian motion. The asymptotic covariance Σ_0 of the underlying process is a solution of the algebraic Liapunov equation $R\Sigma_0 + \Sigma_0 R' = -\Sigma$, and has the following representation:

$$\Sigma_0 = \int_0^{\infty} \exp(Rt)\Sigma\exp(R't)\, dt. \tag{26}$$

An immediate consequence is that as $n \to \infty$

$$(\varphi_n - \varphi^*)/\sqrt{\varepsilon_n} \sim N(0,\, \Sigma_0)$$

i.e., it is asymptotically normal with asymptotic covariance Σ_0. The result we have obtained is more general than that of [35]. First it is from a stochastic process point of view, and focuses on trajectories of the normalized sequence of the estimation errors. Second, it coves a broad range of step-size sequences.

Note that the step sizes have a major influence on the rate of convergence. This is clearly seen from the representations of R corresponding to assumptions A4) a) and A4) b), respectively; see also the related work [45] for the rate results for global optimization algorithms.

IV. Convergence Acceleration by Iterate and Observation Averaging

For designing recursive algorithms, to improve the asymptotic efficiency is an important issue. The effort for analyzing stochastic approximation type of algorithm in this direction can be traced back to that of Chung [9]. Suppose that $\varepsilon_n = 1/n^\gamma$ for some $0 < \gamma \le 1$ and that there is a unique asymptotic stable point φ^* in the interior of G. Then $(\varphi_n - \varphi^*)/\sqrt{\varepsilon_n}$ is asymptotically normal. Among the γ's given above, the best one is $\gamma = 1$ as far as the scaling factor is concerned. If one uses $\varepsilon_n = \Gamma/n$, then it can be demonstrated that the best choice of Γ is the inverse of the gradient of $H(\cdot)$ evaluated at $\varphi = \varphi^*$. This quantity is normally not available. One could construct a sequence of estimates, but the amount of computation is often infeasible, especially for many applications we encounter in the hidden Markov estimation. In addition, from a computation point of view, one may not wish to use a rapidly decreasing sequence of step sizes decaying as $\varepsilon_n = O(1/n)$ since this produces a very slow movement in the initial stage.

Taking these into consideration, an approach of using iterate averaging was suggested in Polyak [33] and Ruppert [37] independently. The idea is that after using a larger than $O(1/n)$ step-size sequence in an initial estimation, one takes an average of the resulting iterates yielding asymptotic optimality. Their results were extended in Yin [44] for mixing type of signals, and generalized further in Kushner and Yang [24] together with an explanation on why the approach works well using a two-time scale interpretation. Meanwhile, Bather [2] suggested another approach that requires the use of not only the iterate averaging but also the averaging in the observation. Schwabe [38] examined further this approach. The convergence and asymptotic optimality were obtained in Yin and Yin [46] for correlated noise.

Treating the HMM estimation problem, Rydén [35] suggested to adopt the iterate averaging to improve the efficiency of the estimation scheme. In this paper, motivated by the work [2], we use an averaging approach with averaging in both iterates and observations. This approach seems to have a smoothing effect that is useful for the initial stage of approximation. Henceforth, for notational simplicity, we take the initial time of the iteration to be $n = 1$ and consider an algorithm of the form

$$\begin{aligned} \varphi_{n+1} &= \Pi_G\left(\overline{\varphi}_n + \varepsilon_n n \overline{S}_n\right)\\ \overline{\varphi}_{n+1} &= \overline{\varphi}_n - \frac{1}{n+1}\,\overline{\varphi}_n + \frac{1}{n+1}\varphi_{n+1},\\ \overline{S}_{n+1} &= \overline{S}_n - \frac{1}{n+1}\,\overline{S}_n + \frac{1}{n+1}\,S_{n+1} \end{aligned} \tag{27}$$

where G is the same constraint set as given before. Note that the algorithm above has the two-time scale interpretation; see [24] (see also [4] and [25]). Rewrite the first equation above as

$$\varphi_{n+1} = \overline{\varphi}_n + \varepsilon_n n \overline{S}_n + \varepsilon_n M_n \tag{28}$$

where M_n is the projection term. We proceed to analyze the above algorithm.

A. Convergence

In what follows, we take $\varepsilon_n = 1/n^\gamma$ with $1/2 < \gamma < 1$. More general step size sequences can be treated. The particular form of the step sizes are selected to simplify the argument and notation in the proof. Note that strictly speaking, (25) is not a recursion for φ_n in the usual stochastic approximation setting. First, let us rewrite it in a more convenient form.

In view of the definition in (25), taking difference of $\varphi_{n+1} - \varphi_n$ and using

$$\overline{\varphi}_n - \overline{\varphi}_{n-1} = \frac{1}{n}\,\varphi_n - \frac{1}{n}\,\overline{\varphi}_{n-1}$$

we arrive at

$$\begin{aligned} \varphi_{n+1} = \varphi_n + \frac{1}{n^\gamma}\,S(\tilde{Y}_n;\,\varphi_n) + \frac{1-\gamma}{(n-1)^\gamma}\,\frac{1}{n}\sum_{i=1}^{n-1} S(\tilde{Y}_i;\,\varphi_i)\\ + \frac{1}{(n-1)^\gamma}\,\rho_n \sum_{i=1}^{n-1} S(\tilde{Y}_i;\,\varphi_i) + \frac{1}{n^\gamma}\,M_n,\\ \text{for } n > 1 \end{aligned} \tag{29}$$

where $\rho_n = O(1/n^2)$.

In [46], dealing with an unconstrained algorithm, we used a recursive formulation similar as above, and examined an auxiliary sequence that is known to converge. Then we compared the difference of the estimates with that of the auxiliary process, and upper-bounded their difference by means of Gronwall's inequality. Here we use a somewhat different approach and treat (26) directly. Define $\varphi^n(\cdot)$ and $M^n(\cdot)$ the same as before. We have the following result.

Theorem 4.1: Under the conditions of Theorem 3.4, its conclusions continue to hold for (24).

Proof: Define $F^n(\cdot)$, $\tilde{F}^n(\cdot)$, and $\hat{F}^n(\cdot)$ on $t \in (-\infty, \infty)$ as the piecewise-constant interpolations of the second, the third, and the fourth terms on the right-hand side of the equality sign of (26); for $t \ge 0$, these terms are

$$\begin{aligned} F^n(t) &= \sum_{i=n}^{m(t_n+t)-1} \frac{1}{i^\gamma}\,S(\tilde{Y}_i;\,\varphi_i)\\ \tilde{F}^n(t) &= \sum_{i=n}^{m(t_n+t)-1} \frac{1-\gamma}{(i-1)^\gamma}\,\frac{1}{i}\sum_{j=1}^{i-1} S(\tilde{Y}_j;\,\varphi_j)\\ \hat{F}^n(t) &= \sum_{i=n}^{m(t_n+t)-1} \frac{1}{(i-1)^\gamma}\,\rho_i \sum_{j=1}^{i-1} S(\tilde{Y}_j;\,\varphi_j). \end{aligned}$$

Then we have

$$\varphi^n(t) = \varphi_n + F^n(t) + \tilde{F}^n(t) + \hat{F}^n(t) + M^n(t).$$

By applying Lemma 3.2 to each of the functions above, we conclude that there is a null set $\tilde{N}$ such that for all $\omega \notin \tilde{N}$, $\{\varphi^n(\omega,\,\cdot),\, M^n(\omega,\,\cdot)\}$ is equicontinuous (in the extended sense [25, p. 73]). Extract a convergent subsequence with index n_k and limit $(\varphi(\omega,\,\cdot),\, M(\omega,\,\cdot))$. We proceed to characterize the limit.

Work with a fixed sample path for $\omega \notin \tilde{N}$, and suppress the ω dependence. For $t \ge 0$, with given $\delta > 0$, split $F^{n_k}(\cdot)$ into three terms

$$F_1^{n_k}(t) = \sum_{j=1}^{\lfloor t/\delta \rfloor - 1}\ \sum_{i=m(t_{n_k}+j\delta)}^{m(t_{n_k}+j\delta+\delta)-1} \frac{1}{i^\gamma}\,H\left(\varphi(j\,\delta)\right)$$

$$F_2^{n_k}(t) = \sum_{j=1}^{\lfloor t/\delta \rfloor - 1} \sum_{i=m(t_{n_k}+j\,\delta)}^{m(t_{n_k}+j\,\delta+\delta)-1} \cdot \frac{1}{i^\gamma}\left[S(\tilde{Y}_i;\,\varphi_i) - S\left(\tilde{Y}_i;\,\varphi(j\delta)\right)\right]$$

$$F_3^{n_k}(t) = \sum_{j=1}^{\lfloor t/\delta \rfloor - 1} \sum_{i=m(t_{n_k}+j\,\delta)}^{m(t_{n_k}+j\,\delta+\delta)-1} \cdot \frac{1}{i^\gamma}\left[S\left(\tilde{Y}_i;\,\varphi(j\,\delta)\right) - H\left(\varphi(j\,\delta)\right)\right]$$

where $\lfloor b \rfloor$ denotes the integer part of $b \in \mathbb{R}$. As $\delta \to 0$, it is easy to see that

$$F_1^{n_k}(t) \to \int_0^t H\left(\varphi(s)\right)\,ds.$$

For $F_2^{n_k}(t)$, by A2), we have

$$|F_2^{n_k}(t)| \le \sup_{\substack{j \le t/\delta \\ m(t_{n_k}+j\,\delta) \le i \le m(t_{n_k}+j\,\delta+\delta)-1}} \tilde{\rho}\left(\varphi_i - \varphi(j\,\delta)\right) \cdot \sum_{i=m(t_{n_k})}^{m(t_{n_k}+t)-1} \frac{1}{i^\gamma}\,\hat{\rho}(\tilde{Y}_i) \to 0, \qquad \text{as } \delta \to 0 \text{ and } k \to \infty.$$

We next analyze the third term. In view of Lemma 3.2, for each fixed φ

$$\sum_{i=m(t_{n_k}+j\,\delta)}^{m(t_{n_k}+j\,\delta+\delta)-1} \frac{1}{i^\gamma}\left[S(\tilde{Y}_i;\,\varphi) - H(\varphi)\right] \to 0 \text{ w.p. 1.}$$

What we need to do now is to approximate $\varphi(j\,\delta)$ by some fixed φ. To do so, for given $\eta > 0$, let $\{B_l^\eta;\, l \le l_\eta\}$ be a finite collection of disjoint sets with diameter smaller than η, and $\varphi_l^\eta \in B_l^\eta$ for $l \le l_\eta$, and $\bigcup_{l=1}^{l_\eta} B_l^\eta = G$. Write

$$\sum_{i=m(t_{n_k}+j\,\delta)}^{m(t_{n_k}+j\,\delta+\delta)-1} \frac{1}{i^\gamma}\left[S\left(\tilde{Y}_i;\,\varphi(j\,\delta)\right) - H\left(\varphi(j\,\delta)\right)\right] = \sum_{l=1}^{l_\eta} I_{\{\varphi(j\,\delta) \in B_l^\eta\}} \sum_{i=m(t_{n_k}+j\,\delta)}^{m(t_{n_k}+j\,\delta+\delta)-1} \frac{1}{i^\gamma}\left[S(\tilde{Y}_i;\,\varphi_l^\eta) - H(\varphi_l^\eta)\right].$$

For fixed $\eta > 0$, as $k \to \infty$, the above term goes to 0 by Lemma 3.2. It then follows that the limit is zero as $k \to \infty$ and then $\eta \to 0$.

Using the same technique, we can show $\tilde{F}^{n_k}(t) \to 0$ and $\hat{F}^{n_k}(t) \to 0$ as $k \to \infty$. The desired limit then follows. □

B. Asymptotic Optimality

To proceed, we demonstrate the averaging algorithm is asymptotically efficient. In what follows, assume that $\varphi_n \to \varphi^*$ w.p. 1, and $\varphi^* \in G^0$. Thus, the boundary of G, namely, ∂G, is reached only a finite number of times. Without loss of generality, we drop the reflection term M_n and assume the iterates are bounded and in G^0.

Estimate of $E|\varphi_n - \varphi^*|^2$: The estimate is of stability type. We use the perturbed Liapunov function method that is to add a small perturbation to a Liapunov function. The purpose of the addition of the small perturbation is to result in desired cancellations. Define the perturbations $\tilde{P}_n$ and $\hat{P}_n$ by

$$\tilde{P}_n = \sum_{i=n}^{\infty} \frac{1}{i^\gamma} B_{i(n+1)} E_n S(\tilde{Y}_i;\,\varphi^*)$$

$$\hat{P}_n = \sum_{i=n}^{\infty} \frac{1}{i^\gamma} B_{i(n+1)} E_n \left[S_\varphi(\tilde{Y}_i;\,\varphi^*) - H_\varphi(\varphi^*)\right]$$

$$\overline{P}_n = \sum_{i=n}^{\infty} \frac{1}{i^\gamma} B_{i(n+1)} E_n \xi_i \tag{30}$$

respectively, where

$$\xi_n = \zeta_n \left(1 + O\left(\frac{1}{n}\right)\right)$$

and

$$\zeta_n = \frac{1-\gamma}{n} \sum_{i=1}^{n-1} S(\tilde{Y}_i;\,\varphi_i) + \rho_n \sum_{i=1}^{n-1} S(\tilde{Y}_i;\,\varphi_i).$$

Their use will be clear from the subsequent development.

Theorem 4.2: Assume that A1), A2), and A5)–A8) hold, for sufficiently small $\eta > 0$

$$E\left|S(\tilde{Y}_n;\,\varphi_n)\right| I_{\{|\varphi_n - \varphi^*| \le \eta\}} < \infty \tag{31}$$

and

$$\begin{aligned}
&\sup_n E|\hat{P}_n|\left(1 + |S(\tilde{Y}_n;\,\varphi_n)|\right) I_{\{|\varphi_n - \varphi^*| \le \eta\}} \le \frac{\kappa}{n^\gamma} \\
&\sup_n E|\tilde{P}_n|\left(1 + |S(\tilde{Y}_n;\,\varphi_n)| + |S(\tilde{Y}_n;\,\varphi_n)|^2\right) \\
&\qquad \cdot I_{\{|\varphi_n - \varphi^*| \le \eta\}} \le \frac{\kappa}{n^\gamma} \\
&\sup_n E|\overline{P}_n|\left(1 + |\xi_n| + |\xi_n|^2\right) I_{\{|\varphi_n - \varphi^*| \le \eta\}} \le \frac{\kappa}{n^\gamma}.
\end{aligned} \tag{32}$$

Then $E|\varphi_n - \varphi^*|^2 = O(1/n^\gamma)$ and $E|\overline{\varphi}_n - \varphi^*|^2 = O(1/n^{1/2})$ for n large enough.

Proof: Use a Taylor expansion

$$S(\tilde{Y};\,\varphi) = S(\tilde{Y};\,\varphi^*) + S_\varphi(\tilde{Y};\,\varphi^*)(\varphi - \varphi^*) + \pi(\varphi,\,\tilde{Y})(\varphi - \varphi^*)$$

where

$$\pi(\varphi,\,\tilde{Y}) = \int_0^1 \left[S_\varphi\left(\tilde{Y};\,\varphi^* + s(\varphi - \varphi^*)\right) - S_\varphi(\tilde{Y};\,\varphi^*)\right] ds.$$

Rewrite (26) as

$$\begin{aligned}
\varphi_{n+1} = \varphi_n + \frac{1}{n^\gamma}\Big[&H_\varphi(\varphi^*)(\varphi_n - \varphi^*) + S(\tilde{Y}_n;\,\varphi^*) \\
&+ \left[S_\varphi(\tilde{Y}_n;\,\varphi^*) - H_\varphi(\varphi^*)\right] \\
&\cdot (\varphi_n - \varphi^*) + \pi(\varphi_n,\,\tilde{Y}_n)(\varphi_n - \varphi^*) + \xi_n\Big]
\end{aligned} \tag{33}$$

The w.p. 1 convergence of φ_n to φ^*, A2), and the uniform integrability of $\{\tilde{Y}_n\}$ imply that $\zeta_n \to 0$ w.p. 1. Let $\eta > 0$ be sufficiently small. Given $\eta_0 > 0$ small enough, there is an $N_0 = N_{\eta,\,\eta_0}$ such that for all $n \ge N_{\eta,\,\eta_0}$

$$|\varphi_n| \le \eta \quad \text{and} \quad |\xi_n| \le \eta \text{ w.p.} \ge 1 - \eta_0. \tag{34}$$

By modifying the process on a set of probability at most η_0, we may suppose that (31) holds for all $n \ge N_{\eta,\,\eta_0}$ and all conditions hold for the modified process. Denote the modified process by $\varphi_n^{\eta_0}$. The tightness of $\{n^{\gamma/2}\varphi_n^{\eta_0}\}$ will imply the tightness of $\{n^{\gamma/2}\varphi_n\}$. Thus, for the tightness proof, without loss of generality, assume the original process itself $\{\varphi_n\}$ satisfies inequality (31).

Since $H_\varphi(\varphi^*)$ is Hurwitz, by virtue of the inverse Liapunov stability theorem, for any symmetric and positive definite matrix P

$$QH_\varphi(\varphi^*) + H'_\varphi(\varphi^*)Q = -P$$

has a unique solution that is symmetric and positive definite. Choose a pair of symmetric and positive-definite matrices Q and P such that $P \geq \lambda Q$ where $\lambda = (\mu_{\max}(P)/\mu_{\min}(Q))$ and $\mu_{\max}(A)$ and $\mu_{\min}(A)$ denote the maximum and minimal eigenvalues of a symmetric and positive-definite matrix A, respectively.

Define a Liapunov function by

$$V(\varphi) = (\varphi - \varphi^*)'Q(\varphi - \varphi^*).$$

Then

$$\begin{aligned}
&E_nV(\varphi_{n+1}) - V(\varphi_n)\\
&\quad= \frac{2}{n^\gamma}(\varphi_n - \varphi^*)'QH_\varphi(\varphi^*)(\varphi_n - \varphi^*)\\
&\qquad+ O\left(\frac{1}{n^\gamma}\right)\left(|\varphi_n - \varphi^*|^2\right)\left|\pi(\varphi_n, \tilde{Y}_n)\right|\\
&\qquad+ \frac{2}{n^\gamma}(\varphi_n - \varphi^*)'QS(\tilde{Y}_n; \varphi^*)\\
&\qquad+ \frac{2}{n^\gamma}(\varphi_n - \varphi^*)'Q\left[S_\varphi(\tilde{Y}_n; \varphi^*) - H_\varphi(\varphi^*)\right](\varphi_n - \varphi^*)\\
&\qquad+ \frac{2}{n^\gamma}(\varphi_n - \varphi^*)'Q\xi_n\\
&\qquad+ O\left(\frac{1}{n^{2\gamma}}\right)E_n\left|S(\tilde{Y}_n; \varphi_n) + \xi_n\right|^2. \qquad (35)
\end{aligned}$$

Given $\eta > 0$, denote η_1 to be such that $\eta_1 \to 0$ as $\eta \to 0$ and

$$\sup_{n,\,|\varphi_n - \varphi^*| \leq \eta}\left[\left|\pi(\varphi_n, \tilde{Y}_n)\right| + |\xi_n|\right] = \eta_1.$$

Define the perturbed Liapunov function by

$$\begin{aligned}
V_n(\varphi) &\overset{\text{def}}{=} V(\varphi_n) + \delta V_n(\varphi_n)\\
&= V(\varphi_n) + 2(\varphi_n - \varphi^*)'Q\hat{P}_n(\varphi_n - \varphi^*)\\
&\quad + 2(\varphi_n - \varphi^*)'Q[\tilde{P}_n + \overline{P}_n].
\end{aligned}$$

Then in the calculation of $E_n\delta V_{n+1}(\varphi_{n+1}) - \delta V_n(\varphi_n)$, three negative terms cancel that of the terms on the third, the fourth, and the fifth lines of (32). Since

$$\begin{aligned}
&2(\varphi_n - \varphi^*)'QH_\varphi(\varphi^*)(\varphi_n - \varphi^*)\\
&\quad= (\varphi_n - \varphi^*)'\left(QH_\varphi(\varphi^*) + H'_\varphi(\varphi^*)Q\right)(\varphi_n - \varphi^*)\\
&\quad= -(\varphi_n - \varphi^*)'P(\varphi_n - \varphi^*) \leq -\lambda V(\varphi_n)
\end{aligned}$$

we obtain

$$\begin{aligned}
&E_nV_{n+1}(\varphi_{n+1}) - V_n(\varphi_n)\\
&\quad\leq -\frac{1}{n^\gamma}\lambda V(\varphi_n) + O\left(\frac{1}{n^{2\gamma}}\right)E_n\left|S(\tilde{Y}_n; \varphi_n)\right|^2\\
&\qquad+ \eta_1 O\left(\frac{1}{n^\gamma}\right)|\varphi_n - \varphi^*|^2 + \frac{1}{n^\gamma}|\varphi_n - \varphi^*|^2|E_n\hat{P}_{n+1}|\\
&\qquad+ E_n|\hat{P}_{n+1}|\left(O\left(\frac{1}{n^\gamma}\right)|\varphi_n - \varphi^*|\left|S(\tilde{Y}_n; \varphi_n)\right|\right.\\
&\qquad\qquad\left. + O\left(\frac{1}{n^{2\gamma}}\right)\left|S(\tilde{Y}_n; \varphi_n)\right|^2\right)\\
&\qquad+ O\left(\frac{1}{n^\gamma}\right)|E_n\tilde{P}_{n+1}||\varphi_n - \varphi^*|\\
&\qquad+ E_nO\left(\frac{1}{n^\gamma}\right)\left|S(\tilde{Y}_n; \varphi_n)\right||\tilde{P}_{n+1}|\\
&\qquad+ \frac{1}{n^\gamma}|\varphi_n - \varphi^*|^2|E_n\overline{P}_{n+1}| + E_n|\overline{P}_{n+1}|\\
&\qquad\cdot\left(O\left(\frac{1}{n^\gamma}\right)|\varphi_n - \varphi^*||\xi_n| + O\left(\frac{1}{n^{2\gamma}}\right)|\xi_n|^2\right). \qquad (36)
\end{aligned}$$

Taking expectation and using the assumptions, detailed computation then leads to

$$EV_{n+1}(\varphi_{n+1}) \leq -\frac{\lambda}{n^\gamma}EV(\varphi_n) + \lambda_\eta\frac{1}{n^\gamma}E|\varphi_n - \varphi^*|^2 + O\left(\frac{1}{n^{2\gamma}}\right)$$

where $\lambda_\eta \to 0$ as $\eta \to 0$. Recall that $|\varphi_n - \varphi^*| \leq \eta$ for sufficiently small η. Using (29), $E\,\delta V_n(\varphi_n) \leq \kappa/n^\gamma$. It follows that for $\lambda_1 < \lambda$

$$EV_{n+1}(\varphi_{n+1}) \leq \left(1 - \lambda_1\frac{1}{n^\gamma}\right)EV_n(\varphi_n) + O\left(\frac{1}{n^{2\gamma}}\right).$$

Iterating on the above inequality

$$\begin{aligned}
EV_n(\varphi_{n+1}) &\leq \kappa\prod_{i=N_0+1}^{n}\left(1 - \frac{\lambda_1}{i^\gamma}\right)EV_{N_0}(\varphi_{N_0})\\
&\qquad+ \kappa\sum_{i=N_0}^{n}\frac{1}{i^{2\gamma}}\prod_{j=i+1}^{n}\left(1 - \frac{\lambda_1}{j^\gamma}\right) = O\left(\frac{1}{n^\gamma}\right).
\end{aligned}$$

This, in turn, yields that $EV(\varphi_n) \leq \kappa/n^\gamma$ as desired.

Thus, the estimate for $E|\varphi_n - \varphi^*|^2$ is obtained. The estimate for $E|\overline{\varphi}_n - \varphi^*|^2$ can be obtained similarly with the use of the recursion for $\overline{\varphi}_n$. □

Asymptotic Normality: First, we deduce an asymptotic equivalency which indicates that $\sqrt{n}(\overline{\varphi}_n - \varphi^*)$ has a very simple form. Then the asymptotic normality follows.

Note that the algorithm as a recursion for φ_n is

$$\overline{\varphi}_{n+1} = \overline{\varphi}_n + \frac{1}{n^\gamma(n+1)}\sum_{i=1}^{n}S(\tilde{Y}_i; \varphi_i).$$

Define $\tilde{\varphi}_n = \varphi_n - \varphi^*$ and $\hat{\varphi}_n = \overline{\varphi}_n - \varphi^*$. We then obtain

$$\begin{aligned}
\hat{\varphi}_{n+1} &= \hat{\varphi}_n + \frac{H_\varphi(\varphi^*)}{n^\gamma}\hat{\varphi}_n - \frac{H_\varphi(\varphi^*)}{n^\gamma(n+1)}\hat{\varphi}_n\\
&\quad+ \frac{1}{n^\gamma(n+1)}\sum_{i=1}^{n}\left[S(\tilde{Y}_i; \varphi^*) - H(\tilde{\varphi}_i) + O\left(|\tilde{\varphi}_i|^2\right)\right]. \qquad (37)
\end{aligned}$$

Define a matrix-valued product as

$$A_{nk} = \begin{cases} \prod_{i=k+1}^{n} (I + H_\varphi(\varphi^*)/i^\gamma), & k < n \\ I, & k = n. \end{cases}$$

Writing down the solution of (34) in its variational form leads to

$$\begin{aligned} \sqrt{n}\hat{\varphi}_{n+1} = \sqrt{n}A_{n0}\hat{\varphi}_1 - \sqrt{n}\sum_{k=1}^{n} \frac{1}{k^\gamma(k+1)} A_{nk}H_\varphi(\varphi^*)\hat{\varphi}_k \\ + \sqrt{n}\sum_{k=1}^{n} \frac{1}{k^\gamma(k+1)} A_{nk} \sum_{i=1}^{k} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right] \\ + \sqrt{n}\sum_{k=1}^{n} \frac{1}{k^\gamma(k+1)} A_{nk} \sum_{i=1}^{k} O(|\tilde{\varphi}_k|^2). \end{aligned} \tag{38}$$

Lemma 4.3: Assume the conditions of Theorem 4.2. Then the following assertions hold:

a) As $n \to \infty$

$$\sqrt{n}\hat{\varphi}_{n+1} = -\frac{H_\varphi^{-1}(\varphi^*)}{\sqrt{n}} \sum_{i=1}^{n} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right] + o(1)$$

where $o(1) \to 0$ in probability as $n \to \infty$.

b) Define

$$W^n(t) = \frac{\lfloor nt \rfloor}{\sqrt{n}} \left(\overline{\varphi}_{\lfloor nt \rfloor + 1} - \varphi^*\right), \qquad \text{for } t \in [0,\, 1].$$

Then

$$W^n(t) = -\frac{H_\varphi(\varphi^*)}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right] + o(1)$$

where $o(1) \to 0$ in probability uniformly in t.

Proof: We prove only the first assertion. The second one can be proved analogously. Examine (35) term by term. Let us begin with the last term. Using Theorem 4.2 and the boundedness of $\{\varphi_n\}$

$$\begin{aligned} E\left|\sqrt{n}\sum_{k=1}^{n} \frac{1}{k^\gamma(k+1)} A_{nk} \sum_{i=1}^{k} O\left(|\tilde{\varphi}_i|^2\right)\right| \\ \leq \kappa\sqrt{n}\sum_{k=1}^{n} \frac{1}{k^\gamma} |A_{nk}| O\left(E|\hat{\varphi}_k|^2\right) \\ \leq \kappa\sqrt{n}O(n^{-\gamma}) \to 0, \qquad \text{as } n \to \infty. \end{aligned}$$

We also have (by using Theorem 4.2),

$$E\left|\sqrt{n}A_{n0}\hat{\varphi}_1\right| \leq \kappa\sqrt{n}|A_{n0}|E|\hat{\varphi}_1| \to 0, \qquad \text{as } n \to \infty$$

and

$$\begin{aligned} E\left|\sqrt{n}\sum_{k=1}^{n} \frac{1}{k^\gamma(k+1)} A_{nk}H_\varphi(\varphi^*)\hat{\varphi}_k\right| \\ \leq \kappa\sqrt{n}\sum_{k=1}^{n} \frac{1}{k^\gamma(k+1)} |A_{nk}||H_\varphi(\varphi^*)|E|\hat{\varphi}_k| \to 0, \\ \text{as } n \to \infty. \end{aligned}$$

For the term on the second line of (35), using a partial summation

$$\begin{aligned} \sqrt{n}\sum_{k=1}^{n} \frac{1}{k^\gamma k} A_{nk} \sum_{i=1}^{k} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right] \\ = \left[\sum_{k=1}^{n} \frac{1}{k^\gamma} A_{nk}\right] \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right]\right] \\ + \sqrt{n} \sum_{k=1}^{n-1} \left[\sum_{i=1}^{k} \frac{1}{i^\gamma} A_{ni}\right] \left[\frac{1}{k} \sum_{i=1}^{k} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right] \right. \\ \left. - \frac{1}{k+1} \sum_{i=1}^{k+1} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right]\right]. \end{aligned}$$

Since

$$\sum_{k=1}^{n} \frac{1}{k^\gamma} A_{nk} = -H_\varphi^{-1}(\varphi^*)(I - A_{n0})$$

$$\begin{aligned} \left[\sum_{k=1}^{n} \frac{1}{k^\gamma} A_{nk}\right] \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right]\right] \\ = -\frac{H_\varphi^{-1}(\varphi^*)}{\sqrt{n}} \sum_{i=1}^{n} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right] + o(1) \end{aligned}$$

where $o(1) \to 0$ in probability as $n \to \infty$. Similarly, we have

$$\begin{aligned} \sqrt{n}\sum_{k=1}^{n-1} \left[\sum_{i=1}^{k} \frac{1}{i^\gamma} A_{ni}\right] \\ \cdot \left[\frac{1}{k} \sum_{i=1}^{k} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right] \right. \\ \left. - \frac{1}{k+1} \sum_{i=1}^{k+1} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right]\right] \\ = \sqrt{n}H_\varphi^{-1}(\varphi^*) \sum_{k=1}^{n-1} (A_{nk} - A_{n0}) \\ \cdot \left[\frac{1}{k(k+1)} \sum_{i=1}^{k} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right] \right. \\ \left. - \frac{1}{(k+1)} \left[S(\tilde{Y}_{k+1};\, \varphi_{k+1}) - H(\varphi_{k+1})\right]\right] \\ \to 0 \text{ in probability.} \end{aligned}$$

Thus, asymptotically the term on the second line of (35) is given by

$$-\frac{H_\varphi^{-1}(\varphi^*)}{\sqrt{n}} \sum_{i=1}^{n} \left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right] + o(1). \qquad \square$$

Theorem 4.4: Under the conditions of Lemma 4.3, $W^n(\cdot)$ converges weakly to a Brownian motion with the optimal covariance $H_\varphi^{-1}(\varphi^*)\Sigma(H_\varphi^{-1}(\varphi^*))'$, where Σ is given by (21).

Proof: By virtue of the argument as in [24], [25], it can be shown that $(1/\sqrt{n})\sum_{i=1}^{\lfloor nt \rfloor} S_i$ converges weakly to a Brownian motion. Owing to $\varphi_n \to \varphi^*$ and the continuity of $H(\cdot)$, $H(\varphi_n) \to H(\varphi^*) = 0$. Likewise, it follows from A2) and the boundedness of the iterates, $E[S(\tilde{Y}_n;\, \varphi_n) - S(\tilde{Y}_n;\, \varphi^*)] \rightarrow 0$ as $n \rightarrow \infty$. Therefore, the covariance of the resulting Brownian motion

$$\lim_n \frac{\lfloor nt \rfloor}{n} \frac{1}{\lfloor nt \rfloor} \sum_{i=1}^{\lfloor nt \rfloor} \sum_{j=1}^{\lfloor nt \rfloor} \cdot E\left[S(\tilde{Y}_i;\, \varphi_i) - H(\varphi_i)\right]\left[S(\tilde{Y}_j;\, \varphi_j) - H(\varphi_j)\right]'$$
$$= \lim_n t \frac{1}{\lfloor nt \rfloor} \sum_{i=1}^{\lfloor nt \rfloor} \sum_{j=1}^{\lfloor nt \rfloor} ES(\tilde{Y}_i;\, \varphi^*)S(\tilde{Y}_j;\, \varphi^*) = t\Sigma.$$

Finally, by Slutsky's theorem, the desired result follows. □

Remark 4.5: An interesting problem to study concerns the rates of convergence taking into consideration the computational budget devoted to the underlying computation. Such problems were dealt with in [26]; see also the subsequent generalization in [41].

V. Constant Step-Size Tracking Algorithms

In this section, we study algorithms with a constant step size, i.e., $\varepsilon_n = \varepsilon > 0$. These algorithms can be used to track AR model with Markov regime whose parameters vary slowly with time. The pertinent notion of convergence is in the sense of weak convergence (see [14], [23], [25]). The algorithm of interest is

$$\varphi_{n+1}^{\varepsilon} = \Pi_G\left(\varphi_n^{\varepsilon} + \varepsilon S(\tilde{Y}_n;\, \varphi_n^{\varepsilon})\right).$$

Again a constraint set is used for the estimation scheme. In the subsequent development, to save some notation, we often write φ_n in lieu of φ_n^{ε} for simplicity, and retain the ε dependence whenever necessary. To proceed, rewrite the recursion as

$$\varphi_{n+1} = \varphi_n + \varepsilon S(\tilde{Y}_n;\, \varphi_n) + \varepsilon M_n \tag{39}$$

where M_n is the reflection or projection term.

Define the piecewise-constant interpolations by

$$\varphi^{\varepsilon}(t) = \begin{cases} \varphi_0^{\varepsilon}, & \text{for } t < 0 \\ \varphi_n^{\varepsilon}, & \text{for } t \geq 0 \text{ and } t \in [n\varepsilon,\, n\varepsilon + \varepsilon) \end{cases}$$

and

$$M^{\varepsilon}(t) = \begin{cases} 0, & \text{for } t < 0 \\ \varepsilon \sum\limits_{i=0}^{\lfloor t/\varepsilon \rfloor - 1} M_i^{\varepsilon}, & \text{for } t \geq 0 \end{cases}$$

where $\lfloor t/\varepsilon \rfloor$ is the integer part of t/ε.

Theorem 5.1: Suppose that φ_0^{ε} converges weakly to φ_0 as $\varepsilon \rightarrow 0$, that for each $\tilde{M} > 0$, the stationary sequence $\{S(\tilde{Y}_j;\, \varphi);\, |\varphi| \leq \tilde{M}\}$ is uniformly integrable, that $S(\tilde{Y};\cdot)$ is continuous for each $\tilde{Y}$, that $ES(\tilde{Y}_j;\, \varphi) = H(\varphi)$ and $H(\varphi)$ is continuous, and that for each φ and each μ, as $n \rightarrow \infty$

$$\frac{1}{n} \sum_{i=\mu}^{\mu+n-1} E_{\mu} S(\tilde{Y}_i;\, \varphi) \rightarrow H(\varphi) \text{ in probability.} \tag{40}$$

Then $\varphi^{\varepsilon}(\cdot)$ converges weakly to $\varphi(\cdot)$ that is a solution of the ODE (15), provided the ODE has a unique solution for each initial condition.

Let $\{q^{\varepsilon}\}$ be a sequence of real numbers such that $\varepsilon q^{\varepsilon} \rightarrow \infty$ as $\varepsilon \rightarrow 0$. Then for almost all ω, $\varphi(\omega,\, \cdot)$ the limit of $\varphi(\varepsilon q_{\varepsilon} + \cdot)$ belongs to an invariant set of (15). If L_G is asymptotically stable, then the invariant set is in L_G. In addition, suppose that $\varphi^* \in G^*$ is the unique point such that $H(\varphi) = 0$. Then $\theta^{\varepsilon}(\varepsilon q^{\varepsilon} + \cdot)$ converges weakly to φ^*.

Remark 5.2: Note that compared to the w.p. 1 convergence, the conditions here are much weaker. In fact, only weak ergodicity in the form of (37) is needed. If the stronger geometric ergodicity holds (see the sufficient conditions given in Section III), then (37) is automatically satisfied. The proof of the theorem essentially follows from the development of [25, Ch. 8]. Owing to the projection, $\{\varphi_n^{\varepsilon}\}$ is bounded, so it is tight. Then all the conditions in [25, Theorem 8.2.2] are satisfied. The assertion follows. To proceed, we state a rate of convergence result below.

Theorem 5.3: Suppose that the conditions of Theorem 5.1 hold, that there is a nondecreasing sequence of real numbers $\{\tilde{t}_{\varepsilon}\}$ satisfying $\tilde{t}_{\varepsilon} \rightarrow \infty$ as $\varepsilon \rightarrow 0$ such that $\varphi^{\varepsilon}(\tilde{t}_{\varepsilon} + \cdot)$ converges weakly to the process with constant value $\varphi^* \in G^0$, that there exists $t_{\varepsilon} \geq \tilde{t}_{\varepsilon}$ such that $\{(\varphi_{n+\lfloor t_{\varepsilon}/\varepsilon \rfloor}^{\varepsilon} - \varphi^*)/\sqrt{\varepsilon}\}$ is tight, and that A5), A7) a), and A8) are satisfied. Define

$$U_n^{\varepsilon} = \frac{\varphi_{\lfloor t_{\varepsilon}/\varepsilon \rfloor + n}^{\varepsilon} - \varphi^*}{\sqrt{\varepsilon}}$$
$$U^{\varepsilon}(t) = U_n, \qquad \text{for } t \in [n\varepsilon, n\varepsilon + \varepsilon)$$
$$W^{\varepsilon}(t) = \sqrt{\varepsilon} \sum_{i=\lfloor t_{\varepsilon} \rfloor}^{\lfloor t_{\varepsilon}/\varepsilon \rfloor + \lfloor t/\varepsilon \rfloor - 1} S(\tilde{Y}_i;\, \varphi^*), \qquad \text{for } t \geq 0.$$

Then $(U^{\varepsilon}(\cdot),\, W^{\varepsilon}(\cdot))$ converges weakly in $D^{2r(r+1)}[0,\, \infty)$ to $(U(\cdot),\, W(\cdot))$, and

$$dU = H_{\varphi}(\varphi^*)U\,dt + dW$$

where $W(\cdot)$ is a Brownian motion having covariance Σt with Σ given by (21).

VI. Numerical Examples

We refer the reader to [10] for several numerical examples that illustrate the performance of the RMLE and the recursive expectation–maximization (EM) algorithm for HMM parameter estimation. Also [11] presents several numerical examples that illustrate the performance of the recursive EM algorithm in estimating AR processes with Markov regime. (The recursive EM algorithm is identical to the RMLE algorithm apart from the fact that it uses a different step size). Our aim here is to illustrate the performance of the RMLE algorithm for parameter estimation of AR models with Markov regime in two numerical examples. In [21], off-line ML parameter estimation was performed on these two numerical examples.

Example 1: Linear AR Model With Gaussian Noise: Consider a second-order ($d = 2$) AR model of the type

$$Y_n = -b_{X_n,1}Y_{n-1} - b_{X_n,2}Y_{n-2} + \sigma_{X_n}e_n.$$

Let $r = 2$, $n = 20\,000$, and let the true parameters be

$$A^0 = \begin{bmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{bmatrix} \quad B^0 = \begin{bmatrix} 0.5 & 0.5 \\ -0.5 & 0.8 \end{bmatrix} \quad \sigma^0 = \begin{bmatrix} 0.3 \\ 0.8 \end{bmatrix}.$$

TABLE I
LINEAR CASE WITH NORMAL ERRORS

Algorithm	$\begin{bmatrix}\hat{a}_{11}\\ \hat{a}_{22}\end{bmatrix}$	$\hat{B}$		$\hat{\sigma}$
$\varepsilon_k = 0.3/k$	0.8987 (0.0178)	0.4793 (0.0187)	0.4718 (0.0185)	0.2841 (0.0055)
	0.9189 (0.0096)	−0.5023 (0.0152)	0.8020 (0.0141)	0.7862 (0.0115)
Averaging	0.9408 (0.0052)	0.5009 (0.0118)	0.5003 (0.0131)	0.2955 (0.0050)
$(\varepsilon_k = 0.1/k^{0.7})$	0.9440 (0.0053)	−0.5014 (0.0101)	0.7993 (0.0103)	0.7957 (0.0083)

TABLE II
NONLINEAR AUTOREGRESSION WITH NORMAL ERRORS

Algorithm	$\begin{bmatrix}\hat{a}_{11}\\ \hat{a}_{22}\end{bmatrix}$	$\hat{B}$	$\hat{\sigma}$
$\varepsilon_k = 0.1/k$ (for $\hat{A}$ and $\hat{B}$)	0.8962 (0.0060)	0.4022 (0.0187)	0.1002 (0.0016)
$10^{-3}/k$ for $\hat{\sigma}$	0.9035 (0.0085)	0.8953 (0.0179)	0.1006 (0.0025)
Averaging	0.8987 (0.0040)	0.4009 (0.0051)	0.1002 (0.0009)
$\varepsilon_k = 10^{-2}/k^{0.7}, 10^{-3}/k^{0.8}$	0.9010 (0.0039)	0.8965 (0.0089)	0.1005 (0.0013)

The parameter vector can be taken as

$$\varphi = (a_{11},\, a_{22},\, b_{11},\, b_{12},\, b_{21},\, b_{22},\, \sigma_1 \sigma_2)$$

i.e., $p = 8$. Fifty independent sample paths based on the above model were generated. For each sample path, the RMLE algorithm was run initialized at

$$\varphi_0 = [0.5,\, 0.5,\, 0.0,\, 0.0,\, 0.0,\, 0.0,\, 0.1,\, 0.2].$$

Table I gives the sample means and standard deviations (in parenthesis) over these 50 replications for various step sizes with and without averaging.

Comments: The best results were obtained for $\gamma = 0.7$. We found that for the algorithms with iterate averaging and step sizes of the form k^γ, $\gamma < 0.6$, the RMLE algorithm became numerically ill-conditioned for some sample paths. Averaging of both observations and iterates appears to have better transient characteristics.

Fixed-Step Size Tracking: The following time-varying linear AR model was simulated:

$$A^0 = \begin{bmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{bmatrix} \quad B^0 = \begin{bmatrix} 0.5 & 0.5 \\ -0.5 & 0.8 \end{bmatrix} \quad \sigma^0 = \begin{bmatrix} 0.3 \\ 0.8 \end{bmatrix} \qquad n < 25\,000$$

$$A^0 = \begin{bmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix} \quad B^0 = \begin{bmatrix} 0.1 & -0.4 \\ -0.9 & 0.2 \end{bmatrix} \quad \sigma^0 = \begin{bmatrix} 0.1 \\ 0.3 \end{bmatrix} \qquad 25\,000 \le n \le 50\,000$$

Fig. 1 illustrates the tracking performance of the RMLE algorithm for step sizes of $\epsilon = 10^{-3}$ and $\epsilon = 3 \times 10^{-3}$. The RMLE algorithm was initialized at

$$\varphi_0 = [0.5,\, 0.5,\, 0.0,\, 0.0,\, 0.0,\, 0.0,\, 0.1,\, 0.2].$$

Example 2: Nonlinear Autoregression With Gaussian Noise: Here we consider a first-order $(d = 1)$ nonlinear autoregression of the type

$$Y_n = \exp(-b_{X_n} Y_{n-1}^2) + \sigma_{X_n} e_n$$

where $\{e_k\}$ is an i.i.d. sequence of standard normal random variables.

Let $r = 2$, $n = 40\,000$, and the true parameters be

$$A^0 = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \qquad B^0 = \begin{bmatrix} 0.4 \\ 0.9 \end{bmatrix} \qquad \sigma^0 = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}.$$

The initial parameter estimate

$$\varphi_0 = (a_{11},\, a_{22},\, b_1(1),\, b_2(1),\, \sigma_1(1),\, \sigma_2(1))$$

was chosen as $(0.5,\, 0.5,\, 0.7,\, 0.3,\, 0.5,\, 0.5)$.

Table II gives the sample means and standard deviations (in parenthesis) over these 50 replications for various step sizes with and without averaging.

Here b_2 is more difficult to estimate than b_1, which might be explained by the inequality $b_2 > b_1$; the exponential function $\exp(-b_i y^2)$ decays faster and is thus smaller in comparison to the noise for $i = 2$.

By conducting several numerical experiments for the above nonlinear autoregressive model we noticed that the convergence of the RMLE algorithm was sensitive to initialization of B_0. The closer the initial value $b_1(1)$ was picked to $b_2(1)$, the slower the initial convergence of the algorithm. For initializations

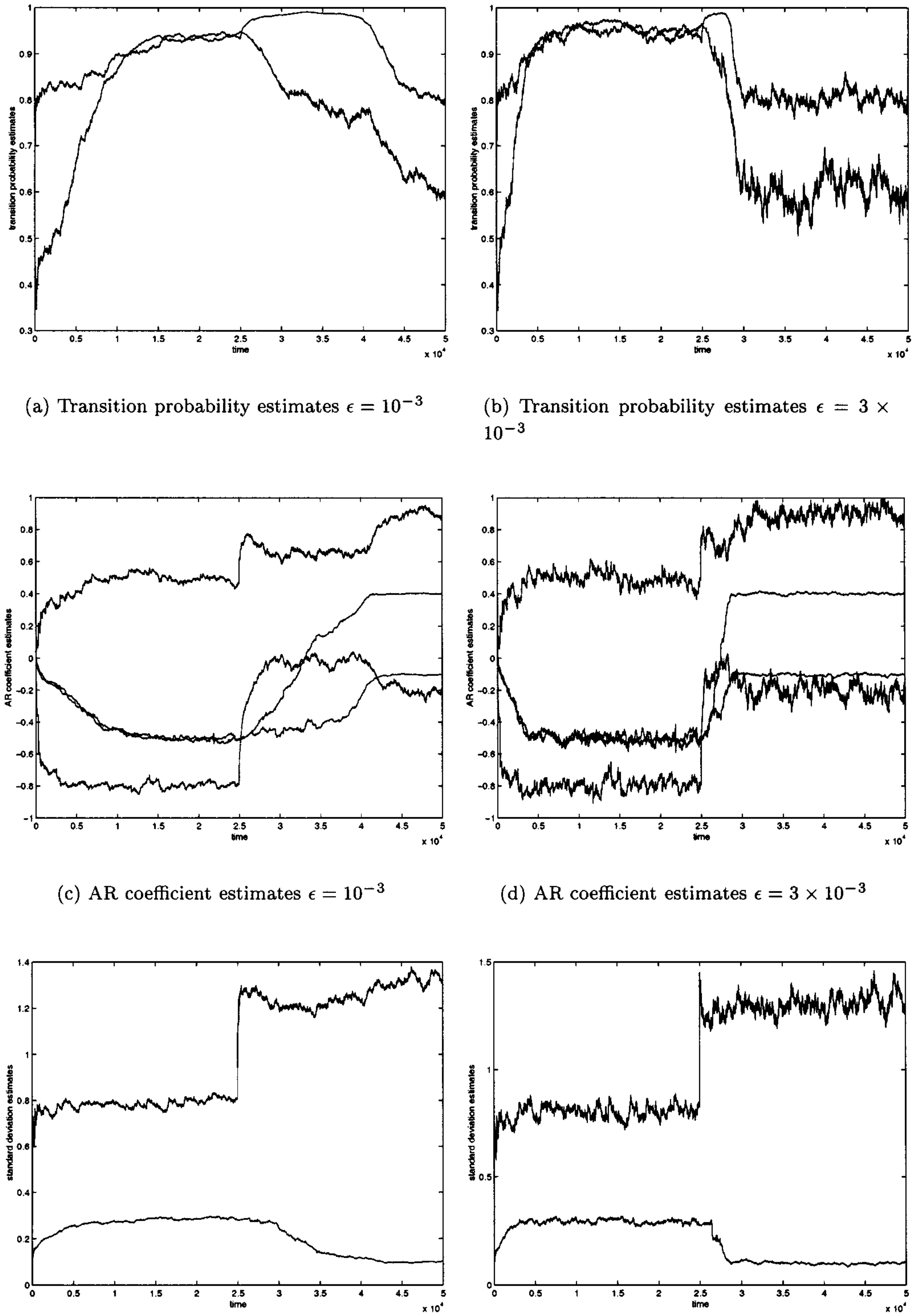
$$|b_1(1) - b_2(1)| > 0.05$$

and in the region $0 < b_1(1) < 2$ and $0 < b_2(1) < 2$ the algorithm converged to the true parameter values within 40 000 time points.

Fixed-Step Size Tracking: The following time-varying version of the above nonlinear AR model was simulated:

$$A^0 = \begin{bmatrix} 0.90 & 0.10 \\ 0.10 & 0.90 \end{bmatrix} \qquad B^0 = \begin{bmatrix} 0.4 \\ 0.9 \end{bmatrix} \qquad \sigma^0 = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} \qquad n < 25\,000$$

$$A^0 = \begin{bmatrix} 0.6 & 0.4 \\ 0.6 & 0.4 \end{bmatrix} \qquad B^0 = \begin{bmatrix} 0.2 \\ 0.7 \end{bmatrix} \qquad \sigma^0 = \begin{bmatrix} 0.3 \\ 0.8 \end{bmatrix} \qquad 25\,000 \le n \le 60\,000.$$



(a) Transition probability estimates $\epsilon = 10^{-3}$

(b) Transition probability estimates $\epsilon = 3 \times 10^{-3}$

(c) AR coefficient estimates $\epsilon = 10^{-3}$

(d) AR coefficient estimates $\epsilon = 3 \times 10^{-3}$

(e) Standard deviation estimate $\epsilon = 10^{-3}$

(f) Standard deviation estimate $\epsilon = 3 \times 10^{-3}$

Fig. 1. Tracking performance of RMLE for linear AR model. Step sizes are $\epsilon = 10^{-3}$ and 3×10^{-3}, respectively. The parameters are specified in Section VI.

Fig. 2 illustrates the tracking performance of the RMLE algorithm for step sizes of

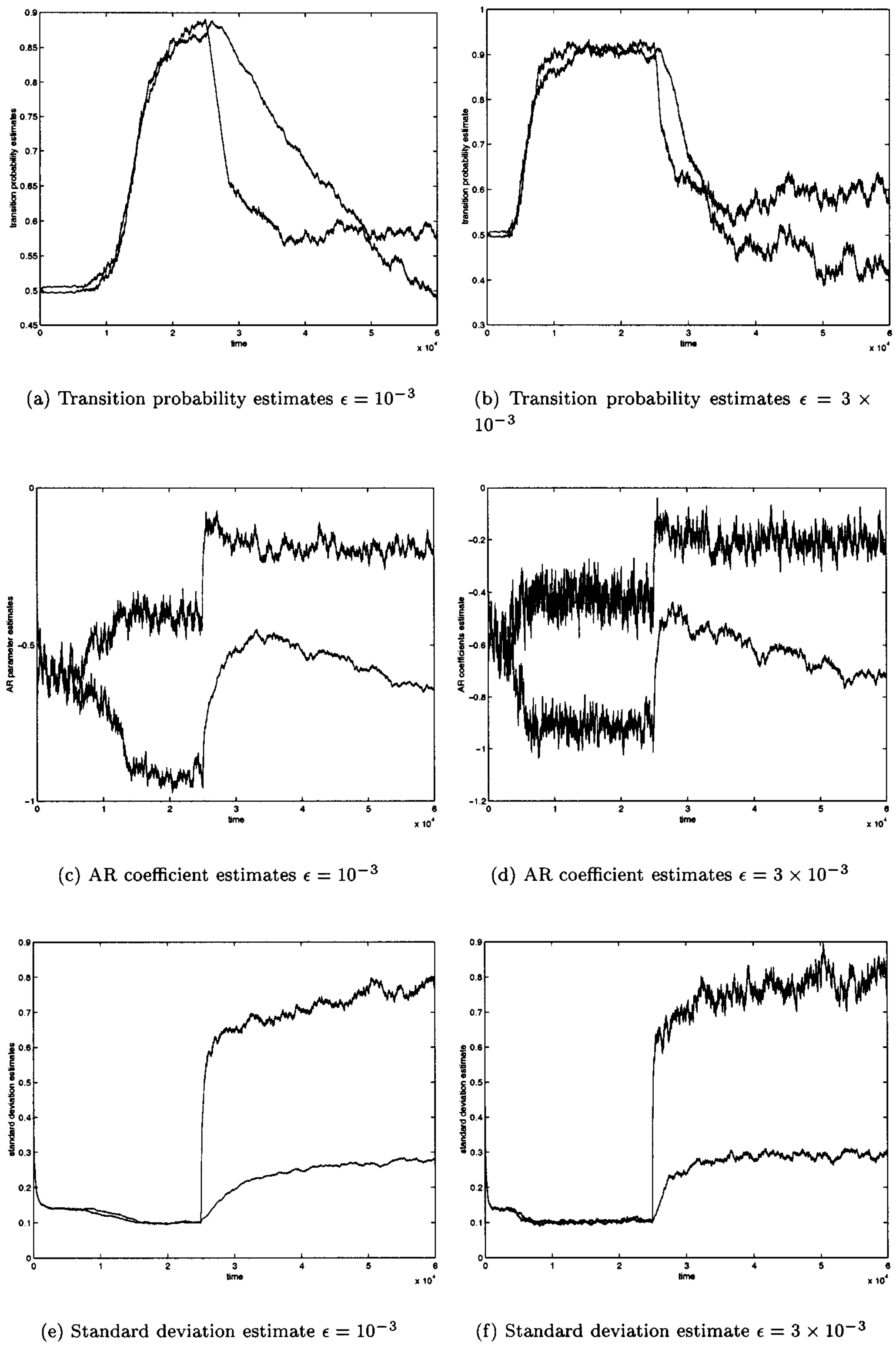
$$\epsilon = 10^{-3} \quad \text{and} \quad \epsilon = 3 \times 10^{-3}.$$

The RMLE algorithm was initialized at

$$\varphi_0 = [0.5, 0.5, 0.0, 0.0, 0.0, 0.0, 0.1, 0.2].$$

VII. Conclusion and Extensions

We have focused on developing asymptotic properties of recursive estimators of stochastic approximation type for hidden Markov estimation. Convergence and rate of convergence results are obtained for both decreasing and constant step-size algorithms. In addition, we have demonstrated that algorithms



(a) Transition probability estimates $\epsilon = 10^{-3}$

(b) Transition probability estimates $\epsilon = 3 \times 10^{-3}$

(c) AR coefficient estimates $\epsilon = 10^{-3}$

(d) AR coefficient estimates $\epsilon = 3 \times 10^{-3}$

(e) Standard deviation estimate $\epsilon = 10^{-3}$

(f) Standard deviation estimate $\epsilon = 3 \times 10^{-3}$

Fig. 2. Tracking performance of RMLE for nonlinear AR model. Step sizes are $\epsilon = 10^{-3}$ and 3×10^{-3}, respectively. The parameters are specified in Section VI.

with averaging in both iterates and observations are asymptotically optimal in the sense they have the best scaling factor and achieve the "smallest possible" variances. For future research, it is both interesting and important to design continuous-time recursive estimators for hidden Markov estimation. It will be of interest from a practical point of view to consider problems under simulation based setting. Recent efforts in this direction can be found in the work of Ho and Cao [18], Konda and Borkar [8], L'Ecuyer and Yin [26], Tang, L'Ecuyer, and Chen [40] among others.

ACKNOWLEDGMENT

The authors gratefully thank the reviewers and the Associated Editor for the detailed comments and suggestions, which lead to much improvement of the paper.

REFERENCES

[1] A. Arapostathis and S. I. Marcus, "Analysis of an identification algorithm arising in the adaptive estimation of Markov chains," *Math. Control, Signals Syst.*, vol. 3, pp. 1–29, 1990.

[2] J. A. Bather, "Stochastic approximation: A generalization of the Robbins–Monro procedure," in *Proc. 4th Prague Symp. Asymptotic Statist.*, P. Mandl and M. Hušková, Eds., 1989, pp. 13–27.

[3] V. S. Borkar, "On white noise representations in stochastic realization theory," *SIAM J. Contr. Optim.*, vol. 31, pp. 1093–1102, 1993.

[4] ——, "Stochastic approximation with two time scales," *Syst. Contr. Lett.*, vol. 29, pp. 291–294, 1997.

[5] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968.

[6] P. Bougerol and N. Picard, "Strict stationarity of generalized autoregressive processes," *Ann. Probab.*, vol. 20, pp. 1714–1730, 1992.

[7] A. Brandt, "The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients," *Adv. Appl. Proban.*, vol. 18, pp. 211–220, 1986.

[8] V. R. Konda and V. S. Borkar, "Action-critic-type learning algorithms for Markov decision processes," *SIAM J. Contr. Optim.*, vol. 38, pp. 94–123, 1999.

[9] K. L. Chung, "On a stochastic approximation method," *Ann. Math. Statist.*, vol. 25, pp. 463–483, 1954.

[10] I. Collings, V. Krishnamurthy, and J. B. Moore, "On-line identification of hidden Markov models via recursive prediction error techniques," *IEEE Trans. Signal Processing*, vol. 42, pp. 3535–3539, 1994.

[11] S. Dey, V. Krishnamurthy, and T. Salmon-Legagneur, "Estimation of Markov modulated time-series via the EM algorithm," *IEEE Signal Processing*, vol. 1, pp. 153–155, 1994.

[12] J. L. Doob, *Stochastic Processes*. New York: Wiley, 1953.

[13] R. Douc, E. Moulines, and T. Rydén. (2001) Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. Preprint. [Online]. Available: http://www.maths.lth.se/matstat/staff/tobias/hmmarasnorm.ps

[14] S. N. Ethier and T. G. Kurtz, *Markov Processes, Characterization and Convergence*. New York: Wiley, 1986.

[15] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, pp. 357–384, 1989.

[16] ——, "Analysis of time series subject to changes in regime," *J. Econometrics*, vol. 45, pp. 39–70, 1990.

[17] J. D. Hamilton and R. Susmel, "Autoregressive conditional heteroskedasticity and changes in regime," *J. Econometrics*, vol. 64, pp. 307–333, 1994.

[18] Y. C. Ho and X. R. Cao, *Perturbation Analysis of Discrete Event Dynamic Systems*. Boston, MA: Kluwer, 1991.

[19] U. Holst, G. Lindgren, J. Holst, and M. Thuvesholmen, "Recursive estimation in switching autoregressions with a Markov regime," *Time Series Anal.*, vol. 15, pp. 489–506, 1994.

[20] V. Krishnamurthy and J. B. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback–Leibler information measure," *IEEE Trans. Signal Processing*, vol. 41, pp. 2557–2573, 1993.

[21] V. Krishnamurthy and T. Rydén, "Consistent estimation of linear and nonlinear autoregressive models with Markov regime," *Time Series Anal.*, vol. 19, pp. 291–307, 1998.

[22] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. New York: Springer-Verlag, 1978.

[23] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, With Applications to Stochastic Systems Theory*. Cambridge, MA: MIT Press, 1984.

[24] H. J. Kushner and J. Yang, "Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes," *SIAM J. Contr. Optim.*, vol. 31, pp. 1045–1062, 1993.

[25] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*. New York: Springer-Verlag, 1997.

[26] P. L'Ecuyer and G. Yin, "Budget-dependent convergence rate of stochastic approximation," *SIAM J. Optim.*, vol. 8, pp. 217–247, 1998.

[27] F. LeGland and L. Mevel. (1996, July) Geometric ergodicity in hidden Markov models. PI-1028, IRISA Res. Rep.. [Online]. Available: ftp://ftp.irisa.fr/techreports/1996/PI-1028.ps.gz

[28] ——, "Asymptotic properties of the MLE in hidden Markov models," in *Proc. 4th European Control Conf.*, Brussels, Belgium, July 1–4, 1997.

[29] ——, "Recursive estimation of hidden Markov models," in *Proc. 36th IEEE Conf. Decision Control*, San Diego, CA, Dec. 1997.

[30] B. G. Leroux, "Maximum-likelihood estimation for hidden Markov models," *Stochastic Process Applic.*, vol. 40, pp. 127–143, 1972.

[31] H. B. Mann and A. Wald, "On the statistical treatment of linear stochastic difference equations," *Econometrica*, vol. 11, pp. 173–220, 1943.

[32] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. London, U.K.: Springer-Verlag, 1993.

[33] B. T. Polyak, "New method of stochastic approximation type," *Automation Remote Contr.*, vol. 7, pp. 937–946, 1991.

[34] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, 1989.

[35] T. Rydén, "On recursive estimation for hidden Markov models," *Stochastic Process Applic.*, vol. 66, pp. 79–96, 1997.

[36] ——, "Asymptotic efficient recursive estimation for incomplete data models using the observed information,", to be published.

[37] D. Ruppert, "Stochastic approximation," in *Handbook in Sequential Analysis*, B. K. Ghosh and P. K. Sen, Eds. New York: Marcel Dekker, 1991, pp. 503–529.

[38] R. Schwabe, "Stability results for smoothed stochastic approximation procedures," *Z. Angew. Math. Mech.*, vol. 73, pp. 639–644, 1993.

[39] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[40] "Central limit theorems for stochastic optimization algorithms using infinitesimal perturbation analysis," *Discr. Event Dyn. Syst.*, vol. 10, pp. 5–32, 2000.

[41] Q.-Y. Tang, P. L'Ecuyer, and H.-F. Chen, "Asymptotic efficiency of perturbation analysis-based stochastic approximation with averaging," *SIAM J. Contr. Optim.*, vol. 37, pp. 1822–1847, 1999.

[42] I. J. Wang, E. Chong, and S. R. Kulkarni, "Equivalent necessary and sufficient conditions on noise sequences for stochastic approximation algorithms," *Adv. Appl. Probab.*, vol. 28, pp. 784–801, 1996.

[43] J. Yao and J. G. Attali, "On stability of nonlinear AR processes with Markov switching," unpublished, 1999.

[44] G. Yin, "On extensions of Polyak's averaging approach to stochastic approximation," *Stochastics*, vol. 36, pp. 245–264, 1992.

[45] G. Yin, "Rates of convergence for a class of global stochastic optimization algorithms," *SIAM J. Optim.*, vol. 10, pp. 99–120, 1999.

[46] G. Yin and K. Yin, "Asymptotically optimal rate of convergence of smoothed stochastic recursive algorithms," *Stochastics Stochastic Repts.*, vol. 47, pp. 21–46, 1994.